Statistique Appliquée

Luc Deneire

Iannis Aliferis

École Polytechnique de l'Université de Nice – Sophia Antipolis
Polytech'Nice Sophia
Département d'Électronique, 3e année, 2008–2009

deneire@unice.fr

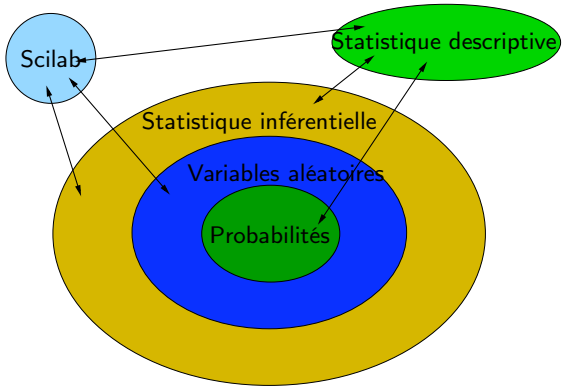
Ce document contient une grande partie des transparents du cours. Cela signifie qu'il n'est en aucun cas complet (auto-suffisant) ; une grande quantité d'information (commentaires, explications, diagrammes, démonstrations etc.) est donnée pendant les séances, oralement ou à l'aide du tableau, en plus de nombreux transparents « extra » qui ne sont pas inclus ici.

 Le logo du logiciel R à droite d'un titre contient un lien vers le script utilisé pour produire les résultats présentés dans le transparent. L'exécution, l'étude et la compréhension des scripts font partie intégrante du cours.

Document préparé avec LaTeX et le package `powerdot`.

Introduction 2

Le cours en bref



3

Plan du cours

- Rappels sur les probabilités
  - différentes définitions
  - probabilité conditionelle
  - indépendance
- Variables aléatoires (discrètes et continues)
  - fonction/densité de probabilité
  - espérance, variance, moments
  - indépendance entre v.a.
- Statistique descriptive
  - moyenne, écart-type, quartiles, ...
  - histogrammes, boîtes à moustaches
- Statistique inférentielle
  - estimation
  - intervalles de confiance
  - tests d'hypothèse

4

Bibliographie

- Probabilités, Variables Aléatoires :
  - P. Bogaert, "Probabilités pour scientifiques et ingénieurs", De Boeck, Bruxelles, 2006
  - D. Bertsekas, J. Tsitsiklis, "Introduction to Probability", Athena Scientific, Belmont, 2002
- Statistique :
  - T.H. Wonnacott, R.J. Wonnacott, "Introductory Statistics", 5th ed., Wiley, 1990
  - R.E. Walpole, R.H. Mayers, "Probability and Statistics for Engineers and Scientists", Prentice Hall International, 1993.
- R (livres disponibles en ligne) :
  - E. Paradis, "R pour les débutants", 2005
  - W. N. Venables, D. M. Smith and the R Development Core Team, "An introduction to R", 2006
  - W. J. Owen, "The R Guide", 2006

5

Évaluation

- 30% (6/20) : contrôle final (semaine 6/2009)
- 30% (6/20) : contrôle intermédiaire (semaine 49)
- 20% (4/20) : Devoir 1 (15/11 $\longrightarrow$ 18/11)
- 20% (4/20) : Devoir 2 (17/01 $\longrightarrow$ 20/01)
  - énoncés en ligne (www.i3s.unice.fr/~ deneire)
  - travail individuel $\longrightarrow$ rédaction individuelle
  - citer les sources / documents / personnes (brièvement à la première page)
  - plagiat $\longrightarrow -$ 20% $= -$ 4/20 (0 ailleurs)

6

Introduction aux probabilités 7

Les probabilités : Pourquoi faire ?

- **az-zahr** mot arabe qui signifie dé
- **hasard** jeu de dés au moyen âge
- **principe d'incertitude** Heisenberg : $\sigma_x.\sigma_p \geq \frac{h}{2}$
- $\Delta E.\Delta t \geq \frac{h}{2}$
- **incertitudes dans les transistors**

8

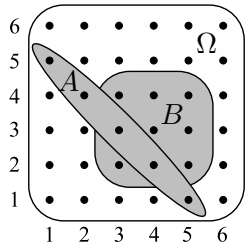
Definitions

- **Expérience aléatoire** : plusieurs résultats possibles
- **Issue** ou **éventualité** ω : un des résultats possibles
- **Univers Ω** : l'ensemble de *tous* les résultats
- **Événement $\mathbf{A}$** : un sous-ensemble de Ω

- Exemple :
  - « Compter le nombre de personnes présentes »
  - $\omega_1 = 1$ (au moins...), $\omega_2 = 70$, etc.
  - $\Omega = \{1, 2, \ldots, N_{\text{max}}\}$
  - $A = \{\text{il y a moins de 5 personnes}\} = \{1, 2, 3, 4\} \subset \Omega$

9

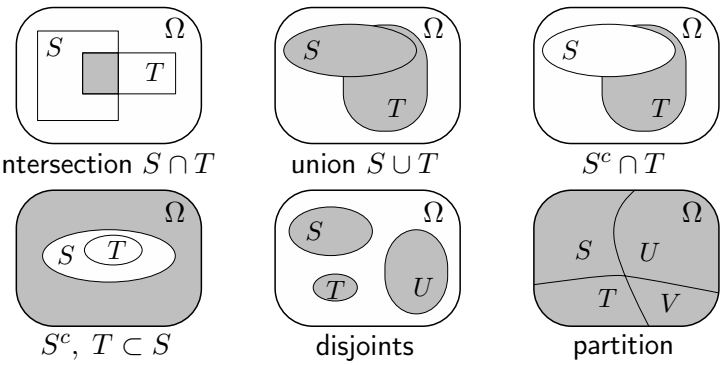
Exemple : lancer deux dés



- $\omega_1 = (1,1)$, $\omega_2 = (3,4)$, $\omega_3 = (4,3)$, ...
- $\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\}$
- $A = \{$la somme est égale à 6$\}$
- $B = \{$le 1er est entre 3 et 5 ; le 2nd entre 2 et 4$\}$

10

Ensembles



- Disjoints : $\bigcap_i S_i = \emptyset$ (mutuellement exclusifs)
- Partition : S_i disjoints *et* $\bigcup_i S_i = \Omega$
- De Morgan 1 : $(\bigcap_i S_i)^c = \bigcup_i S_i^c$
- De Morgan 2 : $(\bigcup_i S_i)^c = \bigcap_i S_i^c$

11

Modèle probabiliste

1. Définir l'ensemble Ω.
2. Attribuer un nombre $P(A) \in [0,1]$ à un événement A.

□ Définition classique (Laplace)

$$P(A) = \frac{\text{nombre de cas équiprobables favorables}}{\text{nombre de cas équiprobables possibles}}$$

□ Définition intuitive (fréquence relative)

$$P(A) = \lim_{n\to\infty} \frac{N_n(A)}{n}$$

□ Définition axiomatique (Kolmogorov)

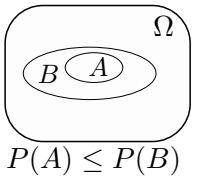
1. $P(A) \geq 0$ pour chaque événement $A \subseteq \Omega$
2. $P(A \cup B) = P(A) + P(B)$ pour A et B disjoints
3. $P(\Omega) = 1$

12

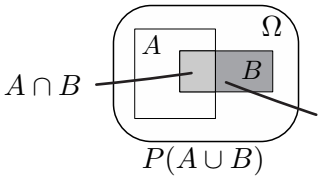
Propriétés

1. $P(A^c) = 1 - P(A)$
   dém. : $P(\Omega) = P(A \cup A^c) \overset{A\cap A^c=\emptyset}{=} P(A) + P(A^c) = 1$
2. $P(\emptyset) = 0 = P(\Omega^c)$
3. Si $A \subset B$, $P(A) \leq P(B)$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
5. $P(A \cup B) \leq P(A) + P(B)$
6. $P(A \cup B \cup C) = P(A) + P(A^c \cap B) + P(A^c \cap B^c \cap C)$

□ Interprétation graphique :



$P(A) \leq P(B)$



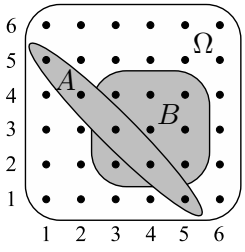
$P(A \cup B)$

13

Probabilité conditionnelle

Attribuer un nombre $P(A|B) \in [0,1]$ à un événement A,
sachant que l'événement $B\,(P(B) \neq 0)$ a été réalisé.

- Exemple : lancer deux dés



- Toutes les issues ω_i $(i = 1, \ldots, 36)$ sont équiprobables
- $P(A) =$ —
- $P(B) =$ —
- $P(A|B) =$ - $= \frac{/36}{/36}$
-

$$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)}}$$

14

Un nouvel Univers

- La probabilité conditionnelle satisfait les trois axiomes :
  1. $P(A|B) = \frac{P(A \cap B)}{P(B)} \geq 0$ pour chaque événement $A \subseteq \Omega$
  2. $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B)$ pour A_1 et A_2 disjoints
  3. $P(\Omega|B) = 1$ (univers Ω)
- Les propriétés générales restent valables, p.ex.,
  $P(A \cup C|B) \leq P(A|B) + P(C|B)$
- On peut remplacer 3. par
  3'. $P(B|B) = \frac{P(B \cap B)}{P(B)} = 1$ (univers B)
- $P(A|B)$: *loi* de probabilité ; $\boxed{\text{univers : } \Omega \to B}$!
- Approche séquentielle :
  - $\boxed{P(A \cap B) = P(B)P(A|B)}$
  - $P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots P\left(A_n|\bigcap_{i=1}^{n-1} A_i\right)$

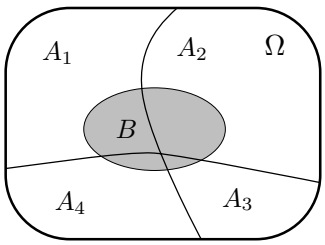
15

Exemple : fausse alarme

- Système radar
  - Avion : Présent / Absent
  - Radar : Détection / Non détection
  - Quatre issues possibles, $\Omega = \{(P,D),(A,D),(P,N),(A,N)\}$
  - $S = \{\text{un avion est présent}\} = \{(P,D),(P,N)\}$
  - $T = \{\text{le radar signale la présence d'un avion}\} = \{(P,D),(A,D)\}$
  - $P(S) = 0.05$ (présence d'un avion)
  - $P(T|S) = 0.99$ (détection si avion présent)
  - $P(T|S^c) = 0.10$ (fausse détection : « détection » si avion absent)
- Quelle est la probabilité d'une fausse alarme ?
  $P(S^c \cap T) =$ $= 0.095$
- Quelle est la probabilité qu'un avion ne soit pas détecté ?
  $P(S \cap T^c) =$ $= 0.0005$

16

Théorème de probabilité totale



- $A_1, A_2, \ldots, A_n$: une partition de Ω
- $B = (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n)$
- $B \cap A_1, B \cap A_2, \ldots, B \cap A_n$: événements disjoints
- $P(B) = P(B \cap A_1) + P(B \cap A_2) + \ldots + P(B \cap A_n)$
  $= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \ldots + P(A_n)P(B|A_n)$
- $\boxed{P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)}$ *Diviser pour régner !*

17

Théorème de Bayes

- « Cause » $A \longrightarrow$ « effet » $B, P(B|A), P(B) \neq 0$
- À partir de $P(B|A)$, calculer $P(A|B)$ (effet $\longrightarrow$ cause)
- $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$

$$\boxed{P(A|B) = P(A)\frac{P(B|A)}{P(B)}}$$

- Plusieurs causes A_i $(i = 1, \ldots, n)$, partition de Ω

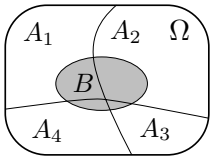
$$P(A_i|B) = P(A_i)\frac{P(B|A_i)}{P(B)}$$

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$$

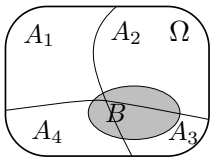
18

Inférence bayésienne

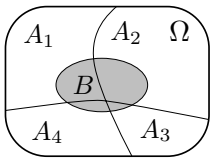
1. $P(A_i|B) = P(A_i)\frac{P(B|A_i)}{P(B)}$
   - $P(A_i)$: *a priori*
   - $P(A_i|B)$: *a posteriori*
   - $P(A_i|B) > P(A_i)$ si $P(B|A_i) > P(B)$
2. $P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$
   - $P(A_i)P(B|A_i) = P(B \cap A_i)$
   - $P(A_i|B) \propto P(B \cap A_i)$



$P(B|A_3) < P(B)$



$P(B|A_3) > P(B)$



$$P(A_2|B) > P(A_1|B) > P(A_4|B) > P(A_3|B)$$

19

Indépendance

1. Entre deux événements A et B :
   - □ $\boxed{P(A \cap B) = P(A)P(B)}$
   - □ si $P(B) \neq 0$, $\boxed{P(A|B) = \dfrac{P(A \cap B)}{P(B)} = P(A)}$
2. Entre deux événements A et B, conditionnés par $C, (P(C) \neq 0)$:
   - □ $P(A \cap B|C) = P(A|C)P(B|C)$
   - □ si $P(B|C) \neq 0, P(A|B \cap C) = P(A|C)$
3. Entre plusieurs événements $A_1, A_2, \ldots, A_n$:
   - □ $P\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} P(A_i)$ pour *chaque* S, sous-ensemble de $\{1, 2, \ldots, n\}$

20

Quelques stratégies

- □ Définir Ω
- □ . . . ou juste *compter* ses éléments. . .
- □ Issues équiprobables : $P(A) = \frac{\text{card}(A)}{\text{card}(\Omega)}$ (Laplace)
- □ Approche séquentielle (+ indépendance)
- □ Probabilité totale (trouver une partition)
- □ $P(B|A) \longrightarrow P(A|B)$: Bayes

21

Compter = multiplier...

- Opération à M étapes,
- chacune pouvant être réalisée selon N_i façons ($i = 1, \ldots, M$).
- Nombre total des réalisations :

$$N = N_1 N_2 \ldots N_M = \prod_{i=1}^{M} N_i$$

1. Permutations de n objets

$$n(n-1)(n-2)\ldots 2 \cdot 1 = \boxed{n!}$$

2. Permutations de k objets choisis parmi n

$$ {}_nP_k = n(n-1)(n-2)\ldots[n-(k-1)] = \boxed{\frac{n!}{(n-k)!}} = {}_nC_k\, k!$$

$({}_nP_n = n! \longrightarrow 0! = 1)$

22

...ou diviser !

3. Combinaisons de k objets choisis parmi n

$$ {}_nC_k = \binom{n}{k} = \frac{{}_nP_k}{k!} = \boxed{\frac{n!}{k!(n-k)!}}$$

4. Repartitions de n objets dans $n_1, n_2, \ldots, n_r$ groupes

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1! n_2! \ldots n_r!}\ ,\ (n_1 + n_2 + \ldots + n_r = n)$$

Méthode générale (*par étape*) :

- n objets : $n!$ permutations
- n_i objets non distincts (identiques ou combinaisons) : diviser par $n_i!$
- répéter pour tous les groupes d'objets

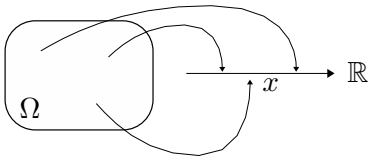
Multiplier pour toutes les étapes.

23

Variable Aléatoire Discrète (une seule) 24

Définition

- Associer *une valeur réelle* x à chaque issue ω d'une expérience aléatoire
- Variable aléatoire discrète (VAD) :
  Nombre de valeurs possibles : fini ou infini dénombrable



- Une variable aléatoire est une fonction ! ($\Omega \to \mathbb{R}$)
- X : la variable aléatoire / x : une valeur possible
- Fonction de probabilité $p_X(x)$:
  $P(\underbrace{\{X = x\}}_{\text{événement} \in \Omega}) \overset{\text{simpl.}}{=} P(X = x) \triangleq p_X(x)$

25

V.A. : à usage unique

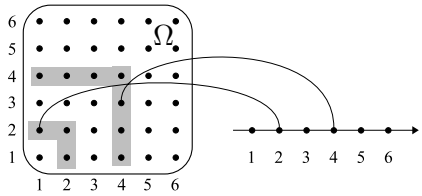
1. On « interroge » la v.a. X
2. L'expérience aléatoire associée est effectuée
3. Une issue $\omega \in \Omega$ est réalisée
4. À l'issue ω correspond une valeur x
5. La v.a. X « répond » avec la valeur x

- Une v.a. X :
  1. représente une expérience aléatoire
     et une association $\Omega \to \mathbb{R}$
  2. est à usage unique : une seule expérience effectuée !
- N v.a. $X_1, X_2, \ldots, X_N$ identiquement distribuées :
  1. représente, chacune, la même expérience aléatoire
     et la même association $\Omega \to \mathbb{R}$
  2. est, chacune, à usage unique :
     la même expérience répetée N fois !

