Notes cours MathSV

M. Bailly-Bechet

Université Claude Bernard Lyon 1 – France

Table des matières

Partie proba/stats du cours

DIAPOS 1-2-3-4-5

1 Probabilités

Avoir de vrais dés !

1.1 Définitions

1.1.1 Evènements

Une expérience est *aléatoire* si on ne peut en prédire le résultat de manière certaine. Bcp cas en biologie par rapport à physique, chimie, etc...

Un *évènement élémentaire* A est le résultat d'une expérience. Par exemple, tirer un 4 sur un dé. On le note $A = \{4\}$.

Un *évènement composé* est un résultat composé de plusieurs évènement élémentaires : tirer un nombre impair. $A = \{1, 3, 5\}$.

L'ensemble de tous les évènement possibles est appelé *univers* et noté Ω : $\Omega = \{1, 2, 3, 4, 5, 6\}$. Cet évènement est parfois dit *certain*.

Un *évènement est inclus* dans un autre si tous ses éléments se retrouvent dans un autre. $A = \{1\}, B = \{1, 3, 5\}, A \subset B$.

Tjs $A \subset A$.

Un évènement *complémentaire* de l'évènement A est noté $\bar{A}$. C'est l'évènement composé de l'univers, sauf les éléments de A. Si $A = \{2, 4\}$, $\bar{A} = \{1, 3, 5, 6\}$.

Le complémentaire de l'univers est l'évènement *impossible*, ou vide, $\varnothing$.

L'union de deux évènements A et B est la réunion de tous les évènements élémentaires les composant, et se note $A \cup B$. Si $A = \{1, 3, 5\}$ et $B = \{1, 2, 4\}$, $A \cup B = \{1, 2, 3, 4, 5\}$.

$A \cup A = A$.

$A \cup \bar{A} = \Omega$.

L'intersection de deux évènements A et B est l'ensemble des évènements élémentaires qu'ils partagent, et se note $A \cap B$. Si $A = \{1, 3, 5\}$ et $B = \{1, 2, 4\}$, $A \cap B = \{1\}$.

$A \cap \bar{A} = \varnothing$.

$A \cap A = A$.

Deux évènements sont dits *compatibles* si leur intersection est non nulle, *incompatibles* si leur intersection est nulle ($A \cap B = \varnothing$).

Un *système complet* d'évènements est un ensemble d'évènements tels que :

$$\forall i, j A_i \cap A_j = \varnothing \quad \cup_i A_i = \Omega$$

1.1.2 Probabilités

Une *probabilité* $p(A)$ est une nombre réel associé à un évènement A.

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

$p(\Omega) = 1$

$p(\varnothing) = 0$

Théorème probabilités totales : $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.

Sur un dé $p(\{3\} \cup \{2\}) = p(3) + p(\{2\})$,

ou $p(\{3, 6\} \cup \{2, 4, 6\}) = p(\{3, 6\}) + p(\{2, 4, 6\}) - p(\{6\})$.

1.2 Probabilités conditionelles

$p(A|B) = \frac{p(A \cap B)}{p(B)}$,

ou $p(A \cap B) = p(A|B)p(B)$.

Exple : proba de faire 9 sur 2 dés (un rouge un bleu) sachant qu'on a un 4 sur le dé bleu.

$\{A\}$: bleu+rouge= 9.

$\{B\}$: bleu= 4.

$p(A|B) = \frac{p(\{4,5\})}{p(\{4\})} = \frac{\frac{1}{36}}{\frac{1}{6}}$.

Mais avec $\{A\}$: bleu+rouge=11, on a : $p(A|B) = \frac{p(\varnothing)}{p(\{4\})} = 0$.

On dite que deux évènements A et B sont indépendants ssi $p(A|B) = p(A)$.

Si les évènements sont indépendants, on a $p(A|B) = p(A)$, donc on peut écrire $p(A \cap B) = p(A|B)p(B) = p(A)p(B)$.

Décomposition de la probabilité d'un évènement A : si $\{B_1, B2, ...B_n\}$ est un système complet d'évènements, alors on a $p(A) = \sum_{i=1}^{n} p(A|B_i)p(B_i)$

Un exemple : dans un lycée, on sait que

30% de Term. L, 25% de S et 45% de ES.