26

Une partition naturelle de l'Univers

- Expérience : lancer deux dés ; X est la valeur maximale
  p.ex. : $p_X(2) = P(\{X = 2\}) = \frac{3}{36}$

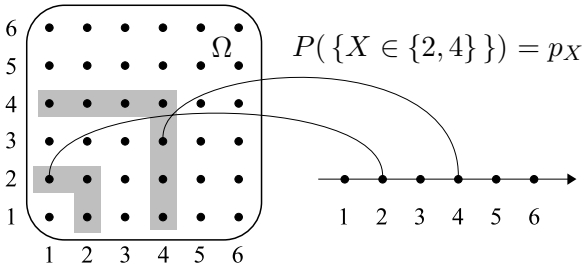


- $\bigcap_x \{X = x\} = \emptyset$
- $\bigcup_x \{X = x\} = \Omega$
- Les événements $\{X = x\}$ forment une partition de Ω

27

Fonction de Probabilité

- Normalisation :
  $\sum_x p_X(x) = \sum_x P(\{X = x\}) \overset{\text{disj.}}{=} P(\bigcup_x \{X = x\}) \overset{\text{part.}}{=} P(\Omega) = 1$
- $P(\{X \in S\}) \overset{\text{disj.}}{=} \sum_{x \in S} p_X(x)$



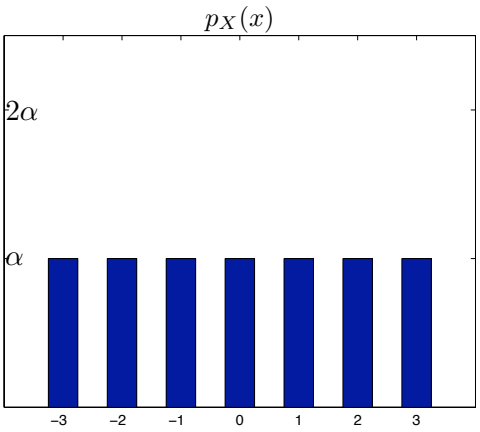
$P(\{X \in \{2,4\}\}) = p_X(2) + p_X(4) = \frac{3}{36} + \frac{7}{36}$

- Comment calculer $p_X(x)$:
  1. Trouver les valeurs possibles ; indiquer les valeurs impossibles
  2. Répérer les issues ω_i constituant l'événement $\{X = x\}$
  3. Additionner les probabilités $P(\omega_i)$

28

Fonction d'une V.A.

- Une fonction d'une V.A. est aussi une V.A.
- $Y = g(X)$
- $p_Y(y) = P(\{Y = y\}) = P\left(\{X \in S\}_{S=\{x|g(x)=y\}}\right) \overset{\text{disj.}}{=} \sum_{\{x|g(x)=y\}} p_X(x)$
- Exemple : X V.A. à distribution uniforme, $x \in \{-3, -2, \ldots, 3\}$; $Y = |X|$



- Normalisation : $\alpha =$

29

Espérance de X

- v.a.d. X ; m valeurs possibles
- classement par ordre : $x_{(1)} < x_{(2)} < \ldots < x_{(m)}$
- Comment calculer une valeur « moyenne » ?

1. Répéter la même expérience n fois !
   (Considérer n v.a. $X_1, X_2, \ldots, X_n$ identiquement distribuées)
2. Prendre la moyenne des n valeurs $x_1, x_2, \ldots, x_n$ obtenues :
   moyenne $= \frac{x_1+x_2+\ldots+x_n}{n}$
   $\overset{\text{regrouper}}{=} \frac{x_{(1)}N_n(x_{(1)})+x_{(2)}N_n(x_{(2)})+\ldots+x_{(m)}N_n(x_{(m)})}{n}$
   $\underset{n\to\infty}{\longrightarrow} x_{(1)}p_X(x_{(1)}) + x_{(2)}p_X(x_{(2)}) + \ldots + x_{(m)}p_X(x_{(m)})$
   $= \sum_x x p_X(x) \triangleq \mathrm{E}[X]$

30

Grandeurs statistiques

- Espérance

$$\boxed{\mu_X = \mathrm{E}[X] = \sum_x x p_X(x)}$$

  centre de gravité de la distribution :
  $\sum_x (x-c)p_X(x) = 0\ ,\ c = \mathrm{E}[X]$
  $p_X(x)$: « masse de probabilité »
- Variance

$$\boxed{\mathrm{var}[X] = \sigma_X^2 = \mathrm{E}\big[(X - \mathrm{E}[X])^2\big] \geq 0}$$

- Écart-type

$$\boxed{\sigma_X = \sqrt{\mathrm{var}[X]}}$$

- n-ième moment (moment d'ordre n) : $\mathrm{E}[X^n]$
- n-ième moment centré : $\mathrm{E}[(X - \mathrm{E}[X])^n]$

31

Espérance de $g(X)$

-

$$\boxed{\mathrm{E}[g(X)] = \sum_x g(x) p_X(x)}$$

- $Y = g(X)\ ,\ p_Y(y) = \sum_{\{x|g(x)=y\}} p_X(x)$
- $\mathrm{E}[g(X)] = \mathrm{E}[Y]$
  $= \sum_y y p_Y(y)$
  $= \sum_y y \sum_{\{x|g(x)=y\}} p_X(x)$
  $= \sum_y \sum_{\{x|g(x)=y\}} y p_X(x)$
  $= \sum_y \sum_{\{x|g(x)=y\}} g(x) p_X(x)$
  $= \sum_x g(x) p_X(x)$

32

Fonction linéaire

□

$$Y = aX + b$$

□

$$\boxed{\mathrm{E}[Y] = a\mathrm{E}[X] + b}$$

$$\boxed{\mathrm{var}[Y] = a^2\mathrm{var}[X]} \quad \boxed{\sigma_Y = |a|\sigma_X}$$

□ $\mathrm{E}[Y] = \mathrm{E}[aX+b] = \sum_x (ax+b)p_X(x) = a\sum_x xp_X(x) + b\sum_x p_X(x)$
$= a\mathrm{E}[X] + b$

□ $\mathrm{var}[Y] = \mathrm{var}[aX+b] = \mathrm{E}\Big[(aX+b-\mathrm{E}[aX+b])^2\Big]$
$= \mathrm{E}\Big[(aX+b-a\mathrm{E}[X]-b)^2\Big] = \mathrm{E}\Big[(aX-a\mathrm{E}[X])^2\Big]$
$= a^2\mathrm{E}\Big[(X-\mathrm{E}[X])^2\Big] = a^2\mathrm{var}[X]$

33

Calcul de la variance

□

$$\boxed{\mathrm{var}[X] = \mathrm{E}\big[X^2\big] - (\mathrm{E}[X])^2 \geq 0}$$

□ $\mathrm{var}[X] = \mathrm{E}\big[(X-\mathrm{E}[X])^2\big] = \sum_x (x-\mathrm{E}[X])^2 p_X(x)$
$= \sum_x \{x^2 - 2x\mathrm{E}[X] + (\mathrm{E}[X])^2\}p_X(x)$
$= \sum_x x^2 p_X(x) - 2\mathrm{E}[X]\sum_x xp_X(x) + (\mathrm{E}[X])^2\sum_x p_X(x)$
$= \mathrm{E}\big[X^2\big] - 2(\mathrm{E}[X])^2 + (\mathrm{E}[X])^2 = \mathrm{E}\big[X^2\big] - (\mathrm{E}[X])^2$

□ $\mathrm{var}[X] = \mathrm{E}[(X - \underbrace{\mathrm{E}[X]}_{\text{cste}})^2] = \mathrm{E}\Big[X^2 - 2X\mathrm{E}[X] + (\mathrm{E}[X])^2\Big]$
$= \mathrm{E}\big[X^2\big] - 2\mathrm{E}[X]\,\mathrm{E}[X] + (\mathrm{E}[X])^2 = \mathrm{E}\big[X^2\big] - (\mathrm{E}[X])^2$

34

Variables Aléatoires Discrètes (deux et plus) 35

Deux variables aléatoires

- X, Y : V.A. associées à la même expérience aléatoire
- Fonction de probabilité conjointe :
  $p_{XY}(x,y) \triangleq P(\underbrace{\{X=x\}}_{\text{événement} \in \Omega} \cap \underbrace{\{Y=y\}}_{\text{événement} \in \Omega}) \overset{\text{sim.}}{=} P(X=x, Y=y)$
- $P((X,Y) \in A) = \sum_{(x,y)\in A} p_{XY}(x,y)$
- Fonctions de probabilité marginales :
  $p_X(x) = \sum_y p_{XY}(x,y)$, $p_Y(y) = \sum_x p_{XY}(x,y)$
- $Z = g(X,Y)$, $p_Z(z) = \sum_{\{(x,y)|g(x,y)=z\}} p_{XY}(x,y)$
  $\mathrm{E}[Z] = \mathrm{E}[g(X,Y)] = \sum_x \sum_y g(x,y) p_{XY}(x,y)$
  $\mathrm{E}[aX + bY + c] = a\mathrm{E}[X] + b\mathrm{E}[Y] + c$
- Généralisation à n variables aléatoires

36

V.A. conditionnées

- V.A. conditionnée par un événement $A, P(A) \neq 0$

$$p_{X|A}(x) = P(\{X=x\}|A) = \frac{P(\{X=x\} \cap A)}{P(A)}$$

  $\bigcap_x (\{X=x\} \cap A) = \emptyset$, $\bigcup_x (\{X=x\} \cap A) = A$
  $P(A) = \sum_x P(\{X=x\} \cap A) \Rightarrow \sum_x p_{X|A}(x) = 1$
- V.A. conditionnée par une autre V.A.

$$p_{X|Y}(x|y) = P(\{X=x\}|\underbrace{\{Y=y\}}_{p_Y(y)\neq 0}) = \frac{P(\{X=x\} \cap \{Y=y\})}{P(\{Y=y\})} = \frac{p_{XY}(x,y)}{p_Y(y)}$$

  $p_Y(y) = \sum_x p_{XY}(x,y) \Rightarrow \sum_x p_{X|Y}(x|y) = 1$
  Approche séquentielle :
  $p_{XY}(x,y) = p_X(x) p_{Y|X}(y|x) = p_Y(y) p_{X|Y}(x|y)$

37

Espérance conditionnelle

- $\mathrm{E}[X|A] \triangleq \sum_x x p_{X|A}(x)$
- $\mathrm{E}[g(X)|A] = \sum_x g(x) p_{X|A}(x)$
- $\mathrm{E}[X|\{Y=y\}] \triangleq \sum_x x p_{X|Y}(x|y)$
- $\mathrm{E}[X] = \sum_y p_Y(y)\mathrm{E}[X|\{Y=y\}]$ (théorème d'espérance totale)
- $A_1, \ldots, A_n$: partition de Ω, $P(A_i) \neq 0$
  $\mathrm{E}[X] = \sum_{i=1}^n P(A_i)\mathrm{E}[X|A_i]$
- $A_1 \cap B, \ldots, A_n \cap B$: partition de B, $P(A_i \cap B) \neq 0$
  $\mathrm{E}[X|B] = \sum_{i=1}^n P(A_i|B)\mathrm{E}[X|A_i \cap B]$

38

Indépendance

- Entre une V.A. X et un événement A :
  - $P(\{X=x\} \cap A) = P(\{X=x\})P(A) = p_X(x)P(A)\ ,\ \forall x$
  - si $P(A) \neq 0\ ,\ p_{X|A}(x) = p_X(x)\ ,\ \forall x$
- Entre deux V.A. X et Y :
  - $\boxed{p_{XY}(x,y)} = P(\{X=x\} \cap \{Y=y\}) = P(\{X=x\})P(\{Y=y\})$
    $\boxed{= p_X(x)p_Y(y)\ ,\ \forall x,y}$
  - $p_{X|Y}(x,y) = p_X(x)\ ,\ \forall x$ et $\forall y, p_Y(y) \neq 0$
  - $\boxed{\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]}\ ,\ \boxed{\mathrm{var}[X+Y] = \mathrm{var}[X] + \mathrm{var}[Y]}$
- Entre n V.A. $X_1, \ldots, X_n$
  - $p_{X_1 \ldots X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1) \ldots p_{X_n}(x_n)\ ,\ \forall x_1, \ldots, x_n$
  - $\mathrm{E}[X_1 \ldots X_n] = \mathrm{E}[X_1] \ldots \mathrm{E}[X_n]$
  - $\mathrm{var}[X_1 + \ldots + X_n] = \mathrm{var}[X_1] + \ldots + \mathrm{var}[X_n]$

39

Deux variables aléatoires indépendantes

Cliquer sur le logo pour télécharger le script R !

- R en 5 points
  1. `x <- 5` équivalent à `x = 5`
     (les deux sont équivalents dans les versions récentes !)
  2. `x = c(1, 2, 3)` : $x = (1, 2, 3)$
     (fonction de concantenation ; on la trouve partout !)
  3. On utilise le point "." dans les noms à la place de "_"
     (`esp.x.fois.y` n'est qu'un nom de variable !)
  4. Obtenir de l'aide sur une commande :
     `?nom_de_la_commande` ou
     `help(nom_de_la_commande)`
  5. Un document très utile :
     Short-refcard.pdf (4 pages)
     (plus la documentation proposée en bibliographie)

40

Fonction de répartition

- $$\boxed{F_X(x) \triangleq P(\{X \leq x\}) = \sum_{x' \leq x} p_X(x')}$$
- Classement par ordre : $x_{(1)} < x_{(2)} < \ldots < x_{(m)}$

$$F_X(x_{(k)}) = P(\{X \leq x_{(k)}\}) = \sum_{i=1}^{k} p_X(x_{(i)})$$

- Propriétés :
  - $F_X(x)$: définie sur $\mathbb{R}$; continue à droite
  - $F_X(x_{(k)}) - F_X(x_{(k)}^-) = p_X(x_{(k)})$
  - Monotone croissante (au sens large) :
    si $x_1 < x_2$, $F_X(x_1) \leq F_X(x_2)$
  - $\lim_{x \to -\infty} F_X(x) = 0$
  - $\lim_{x \to +\infty} F_X(x) = 1$
  - $F_X(x_2) - F_X(x_1) = P(\{x_1 < X \leq x_2\})$

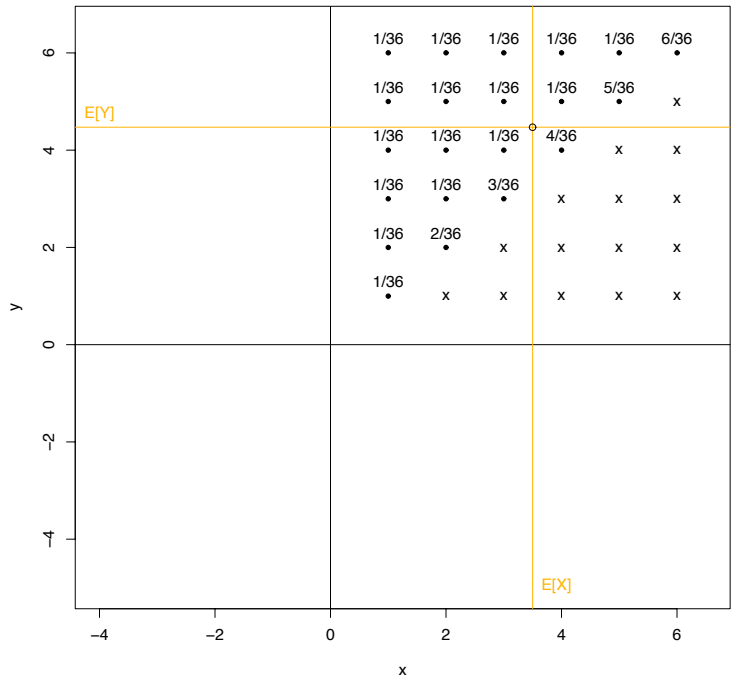
41

Relation linéaire entre deux v.a. ?

- X et Y associées à la même expérience aléatoire
- Est-ce que $Y = aX + b$?
  Si oui :
  - $y = ax + b$ (les valeurs des v.a.)
  - $\mathrm{E}[Y] = a\mathrm{E}[X] + b$ (les espérances des v.a.)
- Comment « mesurer » la dépendance linéaire ?
- Exemple :
  Expérience aléatoire : lancer deux dés
  X : la valeur du premier dé
  Y : la valeur maximale des deux dés

42

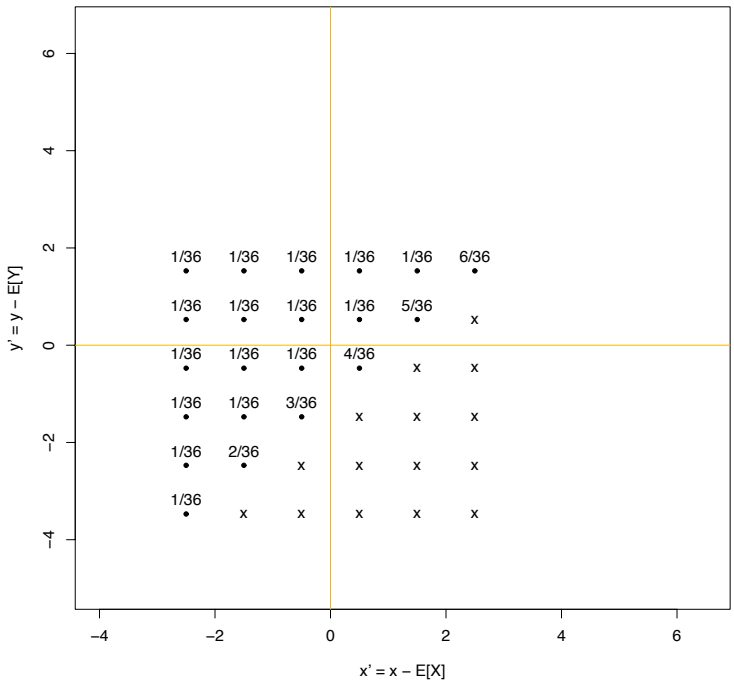
Relation linéaire ? (exploration graphique)



$Y \stackrel{?}{=} aX + b$
Si oui :
$y = ax + b$
$\mathrm{E}[Y] = a\mathrm{E}[X] + b$

43

Relation linéaire ? (exploration graphique 2)



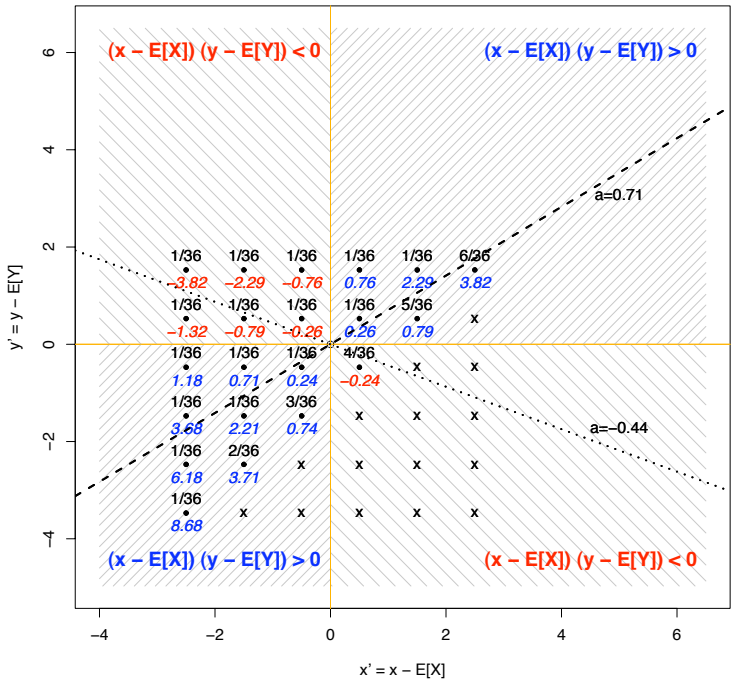
$Y \stackrel{?}{=} aX + b$
Si oui :
$y = ax + b$
$\mathrm{E}[Y] = a\mathrm{E}[X] + b$

$X' = X - \mathrm{E}[X]$
$Y' = Y - \mathrm{E}[Y]$
$Y' = aX'$

$x'y' = ?$

44

Relation linéaire ? (conclusion)



$Y \stackrel{?}{=} aX + b$
Si oui :
$y = ax + b$
$\mathrm{E}[Y] = a\mathrm{E}[X] + b$

$X' = X - \mathrm{E}[X]$
$Y' = Y - \mathrm{E}[Y]$
$Y' = aX'$

$x'y' = ?$

$\mathrm{E}[X'Y'] \approx$

45

Covariance / coefficient de corrélation linéaire

- Covariance

$$\boxed{\mathrm{cov}[X,Y] \triangleq \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]} = \mathrm{R}_{XY} - \mu_X\mu_Y$$

$$\mathrm{cov}[X,X] = \mathrm{E}[(X-\mathrm{E}[X])(X-\mathrm{E}[X])] = \mathrm{E}\big[X^2\big] - \mathrm{E}[X]^2 = \mathrm{var}[X]$$

$\mathrm{cov}[X,Y] \gg 0$ ou $\ll 0$: relation linéaire entre X et Y
Quelles sont les valeurs extrêmes de $\mathrm{cov}[X,Y]$?
Si $Y = aX + b \Rightarrow \mathrm{cov}[X,Y] = \ldots = a\mathrm{var}[X] = a\sigma_X^2 \overset{\sigma_Y = |a|\sigma_X}{=} \mathrm{sign}(a)\sigma_X\sigma_Y$
$-\sigma_X\sigma_Y \leq \mathrm{cov}[X,Y] \leq +\sigma_X\sigma_Y$

- Coefficient de corrélation linéaire

$$\boxed{\rho \triangleq \frac{\mathrm{cov}[X,Y]}{\sigma_X\sigma_Y} \qquad -1 \leq \rho \leq +1 \qquad \rho = \mathrm{sign}(a) \text{ si } Y = aX+b}$$

46

Indépendance / corrélation

- Coefficient de corrélation linéaire

$$\rho = \frac{\mathrm{cov}[X,Y]}{\sigma_X\sigma_Y} = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]}{\sigma_X\sigma_Y}$$

- Corrélation entre X et Y

$$\boxed{R_{XY} \triangleq \mathrm{E}[XY]} = \sum_x \sum_y xy p_{XY}(x,y)$$

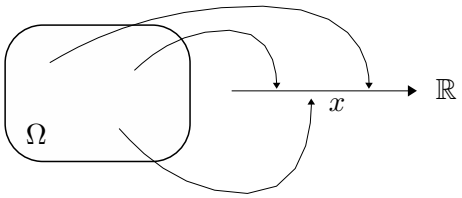
- $\mathrm{E}[XY] \overset{\text{ind.}}{=} \mathrm{E}[X]\,\mathrm{E}[Y] \Rightarrow \rho = 0$
- Si X et Y indépendantes $\Rightarrow$ décorrélées
- Attention (1) : l'inverse n'est pas nécessairement vraie !
  (examiner, p.ex., X et $Y = |X|$ dans le cas où $\mathrm{E}[X] = 0$)
- Attention (2) : « corrélées / décorrélées » se réfère à ρ
  ($\neq$ corrélation R_{XY} !)

47

Variables Aléatoires Continues 48

Définition

- Associer *une valeur réelle* à chaque issue d'une expérience aléatoire
- Nombre de valeurs possibles : infini (non dénombrable)

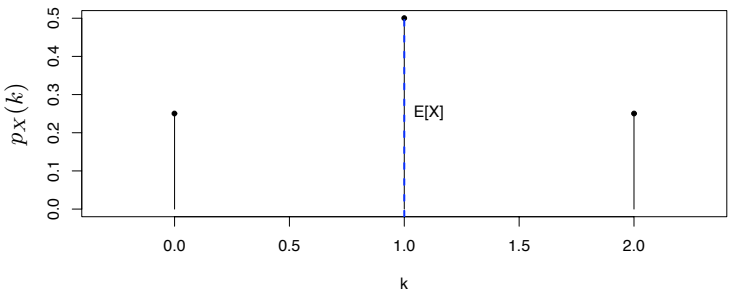


- Exemples :
  - la vitesse d'une voiture
  - le temps entre l'arrivée de deux clients
  - la « position » d'un électron
  - l'énergie d'une particule

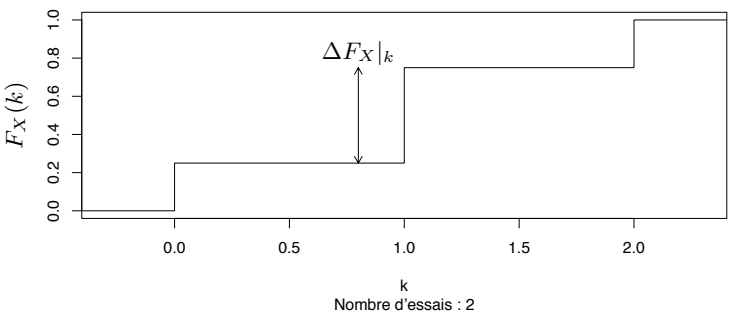
49

Fonction de répartition : v.a.d. vers v.a.c.

Fonction de probabilité, $p_X(k) = P(\{X = k\})$



Fonction de répartition, $F_X(k) = P(\{X \leq k\})$



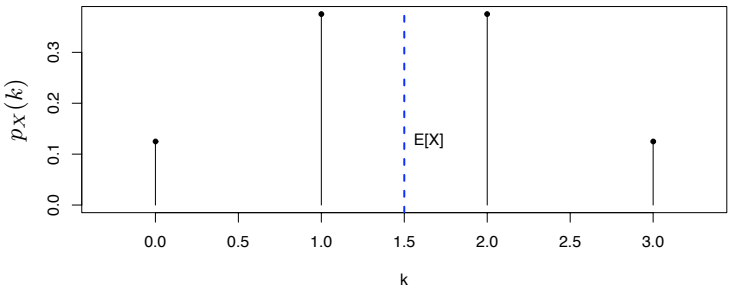
V.A.D.

- $F_X(b) - F_X(a) = P(\{a < X \leq b\})$
- $F_X(k) - F_X(k^-) \triangleq \Delta F_X|_k = P(\{X = k\}) = p_X(k)$

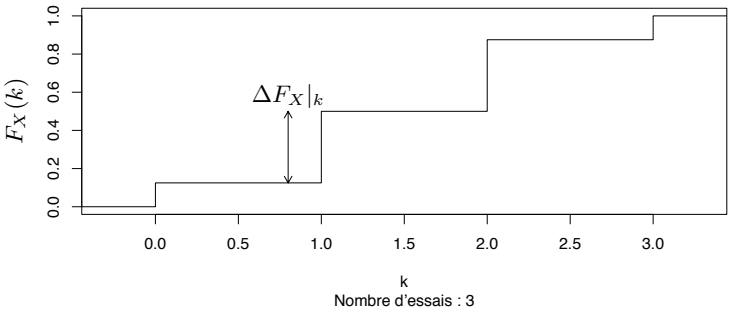
50

Fonction de répartition : v.a.d. vers v.a.c.

Fonction de probabilité, $p_X(k) = P(\{X = k\})$



Fonction de répartition, $F_X(k) = P(\{X \leq k\})$



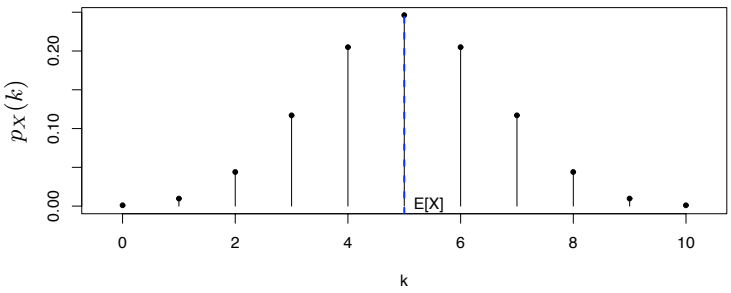
V.A.D.

- $F_X(b) - F_X(a) = P(\{a < X \leq b\})$
- $F_X(k) - F_X(k^-) \triangleq \Delta F_X|_k = P(\{X = k\}) = p_X(k)$

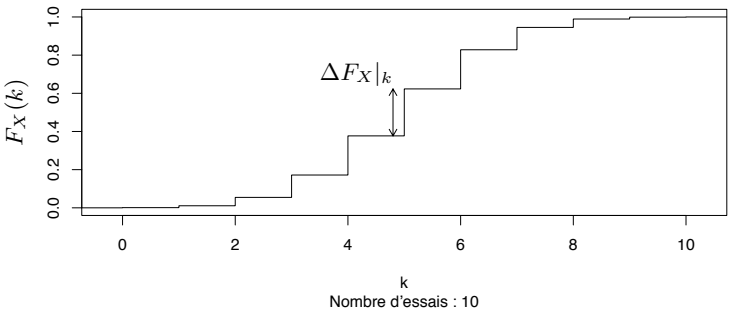
51

Fonction de répartition : v.a.d. vers v.a.c.

Fonction de probabilité, $p_X(k) = P(\{X = k\})$



Fonction de répartition, $F_X(k) = P(\{X \leq k\})$



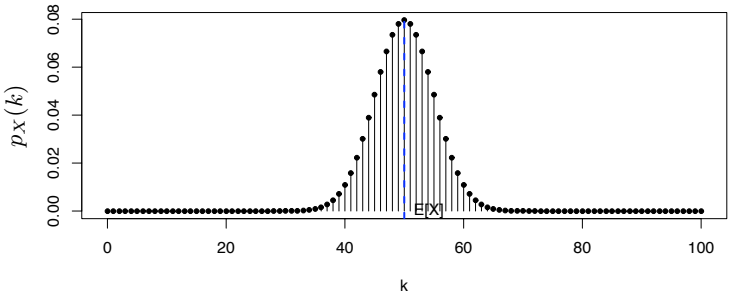
V.A.D.

- $F_X(b) - F_X(a) = P(\{a < X \leq b\})$
- $F_X(k) - F_X(k^-) \triangleq \Delta F_X|_k = P(\{X = k\}) = p_X(k)$

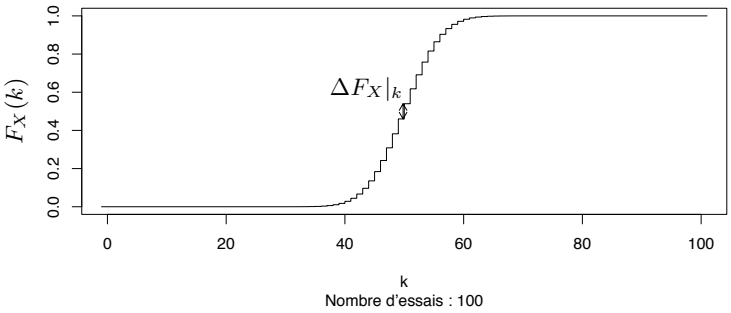
52

Fonction de répartition : v.a.d. vers v.a.c.

Fonction de probabilité, $p_X(k) = P(\{X = k\})$



Fonction de répartition, $F_X(k) = P(\{X \leq k\})$



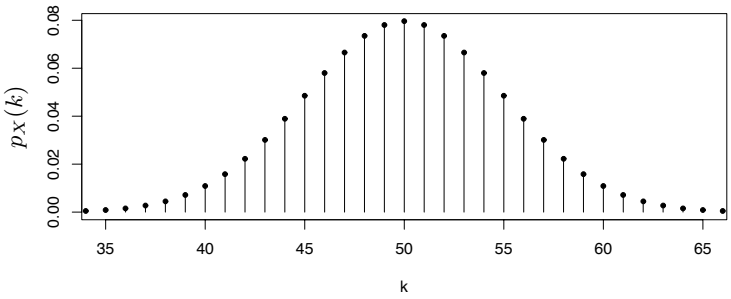
V.A.D.

- $F_X(b) - F_X(a) = P(\{a < X \leq b\})$
- $F_X(k) - F_X(k^-) \triangleq \Delta F_X|_k = P(\{X = k\}) = p_X(k)$

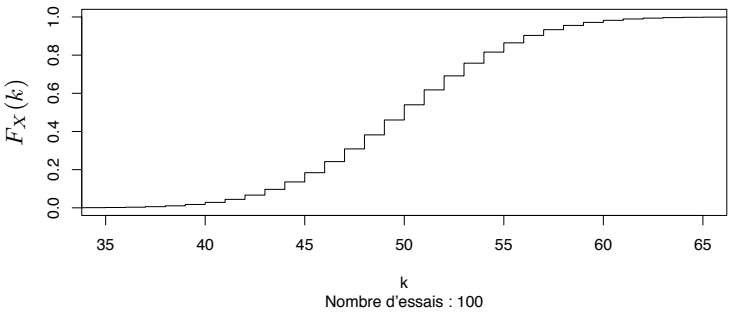
53

Fonction de répartition : v.a.d. vers v.a.c.

Fonction de probabilité, $p_X(k) = P(\{X = k\})$



Fonction de répartition, $F_X(k) = P(\{X \leq k\})$



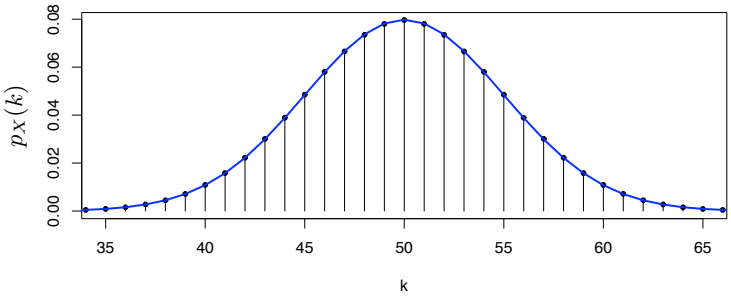
V.A.D.

- $F_X(b) - F_X(a) = P(\{a < X \leq b\})$
- $F_X(k) - F_X(k^-) \triangleq \Delta F_X|_k = P(\{X = k\}) = p_X(k)$

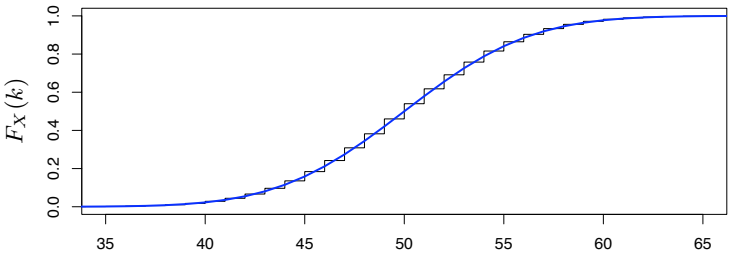
54

Fonction de répartition : v.a.d. vers v.a.c.

Fonction de probabilité, $p_X(k) = P(\{X = k\})$



Fonction de répartition, $F_X(k) = P(\{X \leq k\})$



Nombre d'essais : 100

V.A.D.

- $F_X(b) - F_X(a) = P(\{a < X \leq b\})$
- $F_X(k) - F_X(k^-) \triangleq \Delta F_X|_k = P(\{X = k\}) = p_X(k)$

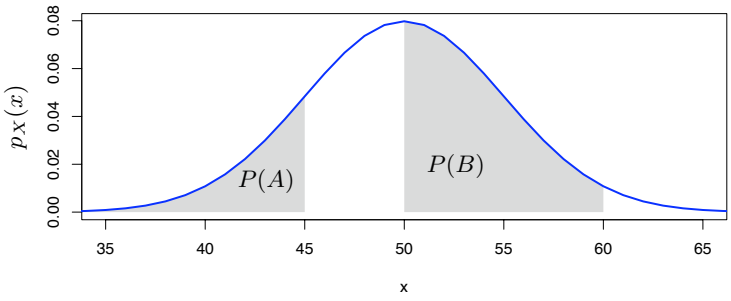
V.A.C.