65% des L font du grec 15% des S font du grec 5% des ES font du grec

Probabilité qu'un élève pris au hasard fasse du grec ?

G = événement "faire du grec"

$P(G) = P(G|L)P(L) + P(G|S)P(S) + P(G|ES)P(ES) = 0.65 * 0.30 + 0.15 * 0.25 + 0.05 * 0.45 = 0.255$

Exemple arbre : illutration

Théorème de Bayes : si $\{B_1, B2, ...B_n\}$ est un système complet d'évènements, alors on a :

$p(B_i|A) = \frac{p(A \cap B_i)}{p(A)} = \frac{p(A \cap B_i)}{\sum_{i=1}^{n} p(A|B_i)p(B_i)} = \frac{p(A|B_i)p(B_i)}{\sum_{i=1}^{n} p(A|B_i)p(B_i)}$

Application : dépistage HIV.

Le dépistage a un résultat positif dans 99% des cas si on est infecté, et dans 1% des cas si on ne l'est pas. Le test est positif : proba que je sois vraiment infecté ?

A : etre positif au test B_1 : etre infecté B_2 : ne pas être infecté

$p(A|B_1) = 0.99$, $p(A|B_2) = 0.01$

$p(B_1|A) = \frac{p(A|B_1)p(B_1)}{p(A|B_1)p(B_1)+p(A|B_2)p(B_2)}$

Il manque $p(B_1)$ et $p(B_2) = 1 - p(B_1)$. On va prendre $p(B_1) = 0.0001$ (valeur proche réalité : rapport INVS2009 1.710^{-4}).

$p(B_1|A) = \frac{0.99*0.0001}{0.99*0.0001+0.01*0.9999} = 0.0098$

Le fait d'être positif au test m'apprend suelement que j'ai beaucoup plus de chances que la population standard d'être infecté, mais mes chances réelles d'être infecté sont très faibles, parce que la majorité des cas ou le test est positif sont des cas ou il se trompe !

1.3 Analyse combinatoire

1.3.1 Permutations

DIAPOS 6-7 Tableau AA

On a une protéine de 8 AA de long. Je sais qu'elle est formée de 8 AA différents. Combien de protéines ? $8 \times 7 \times 6 \times ... = 8!$.

Si parmi les 8 AA j'en ai 6 différents et 2 identiques, pas de distinction entre les 2 identiques. Combien de protéines ? Même nb, mais on ne peut pas distinguer les positions des 2 : $\frac{8!}{2!}$.

Si parmi les 8 3 id : $\frac{8!}{3!}$

Si parmi les 8 2 couples de 2 identiques : $\frac{8!}{2! \times 2!}$.

1.3.2 Arrangements

On veut faire une protéine de 8 AA de long, et on doit choisir parmi 20 AA différents donnés au départ. Si 1 de chaque AA, $20 \times 19 \times 18 \times ... \times 13 = \frac{20!}{(20-8)!} = A_{20}^{8}$. On est dans un tirage sans répétitions.

Si un stock de chaque AA, $20 \times 20 \times 20... = 20^8$. On est dans un tirage avec répétitions.

1.3.3 Combinaisons

Une protéine de 8 AA de long, **tous différents**, se dégrade en solution, et les AA se retrouvent mélangés. On sat qu'au départ la protéine était composée de 8 AA parmi 20. Combien de solutions différentes ?

Le nb de protéines possibles est tjs A_{20}^{8}, mais il y en a bcp qu'on ne distingue pas une fois en solution : ce sont toutes celles formées des mêmes 8 AA dans le désordre. Pour une solution, il y a donc 8! protéines posssibles ; le nb de solutions est donc $\frac{A_{20}^{8}}{8!} = \frac{20!}{8!2!} = C_{20}^{8} = \binom{20}{8}$.

2 Variables aléatoires et lois de probabilités

Une variable aléatoire est le résultat d'un tirage probabiliste. C'est une variable qui peut prendre plusieurs valeurs, avec des probabilités données.