- $F_X(b) - F_X(a) = P(\{a < X \leq b\})$
- $F_X(x + \mathrm{d}x) - F_X(x) \triangleq \mathrm{d}F_X = P(\{x < X \leq x + \mathrm{d}x\})$
- $\frac{\mathrm{d}F_X}{\mathrm{d}x} = \frac{P(\{x<X\leq x+\mathrm{d}x\})}{\mathrm{d}x} \triangleq p_X(x)$ ddp
- $F_X(x) = \int_{-\infty}^{x} p_X(u)\,\mathrm{d}u$

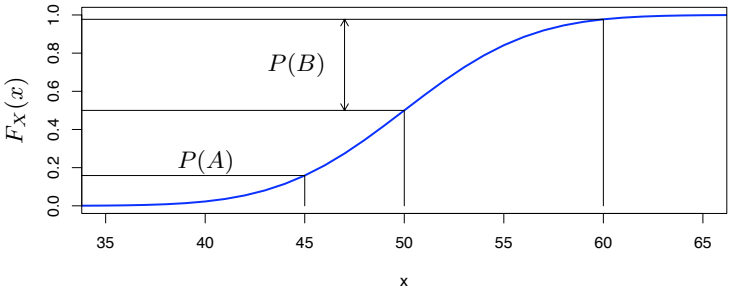
55

Fonction de répartition : v.a.d. vers v.a.c.

ddp $p_X(x) = P(\{x < X \leq x + \mathrm{d}x\})/\mathrm{d}x$



Fonction de répartition, $F_X(x) = P(\{X \leq x\})$



V.A.C.

- $F_X(b) - F_X(a) = P(\{a < X \leq b\})$
- $F_X(x + \mathrm{d}x) - F_X(x) \triangleq \mathrm{d}F_X = P(\{x < X \leq x + \mathrm{d}x\})$
- $\frac{\mathrm{d}F_X}{\mathrm{d}x} = \frac{P(\{x<X\leq x+\mathrm{d}x\})}{\mathrm{d}x} \triangleq p_X(x)$ ddp
- $F_X(x) = \int_{-\infty}^{x} p_X(u)\,\mathrm{d}u$
- $A = \{X \leq 45\}$
- $B = \{50 < X \leq 60\}$

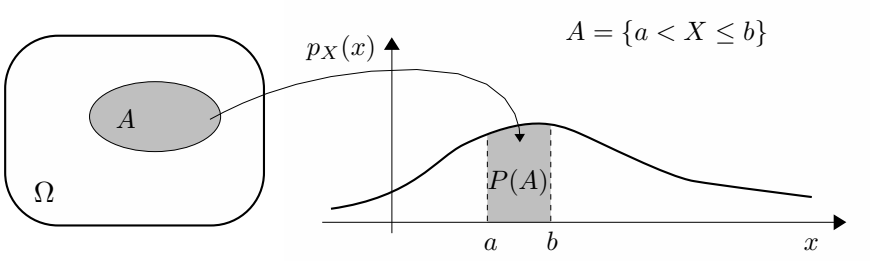
56

Densité de probabilité

□

$$P(\{x < X \leq x + \mathrm{d}x\}) = p_X(x)\,\mathrm{d}x$$

$$\boxed{P(\{a < X \leq b\}) = \int_a^b p_X(x)\,\mathrm{d}x}$$



□ $p_X(x) \geq 0\ ,\ \forall x$
□ $P(\{X = x_0\}) = P(\{x_0 < X \leq x_0\} = \int_{x_0}^{x_0} p_X(x)\,\mathrm{d}x = 0$
□ Normalisation :
$\int_{-\infty}^{+\infty} p_X(x)\,\mathrm{d}x = P(\{-\infty < X < +\infty\}) = P(\Omega) = 1$

57

Fonction de répartition

□

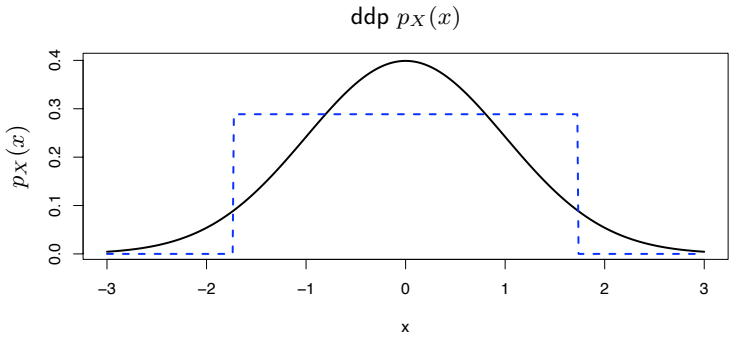
$$\boxed{F_X(x) \triangleq P(\{X \leq x\}) = \int_{-\infty}^{x} p_X(u)\,\mathrm{d}u} \qquad \boxed{p_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}}$$

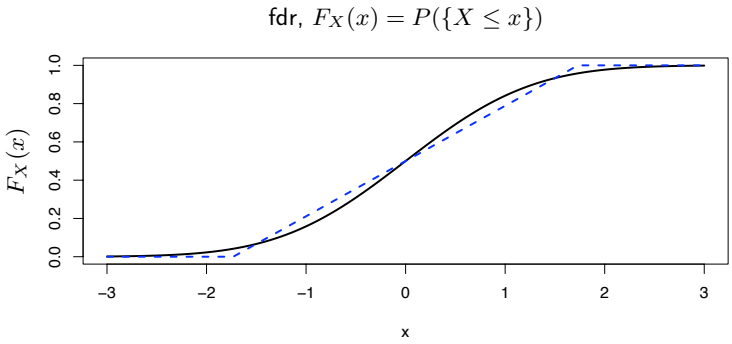
□ Propriétés :

- $F_X(x)$: définie sur $\mathbb{R}$; continue (v.a.c.) / cont. à droite (v.a.d.)
- Monotone croissante (au sens large) :
  si $x_1 < x_2\ ,\ F_X(x_1) \leq F_X(x_2)$
- $\lim_{x\to-\infty} F_X(x) = 0$
- $\lim_{x\to+\infty} F_X(x) = 1$
- $F_X(x_2) - F_X(x_1) = P(\{x_1 < X \leq x_2\})$
- v.a.d. : $F_X(x_{(k)}) - F_X(x_{(k)}^-) = p_X(x_{(k)})$
- v.a.c. : $\mathrm{d}F_X(x) = F_X(x + \mathrm{d}x) - F_X(x) = P(x < X \leq x + \mathrm{d}x)$
- $\frac{\mathrm{d}F_X(x)}{\mathrm{d}x} = \frac{P(x<X\leq x+\mathrm{d}x)}{\mathrm{d}x} \triangleq p_X(x)$

58

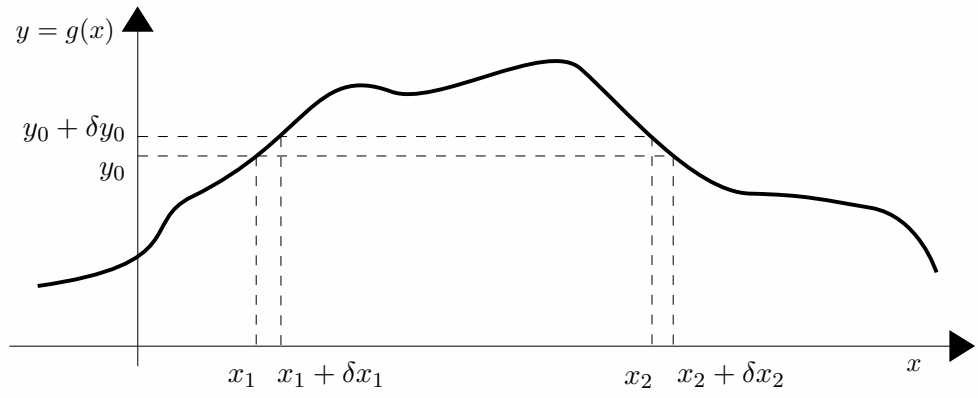
Exemple : v.a. uniforme et v.a. normale





59

Fonction d'une V.A.



- $P(\{y_0 < Y \leq y_0 + \delta y_0\}) = p_Y(y_0) \cdot \delta y_0$
  $= \sum_{\{x_i | g(x_i) = y_0\}} P(\{x_i < X \leq x_i + \delta x_i\})$
  $= \sum_{\{x_i | g(x_i) = y_0\}} p_X(x_i) \cdot \delta x_i$
- $$p_Y(y_0) = \sum_{\{x_i | g(x_i) = y_0\}} p_X(x_i) \frac{1}{\delta y_0 / \delta x_i} = \sum_{\{x_i | g(x_i) = y_0\}} \frac{p_X(x_i)}{|g'(x_i)|}$$

60

Grandeurs statistiques

- Espérance

$$\boxed{\mu_X = \mathrm{E}[X] = \int_{-\infty}^{+\infty} x\, p_X(x)\,\mathrm{d}x}$$

- Espérance de $g(X)$

$$\mu_{g(X)} = \mathrm{E}[g(X)] = \int_{-\infty}^{+\infty} g(x)\, p_X(x)\,\mathrm{d}x$$

- Variance

$$\boxed{\mathrm{var}[X] = \sigma_X^2 = \mathrm{E}\big[(X - \mathrm{E}[X])^2\big] = \mathrm{E}\big[X^2\big] - \mathrm{E}[X]^2}$$

- n-ième moment :

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

- n-ième moment centré :

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

61

Fonction linéaire

-

$$Y = aX + b$$

-

$$\boxed{\mathrm{E}[Y] = a\mathrm{E}[X] + b}$$

$$\boxed{\mathrm{var}[Y] = a^2\mathrm{var}[X]} \qquad \boxed{\sigma_Y = |a|\sigma_X}$$

- $\begin{aligned}\mathrm{E}[Y] &= \mathrm{E}[aX + b] \\ &= \int_{-\infty}^{+\infty} (ax + b)\, p_X(x)\,\mathrm{d}x \\ &= a\int_{-\infty}^{+\infty} x\, p_X(x)\,\mathrm{d}x + b\int_{-\infty}^{+\infty} p_X(x)\,\mathrm{d}x \\ &= a\mathrm{E}[X] + b\end{aligned}$

62

Deux variables aléatoires

- X, Y : V.A. associées à la *même* expérience aléatoire
- Densité de probabilité conjointe $p_{XY}(x, y)$:
- $$P(\{x < X \leq x + \mathrm{d}x\} \cap \{y < Y \leq y + \mathrm{d}y\}) = p_{XY}(x, y)\,\mathrm{d}x\,\mathrm{d}y$$
  $$P(\{a < X \leq b\} \cap \{\, c < Y \leq d\}) = \int_c^d \int_a^b p_{XY}(x, y)\,\mathrm{d}x\,\mathrm{d}y$$
- Densités de probabilité marginales :
  $p_X(x) = \int_{-\infty}^{+\infty} p_{XY}(x, y)\,\mathrm{d}y$, $p_Y(y) = \int_{-\infty}^{+\infty} p_{XY}(x, y)\,\mathrm{d}x$
- $Z = g(X, Y)$
  $\mathrm{E}[Z] = \mathrm{E}[g(X, Y)] = \iint_{-\infty}^{+\infty} g(x, y) p_{XY}(x, y)\,\mathrm{d}x\,\mathrm{d}y$
  $\mathrm{E}[aX + bY + c] = a\mathrm{E}[X] + b\mathrm{E}[Y] + c$
- Généralisation à n variables aléatoires

63

V.A. Conditionnées

- V.A. conditionnée par un événement $A, P(A) \neq 0$
  - ddpc $p_{X|A}(x)$: $P(\{x < X \leq x + \mathrm{d}x\}|A) = p_{X|A}(x)\,\mathrm{d}x$
  - cas spécial : si $A = \{X \in C\}$:
  $$p_{X|\{X \in C\}}(x) = \begin{cases} \frac{p_X(x)}{P(\{X \in C\})} & x \in C \\ 0 & x \notin C \end{cases}$$
- V.A. conditionnée par une V.A.
  -
  $$p_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)}\,, \qquad \forall y \mid p_Y(y) \neq 0$$
  - Approche séquentielle :
  - $p_{XY}(x, y) = p_X(x) p_{Y|X}(y|x) = p_Y(y) p_{X|Y}(x|y)$

64

Espérance conditionnelle

- $\mathrm{E}[X|\{Y=y\}] = \int x\, p_{X|Y}(x|y)\,\mathrm{d}x$
- $\mathrm{E}[X] = \int \mathrm{E}[X|\{Y=y\}]\; p_Y(y)\,\mathrm{d}y$ (théorème d'espérance totale)
- $\mathrm{E}[g(X)|\{Y=y\}] = \int g(x)\; p_{X|Y}(x|y)\,\mathrm{d}x$
- $\mathrm{E}[g(X)] = \int \mathrm{E}[g(X)|\{Y=y\}]\; p_Y(y)\,\mathrm{d}y$
- $\mathrm{E}[g(X,Y)|\{Y=y\}] = \int g(x,y)\; p_{X|Y}(x|y)\,\mathrm{d}x$
- $\mathrm{E}[g(X,Y)] = \int \mathrm{E}[g(X,Y)|\{Y=y\}]\; p_Y(y)\,\mathrm{d}y$

65

Indépendance

- Entre deux V.A. X et Y :
- $\boxed{p_{XY}(x,y) = p_X(x)p_Y(y)\ ,\ \forall x,y}$
  - $p_{X|Y}(x,y) = p_X(x)\ ,\ \forall x$ et $\forall y, p_Y(y) \neq 0$
  - $P(\{X \in A\} \cap \{Y \in B\}) = P(\{X \in A\}) \cdot P(\{Y \in B\})$
  - $\boxed{\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]} \Rightarrow \mathrm{cov}[X,Y] = 0$: v.a. non corrélées
  - $\mathrm{E}[g(X)h(Y)] = \mathrm{E}[g(X)]\,\mathrm{E}[h(Y)]$
  - $\boxed{\mathrm{var}[X+Y] = \mathrm{var}[X] + \mathrm{var}[Y]}$
- Entre n V.A. $X_1, \ldots, X_n$
- $p_{X_1 \ldots X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1) \ldots p_{X_n}(x_n)\ ,\ \forall x_1, \ldots, x_n$
  - $\mathrm{E}[X_1 \ldots X_n] = \mathrm{E}[X_1] \ldots \mathrm{E}[X_n]$
  - $\mathrm{var}[X_1 + \ldots + X_n] = \mathrm{var}[X_1] + \ldots + \mathrm{var}[X_n]$

66

Statistique Descriptive 67

Quelques définitions

- Population statistique : ensemble d'individus à étudier
  - finie
  - infinie
- Individu / unité statistique
- Caractère / variable statistique
  - qualitatif
  - quantitatif (discret / continu)
- Échantillon : sous-ensemble de la population
- Fréquences
  - absolues (effectifs)
  - relatives (proportions)

68

Paramètres statistiques d'un échantillon

- Mesures de tendance centrale (position)
  - Moyenne : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (`mean`)
  - Médiane : partage les valeurs en deux parties (`median`)
  - Quantiles : partagent les valeurs en k parties (`perctl`)
  - Quartiles ($k = 4$) : Q_1, Q_2 (médiane), Q_3 (`quart`)
  - Mode(s) : la (les) valeur(s) avec la plus grande fréquence
- Mesures de dispersion
  - Étendue : $x_{(n)} - x_{(1)}$ (`max` - `min`)
  - Intervalle interquartile (IQR) : $Q_3 - Q_1$ (`iqr`)
  - Variance de l'échantillon : (`variance`)
    $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}$ (attn. si $s/\bar{x} \ll 1$)
  - Écart-type de l'échantillon : s (`stdev`)
  - Écart absolu médian par rapport à la médiane (`mad`)
  - Coefficient de variation : $s/\bar{x}$

69

Exemple : notes TP Élec 2006-2007

- Population : étudiants Élec4, 2006-2007
- Caractère étudié :
  1. option (qualitatif)
  2. moyenne tp (quantitatif)
  3. contrôle final (quantitatif)
- Échantillon : 30 étudiants

70

Statistique Inférentielle : introduction 71

Objectif

Obtenir, à partir de mesures sur une *partie* de la population (échantillon), des informations (de caractère *probabiliste*) sur la *totalité* de celle-ci.

$$\text{Population} \xrightarrow{\text{proba}} \text{Échantillon}$$

$$\text{Échantillon} \xrightarrow{\text{stat. inf.}} \text{Population}$$

72

Échantillonnage : définition

Choisir *au hasard* n individus de la population afin d'étudier un ou plusieurs caractères.

- Deux types d'échantillonnage :
  1. *avec* remplacement de l'individu choisi
     traitement théorique plus simple
  2. *sans* remplacement : échantillonnage exhaustif
     procédure naturelle ou obligatoire (contrôle destructif)
- Population de taille finie + éch. non exhaustif
  $\Rightarrow$ population de taille infinie
- Éch. exhaustif de taille n + Population de taille $N \gg n$
  $\Rightarrow$ échantillonnage non exhaustif

73

Une expérience aléatoire

Choisir *au hasard* un individu de la population. Obtenir une valeur du caractère étudié.

- Valeurs possibles du caractère : celles présentes dans la population
- Probabilité associée : fréquence relative des individus possedant cette valeur dans la population

  À condition que chaque individu ait la même probabilité d'être choisi !

$$\begin{array}{ccccc} \text{population} & \xrightarrow{\text{éch.}} & \text{individu} & \xrightarrow{\text{caract.}} & \text{valeur} \\ \Omega & \xrightarrow{\text{éch.}} & \omega & \xrightarrow{\text{caract.}} & x \end{array}$$

- Expérience aléatoire : choisir au hasard un individu de la population
- Variable aléatoire X associée : le caractère étudié (quantitatif / qualitatif)
- Fonction/densité de probabilité $p_X(x)$: dépend de la population

74

Échantillon : ensemble de variables aléatoires

- « Population $p_X(x)$ » : génère des v.a.
- Observation d'un caractère d'un individu : v.a. X, loi $p_X(x)$
- Échantillonnage de taille n : la même expérience aléatoire répétée n fois !
  ensemble de n v.a. X_i $(i = 1, \ldots, n)$
- Échantillonnage *aléatoire* (non biaisé) : n v.a. *identiques* et *indépendantes* (iid)

$$p_{X_1}(x) = p_{X_2}(x) = \ldots = p_{X_n}(x) = p_X(x)$$

$$p_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) = p_X(x_1) p_X(x_2) \ldots p_X(x_n)$$

  c-à-d : *avec* remplacement + *même probabilité* de choisir chaque individu
- Statistiques : des v.a., fonctions des X_i $(i = 1, \ldots, n)$ d'un échantillon
  (théorie d'échantillonnage : quelles valeurs et quelles probabilités ?)
- Obtenir un échantillon, de taille n :
  ensemble de n valeurs x_i $(i = 1, \ldots, n)$ $\longrightarrow$ Statistique Descriptive !
- Expérience mentale : obtenir une infinité d'échantillons

75

Paramètres statistiques d'un échantillon

- Mesures de tendance centrale (position)
  - Moyenne : $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
  - Médiane : partage les valeurs en deux parties
  - Quantiles : partagent les valeurs en k parties
  - Quartiles $(k = 4)$: Q_1, Q_2 (médiane), Q_3
  - Déciles $(k = 9)$: D_1, D_2, ..., D_5 (médiane), ..., D_9
- Statistiques d'ordre : $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ où $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$
- Mesures de dispersion
  - Étendue : $X_{(n)} - X_{(1)}$
  - Intervalle interquartile (IQR) : $Q_3 - Q_1$
  - Variance de l'échantillon : $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = \frac{n \sum_{i=1}^{n} (X_i)^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)}$ (attn. si $s/\bar{x} \ll 1$)
  - Écart-type de l'échantillon : S
  - Écart absolu médian par rapport à la médiane
  - Coefficient de variation : $S/\bar{X}$

76

Cas spécial : caractère qualitatif (les proportions)

- Étudier un caractère qualitatif à M modalités (réponses possibles)
  - Population : M « types » d'individus ; M fréquences relatives π_j
  - Échantillonnage aléatoire d'un individu :
    v.a.d. X à M valeurs ; probabilités associées π_j $(j = 1, \ldots, M)$
- Autre approche (cas par cas) :
  - Pour chaque modalité du caractère, étudier le nouveau caractère « l'individu présente la modalité j du caractère initial »
  - Réponses possibles : « oui » / « non »
  - Population : 2 « types » d'individus ; fréquences relatives π_j, $1 - \pi_j$
  - Échantillonnage aléatoire d'un individu :
    v.a.d. X à 2 valeurs ($1 =$ « oui », $0 =$ « non ») ;
    probabilités associées π_j, $1 - \pi_j$
  - X : v.a.d. de Bernoulli, de paramètre π_j
  - Échantillon de taille n :
    Moyenne $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \triangleq \hat{P}$ proportion de « oui » dans l'échantillon

77

Statistique inférentielle : feuille de route

Théorie d'échantillonnage : Population $\longrightarrow$ Échantillon
Statistique inférentielle : Échantillon $\longrightarrow$ Population

Échantillon		Population $p_X(x)$
v.a.	valeur	paramètre
une population		
$\bar{X}$ S^2	$m = \bar{x}$ s^2	$\mu_X = \mathrm{E}[X]$ $\sigma_X^2 = \mathrm{var}[X]$
$\hat{P}$	$\hat{p}$	π
deux populations		
$\bar{X}_2 - \bar{X}_1$ S_2^2/S_1^2	$m_2 - m_1 = \bar{x}_2 - \bar{x}_1$ $(s_2/s_1)^2$	$\mu_2 - \mu_1$ $(\sigma_2/\sigma_1)^2$
$\hat{P}_2 - \hat{P}_1$	$\hat{p}_2 - \hat{p}_1$	$\pi_2 - \pi_1$

- Estimer les paramètres de la population
- Calculer des intervalles de confiance
- Formuler des hypothèses et les tester

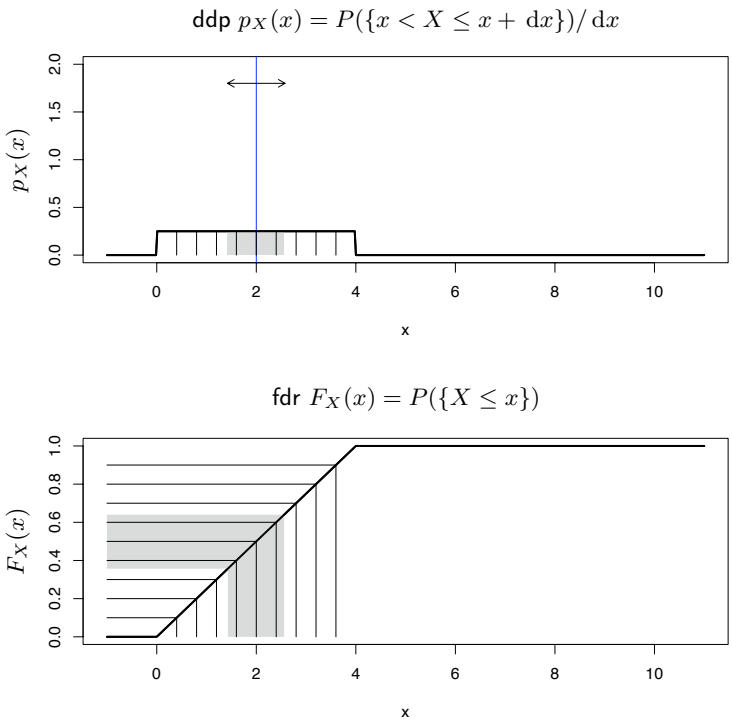
78

Distribution uniforme

- $$p_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{ailleurs} \end{cases}$$
- $$\mathrm{E}[X] = \frac{1}{2}(a+b)$$
- $$\mathrm{var}[X] = \sigma_X^2 = \frac{1}{12}(a-b)^2$$
- $$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$

79

Distribution uniforme



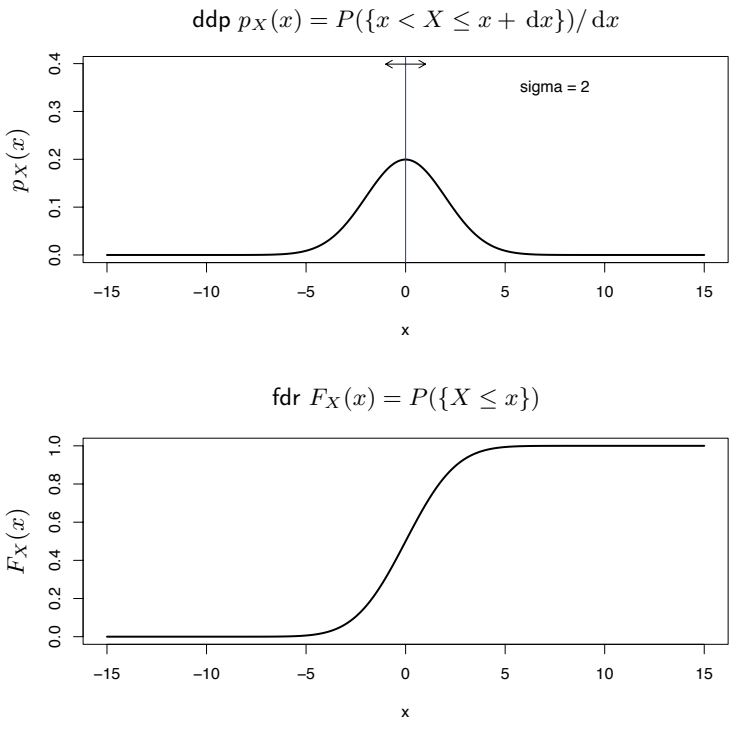
80

Distribution normale (gaussienne)

- $$N(\mu_X, \sigma_X): \quad p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right]$$
- $$\mathrm{E}[X] = \mu_X$$
- $$\mathrm{var}[X] = \sigma_X^2$$
- $$F_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{x'-\mu_X}{\sigma_X}\right)^2\right] \mathrm{d}x'$$

81

Distribution normale



82

Distribution normale standard (centrée réduite)

- $$X = N(\mu_X, \sigma_X) \ : \ P(\{X \leq x = \mu_X + z\sigma_X\}) = F_X(x = \mu_X + z\sigma_X)$$
$$= \frac{1}{\sqrt{2\pi}\sigma_X} \int_{-\infty}^{\mu_X + z\sigma_X} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right] \mathrm{d}x$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{1}{2}u^2\right) \mathrm{d}u = P(\{Z \leq z\}) = F_Z\left(z = \frac{x-\mu_X}{\sigma_X}\right) \triangleq 1 - Q(z)$$
- $\boxed{Z = \dfrac{X - \mu_X}{\sigma_X}}$: normale standard (centrée réduite) $N(0,1)$
- z : exprime l'écart entre x et μ_X en termes (unité de mesure) de σ_X
  toujours sans unité !

83

Distribution normale standard (centrée réduite)

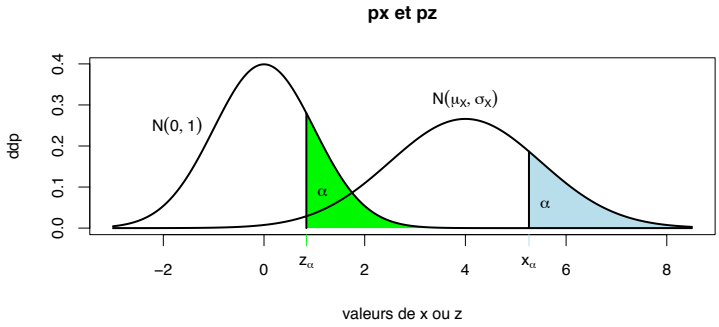
- v.a. centrée réduite : fonction linéaire d'une autre v.a.
- notion générale (pas seulement pour la normale !)

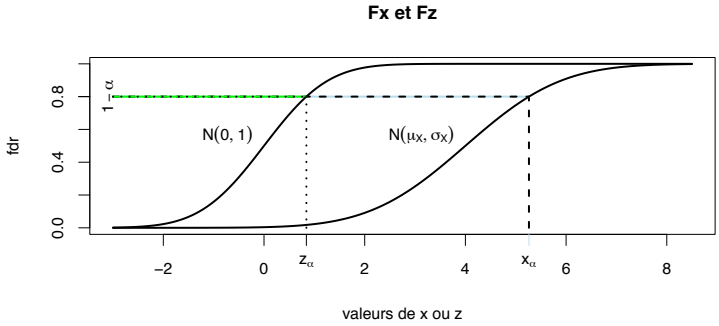
	de X vers Z		de Z vers X	
v.a.	X	$Z = (X - \mu_X)/\sigma_X$	Z	$X = \mu_X + Z\sigma_X$
valeur	x	$z = (x - \mu_X)/\sigma_X$	z	$x = \mu_X + z\sigma_X$
esp.	μ_X	0	0	μ_X
var.	σ_X^2	1	1	σ_X^2
ddp	$p_X(x)$	$\sigma_X p_X(\mu_X + z\sigma_X)$	$p_Z(z)$	$\frac{1}{\sigma_X} p_Z\left(\frac{x-\mu_X}{\sigma_X}\right)$
fdr	$F_X(x) = F_Z\left(\frac{x-\mu_X}{\sigma_X}\right)$		$F_Z(z) = F_X(\mu_X + z\sigma_X)$	

- On peut calculer des probabilités aussi bien en X qu'en Z !

84

Distribution normale standard (centrée réduite)



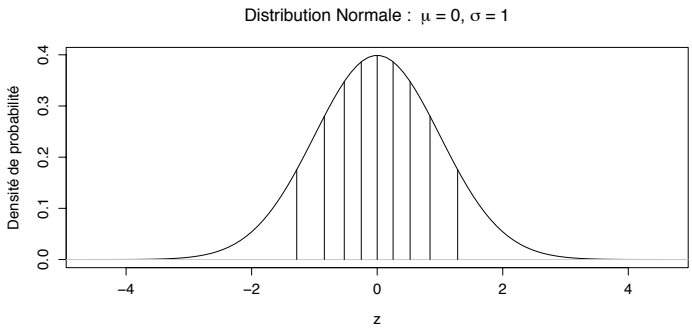


- $X : N(4, 1.5)$
- $Z : N(0, 1)$
- $X = \mu_X + Z\sigma_X$
- « Valeur critique » x_α : $P(\{X > x_\alpha\}) = \alpha$
- $\boxed{F_X(x_\alpha) = 1 - \alpha}$
- « Valeur critique » z_α : $P(\{Z > z_\alpha\}) = \alpha$
- $\boxed{F_Z(z_\alpha) = 1 - \alpha}$
-

$$\boxed{x_\alpha = \mu_X + z_\alpha \sigma_X}$$

85

Distribution normale standard (centrée réduite)



86

Propriétés de la loi normale

1. Deux gaussiennes décoréllées sont indépendantes (l'exception !)
   - □ X_1, X_2 conjointement normales : ddp conjointe $p_{X_1X_2}(x_1, x_2)$:
     $$p_{X_1X_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - \right.$$
     $$\left. -2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{1}{2(1-\rho^2)}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right]$$
   - □ ddp marginales : $X_1 = N(\mu_1, \sigma_1)$ et $X_2 = N(\mu_2, \sigma_2)$
   - □ coefficient de corrélation linéaire : ρ
   - □ $\rho = 0 \Longrightarrow p_{X_1X_2}(x_1, x_2) = p_{X_1}(x_1)p_{X_2}(x_2)$
2. La somme de gaussiennes indépendantes est une gaussienne
   - □ $X_1, X_2, \ldots, X_n$ normales $N(\mu_i, \sigma_i)$, indépendantes
   - □ $X = a_1X_1 + a_2X_2 + \ldots + a_nX_n = \sum_{i=1}^n a_iX_i$
   - □ $\mu_X = a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n = \sum_{i=1}^n a_i\mu_i$
   - □ $\sigma_X^2 \overset{\text{ind}}{=} a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots + a_n^2\sigma_n^2 = \sum_{i=1}^n a_i^2\sigma_i^2$
   - □ $\boxed{X : N(\mu_X, \sigma_X)}$

Somme de deux v.a. indépendantes

- □ X_1, X_2 : v.a. indépendantes (pas nécessairement identiques)
- □ $X = X_1 + X_2$: nouvelle v.a.
- □ Comment trouver $p_X(x)$ à partir de $p_{X_1}(x_1)$ et $p_{X_2}(x_2)$?

1. Cas v.a.d. :
   $$p_X(x) = P(\{X = x\}) \overset{\text{prob. tot.}}{=} \sum_{x_1} P(\{X_1 = x_1\})P(\{X_2 = x - x_1 | X_1 = x_1\})$$
   $$\overset{\text{ind}}{=} \sum_{x_1} P(\{X_1 = x_1\})P(\{X_2 = x - x_1\})$$
   $$= \sum_{x_1} p_{X_1}(x_1)p_{X_2}(x - x_1) = p_{X_1} \star p_{X_2}$$
2. Cas v.a.c. :
   $$p_X(x) = \text{sans démonstration}$$
   $$= \int_{x'} p_{X_1}(x')p_{X_2}(x - x')\,\mathrm{d}x' = p_{X_1} \star p_{X_2}$$

$$\boxed{X = X_1 + X_2 \overset{\text{ind}}{\Longrightarrow} p_X = p_{X_1} \star p_{X_2}}$$

[Théorème limite central]

- $X_1, X_2, \ldots, X_n$: série de v.a. indépendantes
- $p_{X_1}(x) = \ldots = p_{X_n}(x) = p_X(x)$ (même distribution)
- $\mathrm{E}[X_1] = \ldots = \mathrm{E}[X_n] = \mu_X$, $\sigma_{X_1} = \ldots = \sigma_{X_n} = \sigma_X$
-

$$S_n = X_1 + X_2 + \ldots + X_n \ , \ \mathrm{E}[S_n] = n\mu_X \ , \ \sigma^2_{S_n} \overset{\text{ind}}{=} n\sigma^2_X$$

$$\boxed{Z_n = \frac{S_n - \mu_{S_n}}{\sigma_{S_n}}} = \frac{X_1 + X_2 + \ldots + X_n - n\mu_X}{\sqrt{n}\sigma_X} \ , \ \boxed{\mathrm{E}[Z_n] = 0} \ , \ \boxed{\sigma^2_{Z_n} = 1}$$

- TLC :

$$\lim_{n \longrightarrow \infty} P\left(\{Z_n \leq z\}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{1}{2}u^2\right) \mathrm{d}u$$

- TLC : $\boxed{n \to \infty \ : \ Z_n \to N(0,1)}$, $S_n \to N(n\mu_X, \sqrt{n}\sigma_X)$, $\frac{S_n}{n} \to N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$

89

Théorie d'échantillonnage – un échantillon 90

Distribution de la moyenne

- Échantillon aléatoire de taille n ; moyenne $\bar{X}$
- Population normale $N(\mu, \sigma)$
  - $\bar{X}$: normale (combinaison linéaire de v.a. normales)
  - $\mu_{\bar{X}} = \mu$
  - $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ (σ connu)
- Population non normale (σ connu)
  - $n > 30$: $\bar{X} = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ (tlc)
  - $n < 30$: $\bar{X} = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ si $p_X(x)$ « presque » normale
- Presque toujours : $\boxed{\bar{X} = N(\mu, \sigma/\sqrt{n})}$
  - $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \rightarrow N(0,1)$
  - $P(Z > z_\alpha) = \alpha$ (définition de z_α « valeur critique »)
  - $P(Z < -z_\alpha) = \alpha$ (symétrie de la normale)

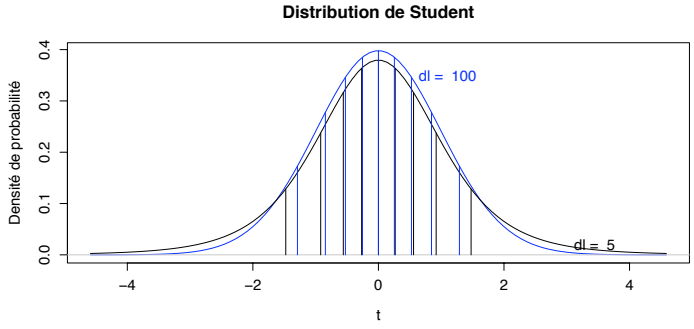
91

Distribution de la moyenne ; σ_X inconnue

- $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to N(0,1)$
- $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{V/(n-1)}} = \frac{Z}{\sqrt{V/\nu}}$
- $V = \frac{(n-1)S^2}{\sigma^2}$: loi du χ^2 à $\nu = n-1$ d.l.
- Condition : population normale
- Z, V indépendantes
- $\boxed{T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \text{ : loi de Student à } \nu = n-1 \text{ d.l.}}$
- $\mathrm{E}[T] = 0$
- $\sigma_T^2 = \frac{\nu}{\nu-2} > 1$ (non définie pour $\nu \leq 2$)
- $P(T > t_\alpha) = \alpha$ (définition de t_α, valeur critique)
- $P(T < -t_\alpha) = \alpha$ (symétrie de la loi t)
- $n \geq 30$: $s \to \sigma$ donc $T \to Z$
- "Student" : W.S. Gosset, 1908

92

Distribution de Student



$\mathrm{E}[T] = 0 \quad , \quad \sigma_T^2 = \frac{\nu}{\nu-2} > 1$ (non définie pour $\nu \leq 2$)

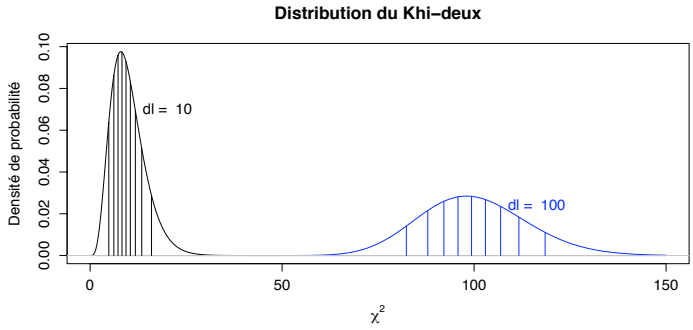
93

Distribution de la variance

- Échantillon aléatoire de taille n ; variance S^2
  - Condition : population normale $N(\mu, \sigma)$
  - $\boxed{X^2 = \dfrac{(n-1)S^2}{\sigma^2}} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
  - $\boxed{X^2 \text{ : v.a. loi du } \chi^2 \text{ à } \nu = n-1 \text{ degrés de liberté (d.l.)}}$
  - $X^2 > 0$
  - $\mathrm{E}\left[X^2\right] = n-1 \longrightarrow \boxed{\mathrm{E}\left[S^2\right] = \sigma^2}$
  - $\sigma^2_{X^2} = 2(n-1) \longrightarrow \sigma^2_{S^2} = 2\sigma^4/(n-1)$
  - $P(X^2 > \chi^2_\alpha(\nu)) = \alpha$ (définition de $\chi^2_\alpha(\nu)$, valeur critique)

94

Distribution du χ^2



$\mathrm{E}\left[X^2\right] = n-1 \quad , \quad \sigma^2_{X^2} = 2(n-1)$

95

Distribution de la proportion

- Population
  - π : proportion d'individus possédant un caractère qualitatif ($\pi \neq 3.14$!)
- Échantillon aléatoire de taille n
  - n v.a. X_i ; $x_i \in \{0, 1\}$: Bernoulli indépendantes, de paramètre π
  - $\sum_{i=1}^{n} X_i$: nombre d'individus possédant le caractère (fréquence)
  - $\boxed{\hat{P} = \frac{1}{n} \sum_{i=1}^{n} X_i \text{ : proportion d'individus (fréquence relative)}}$
- Conditions :
  - $n > 30$ (grand échantillon : théorème limite central)
  - $n\hat{p} \geq 5$ (fréquence de présence du caractère)
  - $n(1 - \hat{p}) = n - n\hat{p} \geq 5$ (fréquence d'absence du caractère)
  - ni $\hat{p} \approx 0$, ni $\hat{p} \approx 1$
- Distribution :
  - $\mu_{\hat{P}} = (n\mu_X)/n = \mu_X = \pi \quad , \quad \sigma^2_{\hat{P}} \overset{\text{ind}}{=} (n\sigma^2_X)/n^2 = \pi(1 - \pi)/n$
  - $\boxed{\hat{P} \text{ : normale } N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)} \rightarrow Z$: normale $N(0, 1)$

96

Théorie d'échantillonnage – deux échantillons 97

Distribution de la différence des moyennes

- Conditions : σ_1, σ_2 connus et
  - populations normales $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$ ou
  - $n_1 > 30$ et $n_2 > 30$, ou
  - populations « presque » normales
- Échantillons aléatoires et indépendants de tailles n_1, n_2 ; moyennes $\bar{X}_1, \bar{X}_2$
  - $\boxed{\bar{X}_1 - \bar{X}_2}$: normale
  - $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$
  - $\sigma^2_{\bar{X}_1 - \bar{X}_2} \overset{\text{ind}}{=} \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
- D'autres cas à examiner ultérieurement...

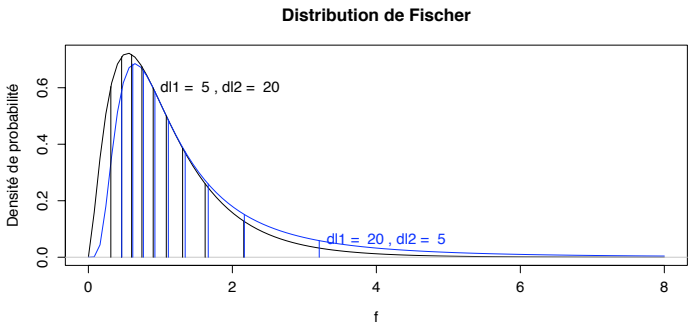
98

Distribution du rapport des variances

- Échantillons aléatoires et indépendants de tailles n_1, n_2
- Provenant de populations normales de variances σ_1^2, σ_2^2
- Variances des échantillons : S_1^2, S_2^2
- $\boxed{F = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}} = \dfrac{V_1/\nu_1}{V_2/\nu_2}$
- $V_i = \dfrac{(n_i-1)S_i^2}{\sigma_i^2}$: v.a. indépendantes, loi du χ^2 à $\nu_i = n_i - 1$ d.l.
- $\boxed{F}$: loi de Fisher (1924) - Snedecor (1934) avec ν_1 et ν_2 d.l.
- $F \geq 0$
- $\mathrm{E}[F] = \dfrac{\nu_2}{\nu_2-2}$ $(\nu_2 > 2)$
- $\sigma_F^2 = \dfrac{\nu_2^2(2\nu_1+2\nu_2-4)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ $(\nu_2 > 4)$
- $P(F > f_\alpha(\nu_1,\nu_2)) = \alpha$ (définition de $f_\alpha(\nu_1,\nu_2)$, v.c.)
- $\boxed{f_\alpha(\nu_1,\nu_2) = \dfrac{1}{f_{1-\alpha}(\nu_2,\nu_1)}}$ (propriété de la loi F)

99

Distribution de Fisher



$f_\alpha(\nu_1,\nu_2) = 1/f_{1-\alpha}(\nu_2,\nu_1)$

100

Estimation – intervalles de confiance 101

Définitions

- Estimation ponctuelle
  - Paramètre à estimer : θ
  - Estimateur : v.a. $\hat{\Theta}$
  - Estimateur non biaisé : $\mathrm{E}[\hat{\Theta}] = \theta$
  - Biais $= \mathrm{E}[\hat{\Theta}] - \theta$
  - Estimateur efficace : sans biais ; de faible variance
  - Estimateur efficace : minimiser l'erreur quadratique moyenne
    $\mathrm{E}\left[(\hat{\Theta} - \theta)^2\right] = \sigma^2_{\hat{\Theta}} + (\text{biais})^2$
  - Estimateur convergent : $n \to \infty \quad : \quad \mathrm{E}[\hat{\Theta}] = \theta$ et $\mathrm{var}[\hat{\Theta}] = 0$
- Estimation par intervalle de confiance
  - v.a. $\hat{\Theta}_L$, $\hat{\Theta}_H$: estimateurs ponctuels
  - $P(\hat{\Theta}_L < \theta < \hat{\Theta}_H) = 1 - \alpha$
  - $\boxed{\hat{\theta}_L < \theta < \hat{\theta}_H}$: intervalle de confiance
  - $1 - \alpha$: niveau de confiance

102

Estimation de la moyenne (1/3)

- Variance σ^2 connue
- $\bar{X}$: normale $N(\mu, \sigma/\sqrt{n})$
- $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$: normale $N(0, 1)$
- $\bar{X}$ estimateur non biaisé et convergent de μ
- $P(Z > z_{\alpha/2}) = \alpha/2$ (définition de $z_{\alpha/2}$)
- $P(Z < -z_{\alpha/2}) = \alpha/2$ (symétrie de la normale)
- $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$
- $P(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha$
- $P(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$
- $P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$
- $\hat{\Theta}_L = \bar{X} - z_{\alpha/2}\sigma_{\bar{X}}$, $\hat{\Theta}_H = \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}$
- $1 - \alpha = 0.95$, $z_{\alpha/2}$ =`qnorm(0.025,mean=0,sd=1,lower.tail=FALSE)`$= 1.96$
- $1 - \alpha = 0.99$, $z_{\alpha/2}$ =`qnorm(0.005,mean=0,sd=1,lower.tail=FALSE)`$= 2.56$

103

Estimation de la moyenne (2/3) : taille de l'échantillon

- $P(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$
- $P(|\bar{X} - \mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$
- $e = |\bar{X} - \mu|$: erreur
- $e_{\text{max}} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$: marge d'erreur à $1-\alpha$
- $n_{\text{min}} = \left(\frac{z_{\alpha/2}\sigma}{e_{\text{max}}}\right)^2$: taille d'échantillon minimale
- $\bar{X} - e_{\text{max}} < \mu < \bar{X} + e_{\text{max}}$ à $1-\alpha$
- Cas particulier : échantillonnage d'une population finie, sans remplacement
  - Population de taille N
  - $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \overset{N\gg 1}{\approx} \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N}} = \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$
  - $n_{\text{min}} = \frac{N z_{\alpha/2}^2 \sigma^2}{N e_{\text{max}}^2 + z_{\alpha/2}^2 \sigma^2}$: taille d'échantillon minimale

104

Estimation de la moyenne (3/3)

- Variance σ^2 inconnue
- Population normale
- $T = (\bar{X} - \mu)/(S/\sqrt{n})$: Student à $n-1$ d.l.
- $P(T > t_{\alpha/2}) = \alpha/2$ (définition de $t_{\alpha/2}$)
- $P(T < -t_{\alpha/2}) = \alpha/2$ (symétrie de la loi t)
- $P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1-\alpha$
- $P(-t_{\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}) = 1-\alpha$
- $P(-t_{\alpha/2}\frac{S}{\sqrt{n}} < \bar{X} - \mu < t_{\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$
- $P(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$
- $\hat{\Theta}_L = \bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}$, $\hat{\Theta}_H = \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}$
- $1-\alpha = 0.95$, $t_{\alpha/2}$ =`qt(0.025,df=29,lower.tail=FALSE)`$= 2.05$
- $1-\alpha = 0.99$, $t_{\alpha/2}$ =`qt(0.005,df=29,lower.tail=FALSE)`$= 2.76$
- Rappel : $n \geq 30$, $T \to Z$
- T : petits échantillons !

105

Estimation de la variance (un échantillon)

- Condition : population normale $N(\mu, \sigma)$
- $X^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
- X^2 : v.a. loi du χ^2 à $\nu = n-1$ degrés de liberté (d.l.)
- $P(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}) = 1-\alpha$
- $P\left(\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right) = 1-\alpha$
- $P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$
- $P\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}}\right) = 1-\alpha$
- Intervalle de confiance :
  $\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}$ à un niveau de confiance de $(1-\alpha)100\%$

106

Proportion = moyenne

- Caractère quantitatif (rappel)
  - Moyenne : $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - $n > 30$, σ connu
  - $\bar{X} = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
- Caractère qualitatif
  - Proportion : $\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - $n > 30$, $n\hat{p} \geq 5$, $n(1-\hat{p}) \geq 5$, ni $\hat{p} \approx 0$, ni $\hat{p} \approx 1$
  - $\hat{P} = N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$
- Les proportions (fréquences relatives) sont des moyennes !
- $\bar{X} \longrightarrow \hat{P}$: remplacer
  - $\mu \longrightarrow \pi$
  - $\sigma \longrightarrow \sqrt{\pi(1-\pi)}$

107

Estimation de la proportion

- Caractère quantitatif (rappel)
  - $P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$
  - Intervalle de confiance à un niveau de confiance de $(1-\alpha)100\%$:
    $\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
  - $n_{\text{min}} = \left(\frac{z_{\alpha/2}\sigma}{e_{\text{max}}}\right)^2$: taille d'échantillon minimale
- Caractère qualitatif
  - $P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} < \pi < \hat{P} + z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right) = 1 - \alpha$
  - Intervalle de confiance à un niveau de confiance de $(1-\alpha)100\%$:
    $\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < \pi < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - $n_{\text{min}} = \left(\frac{z_{\alpha/2}}{e_{\text{max}}}\right)^2 \hat{p}(1-\hat{p})$: taille d'échantillon minimale
    estimer $\hat{p}$ (1er échantillonage, $n \geq 30$) ou prendre $\hat{p} = 0.5$ (pire scénario)

108

Estimation du rapport des variances (deux échantillons)

- Échantillons aléatoires et indépendants de tailles n_1, n_2
- Provenant de populations normales de variances σ_1^2, σ_2^2
- Variances des échantillons : S_1^2, S_2^2
- $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{V_1/\nu_1}{V_2/\nu_2}$
- $V_i = \frac{(n_i-1)S_i^2}{\sigma_i^2}$: v.a. indépendantes, loi du χ^2 à $\nu_i = n_i - 1$ d.l.
- F : loi de Fisher - Snedecor avec ν_1 et ν_2 d.l.
- $P(f_{1-\alpha/2}(\nu_1,\nu_2) < F < f_{\alpha/2}(\nu_1,\nu_2)) = 1 - \alpha$
- $P\left(f_{1-\alpha/2}(\nu_1,\nu_2) < \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} < f_{\alpha/2}(\nu_1,\nu_2)\right) = 1 - \alpha$
- $P\left(\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(\nu_1,\nu_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{f_{1-\alpha/2}(\nu_1,\nu_2)}\right) = 1 - \alpha$
- $P\left(\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(\nu_1,\nu_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{\alpha/2}(\nu_2,\nu_1)\right) = 1 - \alpha$

109

Tests d'hypothèse 110

Définitions

- Hypothèse : énoncé concernant les caractéristiques d'une population
- Hypothèse nulle : fixer un paramètre θ à une valeur particulière θ_0
  - $H_0 : \theta = \theta_0$
- Hypothèse alternative (trois choix possibles)
  - $H_1 : \theta \neq \theta_0$ (test bilatéral)
  - $H_1 : \theta < \theta_0$ (test unilatéral)
  - $H_1 : \theta > \theta_0$ (test unilatéral)
- Test : procédure suivie afin d'accepter/rejeter H_0
- Rejet > Acceptation (non-rejet)
- En pratique : formuler H_0 comme l'opposé de ce qu'on veut démontrer !

111

Types et probabilités d'erreur

Types d'erreur		
décision \ état du monde	H_0 vraie	H_1 vraie
non-rejet de H_0	OK	Type II
rejet de H_0	Type I	OK

- $P(\text{Type I}) = P(\text{rejet de } H_0 | H_0 \text{ vraie}) = \alpha$
- $P(\text{Type II}) = P(\text{non-rejet de } H_0 | H_1 \text{ vraie}) = \beta$

Probabilités d'erreur		
décision \ état du monde	H_0 vraie	H_1 vraie
non-rejet de H_0	$1-\alpha$	β
rejet de H_0	α	$1-\beta$

- α : seuil de signification (calculé dans l'univers de H_0, ok)
- $1-\beta$: puissance du test (calculée dans l'univers de H_1, ? ? ?)
  - Préciser H_1, ensuite calculer une valeur de β liée à cette H_1

112

Tests : la procédure à suivre

1. Formuler les hypothèses H_0 et H_1
2. Choisir le seuil de signification α (typiquement 1% ou 5%)
3. Déterminer la statistique utilisée ainsi que sa distribution
4. Définir la région critique (région de rejet de H_0)
5. Adopter une règle de décision (à partir des valeurs critiques)
6. Prélever un échantillon et faire les calculs
7. Décider

113

Test sur une moyenne (1/3)

1. $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ (test bilatéral)
2. α à définir
3. Statistique à utiliser : $\bar{X}$; distribution :
   $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ si on connaît σ ou n grand (cas présenté dans la suite)
   $T = (\bar{X} - \mu)/(S/\sqrt{n})$ si on ne connaît pas σ et n petit (population normale)
4. $P(\text{non-rejet de } H_0 | H_0 \text{ vraie}) = 1 - \alpha$
   $P(\text{non-rejet de } H_0 | \mu = \mu_0) = 1 - \alpha$
   $P(z_{1-\alpha/2} < Z < z_{\alpha/2} | \mu = \mu_0) = 1 - \alpha$
   $P(-z_{\alpha/2} < Z < z_{\alpha/2} | \mu = \mu_0) = 1 - \alpha$
   $P(-z_{\alpha/2} < (\bar{X} - \mu)/(\sigma/\sqrt{n}) < z_{\alpha/2} | \mu = \mu_0) = 1 - \alpha$
   $P(-z_{\alpha/2} < (\bar{X} - \mu_0)/(\sigma/\sqrt{n}) < z_{\alpha/2}) = 1 - \alpha$
   région critique : $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n}) < -z_{\alpha/2}$ et $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n}) > z_{\alpha/2}$
5. Règle de décision :
   rejeter H_0 si $\bar{x} < \bar{x}_{c1} = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ ou $\bar{x} > \bar{x}_{c2} = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

114

Test sur une moyenne (2/3)

1. $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$ (test unilatéral)
2. α à définir
3. Statistique à utiliser : $\bar{X}$; distribution :
   $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ si on connaît σ ou n grand (cas présenté dans la suite)
   $T = (\bar{X} - \mu)/(S/\sqrt{n})$ si on ne connaît pas σ et n petit (population normale)
4. $P(\text{non-rejet de } H_0 | H_0 \text{ vraie}) = 1 - \alpha$
   $P(\text{non-rejet de } H_0 | \mu = \mu_0) = 1 - \alpha$
   $P(Z < z_\alpha | \mu = \mu_0) = 1 - \alpha$
   $P((\bar{X} - \mu)/(\sigma/\sqrt{n}) < z_\alpha | \mu = \mu_0) = 1 - \alpha$
   $P((\bar{X} - \mu_0)/(\sigma/\sqrt{n}) < z_\alpha) = 1 - \alpha$
   région critique : $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n}) > z_\alpha$
5. Règle de décision :
   rejeter H_0 si $\bar{x} > \bar{x}_c = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$

115

Test sur une moyenne (3/3) : taille de l'échantillon

- $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$ (test unilatéral)
- $\alpha = P(\text{rejet de } H_0 | H_0 \text{ vraie}) = P(\text{rejet de } H_0 | \mu = \mu_0) = P(Z > z_\alpha | \mu = \mu_0)$
  $= P((\bar{X} - \mu)/(\sigma/\sqrt{n}) > z_\alpha | \mu = \mu_0)$
  $= P((\bar{X} - \mu_0)/(\sigma/\sqrt{n}) > z_\alpha)$
- Règle de décision : rejeter H_0 si $\bar{x} > \bar{x}_c = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$
- $\beta = P(\text{rejet de } H_1 | H_1 \text{ vraie}) = P(\text{non-rejet de } H_0 | H_1 \text{ vraie})$
  $= P(\bar{X} < \bar{x}_c | H_1 \text{ vraie})$
- Préciser $H_1 : \mu = \mu_0 + \delta$
- $\beta = P(\bar{X} < \bar{x}_c | \mu = \mu_0 + \delta) = P(Z < (\bar{x}_c - \mu)/(\sigma/\sqrt{n}) | \mu = \mu_0 + \delta)$
- $= P(Z < \frac{\bar{x}_c - \mu_0}{\sigma/\sqrt{n}} - \frac{\delta}{\sigma/\sqrt{n}})$
- $= P(Z < z_\alpha - \frac{\delta}{\sigma/\sqrt{n}})$
- $-z_\beta = z_\alpha - \frac{\delta}{\sigma/\sqrt{n}}$
- $n = (z_\alpha + z_\beta)^2 \frac{\sigma^2}{\delta^2}$

116

Test sur une variance (1/2)

1. $H_0 : \sigma = \sigma_0$, $H_1 : \sigma \neq \sigma_0$ (test bilatéral)
2. α à définir
3. Statistique à utiliser : S ; distribution :
   $X^2 = \frac{(n-1)S^2}{\sigma^2}$, v.a. loi du χ^2 à $\nu = n-1$ degrés de liberté (population normale)
4. $P(\text{non-rejet de } H_0 | H_0 \text{ vraie}) = 1-\alpha$
   $P(\text{non-rejet de } H_0 | \sigma = \sigma_0) = 1-\alpha$
   $P(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2} | \sigma = \sigma_0) = 1-\alpha$
   $P\left(\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{\alpha/2}\right) = 1-\alpha$
   $P\left(\frac{\chi^2_{1-\alpha/2}\sigma_0^2}{(n-1)} < S^2 < \frac{\chi^2_{\alpha/2}\sigma_0^2}{(n-1)}\right) = 1-\alpha$
   région critique : $X^2 < \chi^2_{1-\alpha/2}$ et $X^2 > \chi^2_{\alpha/2}$
5. Règle de décision :
   rejeter H_0 si $s^2 < s^2_{c1} = \chi^2_{1-\alpha/2}\sigma_0^2/(n-1)$ ou $s^2 > s^2_{c2} = \chi^2_{\alpha/2}\sigma_0^2/(n-1)$

117

Test sur une variance (2/2)

1. $H_0 : \sigma = \sigma_0$, $H_1 : \sigma < \sigma_0$ (test unilatéral)
2. α à définir
3. Statistique à utiliser : S ; distribution :
   $X^2 = \frac{(n-1)S^2}{\sigma^2}$, v.a. loi du χ^2 à $\nu = n-1$ degrés de liberté (population normale)
4. $P(\text{non-rejet de } H_0 | H_0 \text{ vraie}) = 1-\alpha$
   $P(\text{non-rejet de } H_0 | \sigma = \sigma_0) = 1-\alpha$
   $P(\chi^2_{1-\alpha} < X^2 | \sigma = \sigma_0) = 1-\alpha$
   $P\left(\chi^2_{1-\alpha} < \frac{(n-1)S^2}{\sigma_0^2}\right) = 1-\alpha$
   $P\left(\frac{\chi^2_{1-\alpha}\sigma_0^2}{(n-1)} < S^2\right) = 1-\alpha$
   région critique : $X^2 < \chi^2_{1-\alpha}$
5. Règle de décision :
   rejeter H_0 si $s^2 < s^2_c = \chi^2_{1-\alpha}\sigma_0^2/(n-1)$

118

Test sur une proportion

1. $H_0 : \pi = \pi_0$, $H_1 : \pi \neq \pi_0$ (test bilatéral)
2. α à définir
3. Statistique à utiliser : $\hat{P}$; distribution :
   $Z = (\hat{P} - \pi)/(\sqrt{\pi(1-\pi)}/\sqrt{n})$
4. $P(\text{non-rejet de } H_0 | H_0 \text{ vraie}) = 1 - \alpha$
   $P(\text{non-rejet de } H_0 | \pi = \pi_0) = 1 - \alpha$
   $P(-z_{\alpha/2} < (\hat{P} - \pi_0)/(\sqrt{\pi_0(1-\pi_0)}/\sqrt{n}) < z_{\alpha/2}) = 1 - \alpha$
   région critique : $Z < -z_{\alpha/2}$ et $Z > z_{\alpha/2}$
5. Règle de décision :
   rejeter H_0 si $\hat{p} < \hat{p}_{c1} = \pi_0 - z_{\alpha/2}\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}$ ou $\hat{p} > \hat{p}_{c1} = \pi_0 + z_{\alpha/2}\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}$

1. $H_0 : \pi = \pi_0$, $H_1 : \pi > \pi_0$ (test unilatéral)
   ...
5. Règle de décision : rejeter H_0 si $z > z_\alpha$
   c.à.d. $\hat{p} > \hat{p}_c = \pi_0 + z_\alpha\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}$

119

Récapitulatif : un échantillon

120

Statistiques d'un échantillon : moyenne

Paramètre θ	μ			
Population Écart-type σ	$\approx$ normale connu	— connu	$\approx$ normale inconnu	
Échantillon	—	$n > 30$	$n > 30$	$n < 30$
Statistique $\hat{\Theta}$	$\bar{X}$			
St. normalisée	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$		$Z = \frac{\bar{X}-\mu}{S/\sqrt{n}}$	$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$
Distribution	$N(0,1)$			Student (ν)
D.L.	—			$n-1$
Mesure $\hat{\theta}$	$\bar{x}$			

121

Statistiques d'un échantillon : proportion, variance

Paramètre θ	π	σ^2
Population	—	$\approx$ normale
Écart-type σ	—	—
Échantillon	$n > 30$ [a]	—
Statistique $\hat{\Theta}$	$\hat{P}$	S^2
St. normalisée	$Z = \frac{\hat{P}-\pi}{\sqrt{\pi(1-\pi)/n}}$	$X^2 = \frac{(n-1)S^2}{\sigma^2}$
Distribution	$N(0,1)$	khi-deux (ν)
D.L.	—	$n-1$
Mesure $\hat{\theta}$	$\hat{p}$	s^2

122

[a] En plus : $n\hat{p} \geq 5$, $n(1-\hat{p}) \geq 5$, ni $\hat{p} \approx 0$, ni $\hat{p} \approx 1$.

Estimation / tests : un échantillon

Stat. norm.	Intervalle de confiance	Test d'hypothèse $H_0 : \theta = \theta_0$		
		$H_1 : \theta \neq \theta_0$	$H_1 : \theta < \theta_0$	$H_1 : \theta > \theta_0$
Z	$-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}}$	$z < -z_{\frac{\alpha}{2}}$ ou $> z_{\frac{\alpha}{2}}$	$z < -z_\alpha$	$z > z_\alpha$
T	$-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}$	$t < -t_{\frac{\alpha}{2}}$ ou $> t_{\frac{\alpha}{2}}$	$t < -t_\alpha$	$t > t_\alpha$
X^2	$\chi^2_{1-\frac{\alpha}{2}} < \chi^2 < \chi^2_{\frac{\alpha}{2}}$	$\chi^2 < \chi^2_{1-\frac{\alpha}{2}}$ ou $> \chi^2_{\frac{\alpha}{2}}$	$\chi^2 < \chi^2_{1-\alpha}$	$\chi^2 > \chi^2_\alpha$
	mettre sous la forme : $\theta_L < \theta < \theta_H$	« entrer dans le monde de H_0 » : $\theta = \theta_0$, calculer z, t, χ^2 à partir des mesures ; décisions de *rejet* de H_0		

- Intervalle de confiance : niveau de confiance $1-\alpha$
- Tests d'hypothèse : seuil de signification α
- Voir tableaux unifiés dans le document « Aide-mémoire ».

123

Intervalles et tests avec deux échantillons 124

Distribution de la différence des moyennes (1/6) - rappel #98

- Conditions : σ_1, σ_2 connus et
  - populations normales $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$ ou
  - $n_1 > 30$ et $n_2 > 30$, ou
  - populations « presque » normales
- Échantillons aléatoires et indépendants de tailles n_1, n_2 ; moyennes $\bar{X}_1, \bar{X}_2$
  - $\boxed{\bar{X}_1 - \bar{X}_2}$: normale
  - $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$
  - $\sigma^2_{\bar{X}_1 - \bar{X}_2} \overset{\text{ind}}{=} \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

125

Distribution de la différence des moyennes (2/6)

- Échantillons aléatoires et indépendants de tailles n_1, n_2
- Populations normales ou grands échantillons ($n_1 > 30$, $n_2 > 30$)
- σ_1, σ_2 : connus
- $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \to N(0,1)$
- Intervalle de confiance : $(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- Test d'hypothèse :

  1. $H_0 : \mu_1 - \mu_2 = d_0$, $H_1 : \mu_1 - \mu_2 \neq d_0$ (test bilatéral)
  5. Règle de décision : rejeter H_0 si $z < -z_{\alpha/2}$ ou $z > z_{\alpha/2}$

     $(\bar{x}_1 - \bar{x}_2) < (\bar{x}_1 - \bar{x}_2)_{c1} = d_0 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ ou

     $(\bar{x}_1 - \bar{x}_2) > (\bar{x}_1 - \bar{x}_2)_{c2} = d_0 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

126

Distribution de la différence des moyennes (3/6)

- Échantillons aléatoires et indépendants de tailles n_1, n_2
- Populations normales et grands échantillons ($n_1 > 30$, $n_2 > 30$)
- σ_1, σ_2 : inconnus
- $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \rightarrow \approx N(0,1)$
- *Équivalent de $T \rightarrow Z$ pour grands échantillons*
- Intervalle de confiance : $(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Test d'hypothèse :

1. $H_0 : \mu_1 - \mu_2 = d_0$, $H_1 : \mu_1 - \mu_2 > d_0$ (test unilatéral)
5. Règle de décision : rejeter H_0 si $z > z_\alpha$
   $(\bar{x}_1 - \bar{x}_2) > (\bar{x}_1 - \bar{x}_2)_c = d_0 + z_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

127

Distribution de la différence des moyennes (4/6)

- Échantillons aléatoires et indépendants de tailles n_1, n_2
- Populations normales et petits échantillons ($n_1 < 30$ ou $n_2 < 30$)
- σ_1, σ_2 : inconnus mais $\sigma_1 = \sigma_2$ (à tester)
- $T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_c^2}{n_1} + \frac{S_c^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightarrow$ Student
- Variance commune : $S_c^2 = \frac{\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$
- T : Student à $(n_1 + n_2 - 2)$ d.l.
- Intervalle de confiance : $(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- Test d'hypothèse : ...
- À propos des conditions :
  - $\sigma_1 \approx \sigma_2$ ou populations $\approx$ normales : OK
  - $\sigma_1 \neq \sigma_2$ et normales : OK si $n_1 = n_2$

128

Distribution de la différence des moyennes (5/6)

- Échantillons aléatoires et indépendants de tailles n_1, n_2
- Populations normales et petits échantillons ($n_1 < 30$ ou $n_2 < 30$)
- σ_1, σ_2 : inconnus et $\sigma_1 \neq \sigma_2$ (à tester)
- $T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \rightarrow$ Student à ν d.l. ; $\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$
- Arrondir ν au nombre entier *inférieur*.
- Intervalle de confiance : $(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Test d'hypothèse :

1. $H_0 : \mu_1 - \mu_2 = d_0$, $H_1 : \mu_1 - \mu_2 < d_0$ (test unilatéral)
5. Règle de décision : rejeter H_0 si $t < t_\alpha$
   $(\bar{x}_1 - \bar{x}_2) < (\bar{x}_1 - \bar{x}_2)_c = d_0 - t_\alpha\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

129

Distribution de la différence des moyennes (6/6)

- Échantillons aléatoires et appariés de tailles $n_1 = n_2 = n$
- Appariés : « avant / après »
- Population : nouvelle v.a. $D = X_1 - X_2$ (μ_D, σ_D)
- Échantillon : calculer $d_i = x_{1i} - x_{2i}$; oublier X_1, X_2 !
- Population normale ou grands échantillons ($n > 30$), σ_D connu :
  $Z = \frac{\bar{D} - \mu_D}{\sigma_D/\sqrt{n}} \rightarrow N(0,1)$
- Population normale et petits échantillons ($n < 30$), σ_D inconnu :
  $T = \frac{\bar{D} - \mu_D}{s_D/\sqrt{n}}$ à $(n-1)$ d.l.
- Intervalle de confiance : $\bar{d} - t_{\alpha/2}\frac{s_D}{\sqrt{n}} < \mu_D < \bar{d} + t_{\alpha/2}\frac{s_D}{\sqrt{n}}$
- Test d'hypothèse : ...
- Échantillons appariés : un seul nouvel échantillon !

130

Distribution de la différence des proportions

- ☐ Échantillons aléatoires et indépendants de tailles n_1, n_2
- ☐ Grands échantillons ($n_1 > 30$, $n_2 > 30$)
- ☐ Proportions : $\hat{P}_i = N(\pi_i, \sqrt{\pi_i(1-\pi_i)}/\sqrt{n_i})$
- ☐ $Z = \frac{(\hat{P}_1 - \hat{P}_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \to N(0,1)$
- ☐ Intervalle de confiance :
  $(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}} < \pi_1 - \pi_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$;
  remplacer $\pi_i(1-\pi_i) \to \hat{p}_i(1-\hat{p}_i)$
- ☐ Test d'hypothèse :

  1. $H_0 : \pi_1 - \pi_2 = d_0$ ($\pi_1 = \pi_2 + d_0$) , $H_1 : \pi_1 - \pi_2 > d_0$ (test unilatéral)
  5. Règle de décision : rejeter H_0 si $z > z_\alpha$
     $(\hat{p}_1 - \hat{p}_2) > (\hat{p}_1 - \hat{p}_2)_c = d_0 + z_\alpha\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$
     Si $d_0 = 0$, $\pi_1 = \pi_2$: remplacer $\pi_j \to \hat{p} = \frac{\sum_{i=1}^{n_1} x_{1i} + \sum_{i=1}^{n_2} x_{2i}}{n_1+n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$
     Si $d_0 \neq 0$: remplacer $\pi_j \to \hat{p}_j$

131

Distribution du rapport des variances (1/2) - rappel #99

- ☐ Échantillons aléatoires et indépendants de tailles n_1, n_2
- ☐ Provenant de populations normales de variances σ_1^2, σ_2^2
- ☐ Variances des échantillons : S_1^2, S_2^2
- ☐ $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{V_1/\nu_1}{V_2/\nu_2}$
- ☐ $V_i = \frac{(n_i-1)S_i^2}{\sigma_i^2}$: v.a. indépendantes, loi du χ^2 à $\nu_i = n_i - 1$ d.l.
- ☐ F : loi de Fisher (1924) - Snedecor (1934) avec ν_1 et ν_2 d.l.
- ☐ $F \geq 0$
- ☐ $P(F > f_\alpha(\nu_1, \nu_2)) = \alpha$ (définition de $f_\alpha(\nu_1, \nu_2)$)
- ☐ $f_\alpha(\nu_1, \nu_2) = \frac{1}{f_{1-\alpha}(\nu_2, \nu_1)}$ (propriété de la loi F)

132

Distribution du rapport des variances (2/2)

- $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2}$
- Intervalle de confiance (niveau de confiance $1-\alpha$) :
  - $f_{1-\alpha/2}(\nu_1,\nu_2) < f < f_{\alpha/2}(\nu_1,\nu_2)$
  - $\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2}(\nu_1,\nu_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}\frac{1}{f_{1-\alpha/2}(\nu_1,\nu_2)}$
- Test d'hypothèse $H_0 : \sigma_1 = \sigma_2$
- Règle de décision : rejeter H_0 si
  - $H_1 : \sigma_1 \neq \sigma_2$
    $f < f_{1-\alpha/2}$ ou $f > f_{\alpha/2}$ c-à-d $s_1^2/s_2^2 < f_{1-\alpha/2}$ ou $s_1^2/s_2^2 > f_{\alpha/2}$
  - $H_1 : \sigma_1 > \sigma_2$
    $f > f_{\alpha}$ c-à-d $s_1^2/s_2^2 > f_{\alpha}$
  - $H_1 : \sigma_1 < \sigma_2$
    $f < f_{1-\alpha}$ c-à-d $s_1^2/s_2^2 < f_{1-\alpha/2}$

133

Récapitulatif : deux échantillons

134

Statistiques de deux (grands) échantillons : moyenne

Paramètre θ	$\mu_2 - \mu_1$		
Populations	$\approx$ normales	—	$\approx$ normales
Écart-types σ_1, σ_2	connus	connus	inconnus
Échantillons	—	$n_1 > 30$ et $n_2 > 30$	$n_1 > 30$ et $n_2 > 30$
Statistique $\hat{\Theta}$	$\bar{X}_2 - \bar{X}_1$		
St. normalisée	$Z = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$		$Z = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
Distribution	$N(0,1)$		
Degrés de liberté	—		
Mesure $\hat{\theta}$	$\bar{x}_2 - \bar{x}_1$		

135

Statistiques de deux (petits) échantillons : moyenne

Paramètre θ	$\mu_2 - \mu_1$	
Populations	$\approx$ normales	
Écart-types σ_1, σ_2	inc., $\sigma_1 = \sigma_2$ ou $n_1 = n_2$	inc., $\sigma_1 \neq \sigma_2$ et $n_1 \neq n_2$
Échantillons	$n_1 < 30$ ou $n_2 < 30$	
Statistique $\hat{\Theta}$	$\bar{X}_2 - \bar{X}_1$	
St. normalisée	$T = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{S_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$T = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
Distribution	Student (ν)	
Degrés de liberté	$n_1 + n_2 - 2$	ν^*
Mesure $\hat{\theta}$	$\bar{x}_2 - \bar{x}_1$	
Rappels	S_c : diapo #128	ν^* : diapo #129

136

Statistiques de deux échantillons : proportion, variance

Paramètre θ	$\pi_2 - \pi_1$	σ_1^2/σ_2^2
Populations	—	$\approx$ normales
Écart-types σ_1, σ_2	—	—
Échantillons	$n_1 > 30$ et $n_2 > 30$ [a]	—
Statistique $\hat{\Theta}$	$\hat{P}_2 - \hat{P}_1$	F
St. normalisée	$Z = \frac{(\hat{P}_2 - \hat{P}_1) - (\pi_2 - \pi_1)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}$	$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$
Distribution	$N(0,1)$	Fischer (ν_1, ν_2)
Degrés de liberté	—	$n_1 - 1, n_2 - 1$
Mesure $\hat{\theta}$	$\hat{p}_2 - \hat{p}_1$	s_1^2/s_2^2

137

[a] En plus : $n_i \hat{p}_i \geq 5$, $n_i(1 - \hat{p}_i) \geq 5$, ni $\hat{p}_i \approx 0$, ni $\hat{p}_i \approx 1$ $(i = 1, 2)$.

Estimation / tests : deux échantillons

Stat. norm.	Intervalle de confiance	Test d'hypothèse $H_0 : \theta = \theta_0$		
		$H_1 : \theta \neq \theta_0$	$H_1 : \theta < \theta_0$	$H_1 : \theta > \theta_0$
Z	$-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}}$	$z < -z_{\frac{\alpha}{2}}$ ou $> z_{\frac{\alpha}{2}}$	$z < -z_\alpha$	$z > z_\alpha$
T	$-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}$	$t < -t_{\frac{\alpha}{2}}$ ou $> t_{\frac{\alpha}{2}}$	$t < -t_\alpha$	$t > t_\alpha$
F	$f_{1-\frac{\alpha}{2}} < f < f_{\frac{\alpha}{2}}$	$f < f_{1-\frac{\alpha}{2}}$ ou $f > f_{\frac{\alpha}{2}}$	$f < f_{1-\alpha}$	$f > f_\alpha$
	mettre sous la forme : $\theta_L < \theta < \theta_H$	« entrer dans le monde de H_0 » : $\theta = \theta_0$, calculer z, t, χ^2 à partir des mesures ; décisions de *rejet* de H_0		

- ☐ Intervalle de confiance : niveau de confiance $1 - \alpha$
- ☐ Tests d'hypothèse : seuil de signification α
- ☐ Voir tableaux unifiés dans le document « Aide-mémoire ».

138

Tests : au délà du seuil de signification 139

Seuil descriptif (p-value)

- Test statistique : « 2. Choisir le seuil de signification α »
- « Typiquement 1% ou 5% »
- Comment choisir ?
- Comment décider ?
- *Pourquoi* choisir α ?
- Tests classiques :
  - Mesurer $\hat{\theta}$; comparer $\hat{\theta}$ aux valeurs critiques $\hat{\theta}_c$
  - Valeurs critiques dépendent de α
- Alternative
  - Calculer α_p (p-value) telle que $\hat{\theta} = \hat{\theta}_c$
  - α_p : rejeter H_0 de façon marginale
- P-value (seuil descriptif) : la plus petite valeur de $\alpha = P(\text{rejeter}H_0|H_0 \text{ vraie})$ qui conduirait au rejet de H_0
- La probabilité de se retrouver « au moins aussi loin » de la H_0 – dans le sens de la H_1 – que l'échantillon examiné, si H_0 est vraie.

140

Seuil descriptif (p-value) : exemple (1/3)

□ Test sur la moyenne, petit échantillon, population normale, σ inconnu

1. $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ (test bilatéral)
2. α à définir
3. Statistique à utiliser : $\bar{X}$; distribution :
   $T = (\bar{X} - \mu)/(S/\sqrt{n})$
4. Région critique : $T < -t_{\alpha/2}$ et $T > t_{\alpha/2}$
5. Règle de décision :
   rejeter H_0 si $t < -t_{\alpha/2}$ ou $> t_{\alpha/2}$
6. Prélever un échantillon et faire les calculs
7. Décider

141

Seuil descriptif (p-value) : exemple (2/3)

6. Prélever un échantillon et faire les calculs
   Population $N(0.5, 1)$, $n = 5$

```
-> x = 0.5+rand(1,5,'normal')
x = 0.4303745 -1.2195277 -0.3570756 2.2734783 -0.5112132
-> mean(x)
ans = 0.1232073
-> stdev(x)
ans = 1.337359
```

   $\mu_0 = 0$, calculer t :

```
-> t = ( mean(x) - 0 ) / ( stdev(x) / sqrt(5) )
t = 0.2060029
```

   $\alpha = 0.05$, calculer $t_c = t_{\alpha/2}$:

```
-> cdft("T",4,1-0.025,0.025)
ans = 2.776445
```

7. Décider : $-t_{\alpha/2} < t < t_{\alpha/2}$, on ne peut pas rejeter $H_0 : \mu = \mu_0 = 0$

142

Seuil descriptif (p-value) : exemple (3/3)

6. Prélever un échantillon et faire les calculs
   $\mu_0 = 0$, calculer t :
   `-> t = ( mean(x) - 0 ) / ( stdev(x) / sqrt(5) )`
   `ans = 0.2060029`
   Quelle est la valeur de α qui donne $t = t_c = t_{\alpha/2}$?
   `-> [P,Q]=cdft('PQ',t,4`
   `Q=0.4234244 P= 0.5765756`
   p-value$/2$ = 0.4234244, p-value $= 0.8468488$
7. Décider : échantillon très probable si H_0 est vraie

143

Test du χ^2 144

Définition – cadre général

Comparer, à l'issue d'une expérience aléatoire, des fréquences expérimentales aux fréquences prévues par la théorie (Pearson, 1900).

- k : nombre de fréquences à comparer (nombre de classes)
- o_i : fréquences Observées (obtenues expérimentalement)
- e_i : fréquences « Espérées » (théoriques, à calculer)
-

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- Loi du χ^2 à ν degrés de liberté ; si $o_i = e_i$, $\chi^2 = 0$, sinon $\chi^2 > 0$
- Calculer χ^2 à partir de o_i, e_i ; obtenir $\alpha = P(X^2 > \chi^2)$, la p-value
- $\nu = k - 1 -$ (nombre de paramètres estimés utilisés dans le calcul de e_i)
- Condition : $e_i \geq 5$ au moins pour 80% des classes ; $e_i > 0$ pour les autres
- Applications : test d'adéquation, d'indépendance, d'homogénéité, de proportions

145

Test d'adéquation (ou d'ajustement)

H_0 : les données expérimentales ont été obtenues à partir d'une population suivant la loi $p_X(x)$ (p.ex., normale, uniforme, etc).

- Example : données sur plusieurs lancers d'un dé (données simulées...)
Face	1	2	3	4	5	6	Total N
Fréquence (o_i)	1037	937	1055	1034	929	1008	6000

```
O = [ 1037 937 1055 1034 929 1008]
```

- H_0 : le dé est bien équilibré ; $p_i = 1/6$, $\boxed{e_i = p_i N} = 1000$

```
e=ones(1,6)*1000
```

- Conditions : OK (sinon grouper des classes voisines)
- Calculer $\chi^2 = 14.624$ (`sum((O-e).^2)/1000`
- $\nu = 6 - 1 - 0 = 5$
- p-value : $P(X^2 > 14.624) =$

```
[P Q]=cdfchi(PQ,sum((O-e).^2)/1000,5)
Q= 0.0120957 P=0.9879047
```

- On peut rejeter H_0 au seuil de signification 5%

146

Test d'indépendance / tableau de contingence

On mesure, sur chaque individu d'un échantillon aléatoire de taille n, deux caractères X et Y, à l et c modalités, respectivement.

H_0 : les deux caractères X et Y sont indépendants.

- Example : le tabac et les jeunes, INPES, baromètre santé 2000 (tr. #20)
Sexe \ Fumeur	Oui	Non	Total
Homme	340 (310)	314 (344)	654
Femme	289 (319)	384 (354)	673
Total	629	698	**1327**
- H_0 : X et Y sont indépendants ; $\pi_{ij} = \pi_i \pi_j$ $(i = 1, \ldots, l\,;\ j = 1, \ldots, c)$
- On *estime* π_i et π_j à partir des fréquences marginales de l'échantillon
- $\pi_{ij} = \pi_i \pi_j \rightarrow \frac{e_{ij}}{n} = \frac{\sum_{j=1}^{c} o_{ij}}{n} \frac{\sum_{i=1}^{l} o_{ij}}{n} \rightarrow \boxed{e_{ij} = \frac{1}{n} \sum_{j=1}^{c} o_{ij} \sum_{i=1}^{l} o_{ij}}$
- Degrés de liberté $\nu = (lc - 1) - 1 - [(l - 1) + (c - 1)] = \boxed{(l - 1)(c - 1)}$
- Conditions : OK (sinon ? augmenter la taille de l'échantillon !)

147

Test d'indépendance : correction de Yates

- Si $\nu = 1$ (tableau 2×2) utiliser :

$$\chi^2 = \sum_{i,k} \frac{(|o_{ij} - e_{ij}| - 0.5)^2}{e_{ij}}$$

- Calculer $\chi^2 = 10.5256$
- $\nu = (2-1)(2-1) = 1$
- p-value : $P(X^2 > 10.5256) =$
  `[P Q]=cdfchi('PQ', 10.5256, 1)`
  `Q=0.0011773 P = 0.998227`
- On peut rejeter H_0 au seuil de signification 1%

148

Test d'homogénéité

À partir de c populations, on obtient c échantillons aléatoires et indépendants, de taille n_j ($j = 1, \ldots, c$). On mesure sur chaque individu le même caractère X, à l modalités.

H_0 : la proportion d'individus appartenant à la i-ème modalité ($i = 1, \ldots, l$), reste la même pour toutes les populations (les populations sont *homogènes* par rapport au caractère étudié).

- Example : notes (fictives) échantillonnées dans trois parcours

Note \ Parcours	I	II	III	Total
$0 \leq x < 6$	32	15	8	55
$6 \leq x < 12$	123	60	43	226
$12 \leq x \leq 20$	145	125	149	419
Total (n_j)	**300**	**200**	**200**	**700**

- H_0 : proportion de chaque modalité constante ;
  $\pi_{i1} = \pi_{i2} = \ldots = \pi_{ic} = \pi_i$ ($i = 1, \ldots, l$)
- On *estime* π_i à partir des fréquences marginales de l'échantillon

149

Test d'homogénéité

Note \ Parcours	I	II	III	Total
$0 \leq x < 6$	32 (23.57)	15 (15.71)	8 (15.71)	55
$6 \leq x < 12$	123 (96.86)	60 (64.57)	43 (64.57)	226
$12 \leq x \leq 20$	145 (179.57)	125 (119.71)	149 (119.71)	419
Total (n_j)	**300**	**200**	**200**	**700**

- H_0 : proportion de chaque modalité constante ;
  $\pi_{i1} = \pi_{i2} = \ldots = \pi_{ic} = \pi_i \ (i = 1, \ldots, l)$
- On *estime* π_i à partir des fréquences marginales de l'échantillon
- $\pi_{ij} = \pi_i \rightarrow \frac{e_{ij}}{n_j} = \frac{\sum_{j=1}^{c} o_{ij}}{n} \rightarrow \boxed{e_{ij} = \frac{1}{n} \sum_{j=1}^{c} o_{ij} \underbrace{\sum_{i=1}^{l} o_{ij}}_{n_j}}$
- Degrés de liberté $\nu = (lc - 1) - 1 - [(l - 1) + (c - 1)] = \boxed{(l - 1)(c - 1)}$
- Conditions : OK (sinon ? augmenter la taille de l'échantillon !)
- Même formule que le test d'indépendance !

150

Test d'homogénéité

- $$\boxed{\chi^2 = \sum_{i,k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}}$$
- Calculer $\chi^2 = 35.4729$
- $\nu = (3 - 1)(3 - 1) = 4$
- p-value : $P(X^2 > 35.4729) =$
  `[P Q]=cdfchi('PQ', 35.4729, 4)`
  `Q=3.714026 10^7` $P = 0.9999996$
- On peut rejeter H_0 pratiquement à n'importe quel seuil de signification !

151

Test de proportions

À partir de c populations, on obtient c échantillons aléatoires et indépendants, de taille n_j ($j = 1, \ldots, c$). On mesure sur chaque individu le même caractère X, à 2 modalités (« oui » / « non »).

H_0 : la proportion de « oui » reste la même pour toutes les populations (cas spécial du test d'homogénéité, $l = 2$).

- Example : nombre de pièces défectueuses et moment de production

Pièces\ Créneau	Matin	Après-midi	Nuit	Total
Défectueuses (« O »)	45 (56.97)	55 (56.67)	70 (56.37)	170
Normales (« N »)	905 (893.03)	890 (888.33)	870 (883.63)	2665
Total (n_j)	**950**	**945**	**940**	**2835**

- $H_0 : \pi_1 = \pi_2 = \ldots = \pi_c = \pi$
- On *estime* π à partir des fréquences marginales de l'échantillon
- « Oui » : $\pi_j = \pi \rightarrow \frac{e_{1j}}{n_j} = \frac{\sum_{j=1}^{c} o_{1j}}{n}$
- « Non » : $1 - \pi_j = 1 - \pi \rightarrow \frac{e_{2j}}{n_j} = \frac{\sum_{j=1}^{c} o_{2j}}{n}$

152

Test de proportions

- $e_{ij} = \frac{n_j}{n} \sum_{j=1}^{c} o_{ij} \rightarrow \boxed{e_{ij} = \frac{1}{n} \sum_{j=1}^{c} o_{ij} \sum_{i=1}^{l} o_{ij}}$
- Même formule que le test d'indépendance / d'homogénéité !
- Degrés de liberté $\boxed{\nu = (2-1)(c-1) = c-1}$
- Conditions : OK (sinon ? augmenter les tailles des échantillons !)
- Calculer $\chi^2 = 6.2339$
- $\nu = (3-1) = 2$
- p-value : $P(X^2 > 6.2339) =$

```
[P Q]=cdfchi('PQ', 6.2339, 2)
Q=0.04429
```

- On peut rejeter H_0 au seuil de signification 5%

153

Test de proportions sans estimation de paramètres

Même contexte qu'avant : c populations, c échantillons, caractère X à deux modalités.

H_0 : les proportions de « oui », $\pi_1, \ldots, \pi_c$, sont égales à $p_1, \ldots, p_c$ (pas d'estimation de paramètres).

- « Oui » : $\pi_j = p_j \to \frac{e_{1j}}{n_j} = p_j$
- « Non » : $1 - \pi_j = 1 - p_j \to \frac{e_{2j}}{n_j} = 1 - p_j$
- $\boxed{\nu = c}$: on ne perd aucun degré de liberté
- Example précédent avec :
  $p_1 = 0.05$, $p_2 = 0.06$, $p_3 = 0.08$ ($\neq 170/2835 \approx 0.06$)
- Calculer $\chi^2 = 0.5836$
- $\nu = 3$
- p-value : $P(X^2 > 0.5836) = 0.9002$
- On ne peut pas rejeter H_0

154

Test d'adéquation à la loi normale (Shapiro–Wilk)

H_0 : les données expérimentales (échantillon de taille n) ont été obtenues à partir d'une population normale.

- Procédure « classique » : test du χ^2 (cf. TD 6)
  1. Répartir les données en classes (histogramme)
  2. Estimer μ et σ avec `cdfnor`
  3a. Calculer les probabilités théoriques p_j des classes
  Calculer les fréquences théoriques $e_j = p_j n$
  Vérifier les conditions sinon regrouper les classes
  3b. Ou répartir en $(M+1)$ classes équiprobables : $e_j = n/(M+1)$
  4. Calculer χ^2 (on perd deux d.l. avec l'estimation de μ et σ !)
- Une grande p-value permet de ne pas rejetter l'hypothèse de normalité

155