En biologie, on observe des caractères sur les individus : ce sont des grandeurs qui peuvent prendre plusieurs états ou *modalités*

En statistiques, on travaille avec des variables aléatoires : ce sont des variables qui peuvent prendre plusieurs *valeurs* avec une certaine probabilité

Caractère biologique (couleur)	$\Leftrightarrow$	Variable aléatoire X
État (bleu, vert, rouge)	$\Leftrightarrow$	valeur x de probabilité $p(X = x)$

Les variables quantitatives sont les variables que l'on peut mesurer explicitement : taille, poids, nombre de pattes...

Les variables qualitatives sont les variables pour lesquelles une mesure est difficile à produire, ou subjective : couleur, type de régime alimentaire, intensité de la douleur...

On distingue parfois :
— les variables quantitatives *discrètes*, ne pouvant prendre qu'un nombre fini de valeurs (par exemple le nombre de jambes d'un individu).
— les variables quantitatives *continues*, pouvant prendre un nombre infini de valeurs (par exemple la taille d'un individu).
Les variables quantitatives peuvent être distinguées par :
— leur espérance notée $\mathbb{E}(X)$ ou μ (valeur moyenne attendue). Une variable d'espérance 0 est dite *centrée*
— leur écart-type notée σ (variabilité attendue des résultats autour de la moyenne ; **exemple des notes des étudiants autour de 10**). Une variable d'écart-type 1 est dite *réduite*. On utilise souvent pour des raisons mathématiques σ^2 ou $\mathbb{V}(X)$, la variance.

2.1 Variables discrètes

La *loi de probabilité* d'une v.a. discrète est la probabilité de chaque résultat possible, notée $p(X = x)$. Si on lance 2 dés, la loi de probabilité de la somme S est :

s	$p(S=s)$
2	$\frac{1}{36}$
3	$\frac{2}{36}$
4	$\frac{3}{36}$
...	...

Pour $s = 4$, le calcul se fait ainsi : les combinaisons permettant d'obtenir 4 avec 2 dés sont 1 et 3, et 2 et 2. Mais une fois les dés jetés, on ne maitrise pas leur ordre : il y a donc 2 façons de faire 1 et 3, dans un ordre ou dans l'autre. Au total il y a donc 3 façons de faire 4 avec 2 dés :

$$p(S=4) = p(B=1 \cap R=3) + p(B=2 \cap R=2) + p(B=3/capR=1).$$

On a toujours, si les résultats possibles sont notés x_i avec $i = 1..N$, $\sum_{i=1}^{N} p(X=x_i) = 1$.

On a toujours

$$P(a < X < b) = \sum_{x=a}^{b} p(X=x).$$

La fonction de répartition d'une variable se déduit de sa loi de probabilité. C'est la probabilité cumulée d'observer n'importe quelle valeur en dessous d'un seuil donné. On a $F(x) = \sum_{y<x} p(X=y)$. Dans le cas des deux dés, on a :

s	$p(S=s)$	F(s)
2	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{2}{36}$	$\frac{3}{36}$
4	$\frac{3}{36}$	$\frac{6}{36}$
...	...	...

L'espérance d'une v.a. est la valeur moyenne d'un tirage de la v.a.

$$\mathbb{E}(X) = \sum_{x \in \Omega} xp(X=x).$$

La *variance* d'une v.a. représente la variabilité attendue des tirages autour de la valeur moyenne.

$$\mathbb{V}(X) = \left(\sum_{x\in\Omega} x^2 p(X=x)\right) - \mathbb{E}^2(X), \tag{1}$$

$$\text{ou } \mathbb{V}(X) = \sum_{x\in\Omega} \left(x - \mathbb{E}(X)\right)^2 p(X=x), \tag{2}$$

$$\text{ou } \mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X). \tag{3}$$

Attention au fait que la variance est mesurée dans de mauvaises unités ; on utilisera l'écart-type en pratique avec $\sigma = \sqrt{\mathbb{V}(X)}$.

2.2 Propriétés de l'espérance et de la variance

Proprietes de l'espérance (propriétes valables pour les 2 meme si la demo est en discret)

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) = \textstyle\sum_{x\in A}\sum_{y\in B}((x+y))p(X=x)p(Y=y)), \tag{4}$$

$$= \textstyle\sum_{x\in A}\sum_{y\in B} xp(X=x)p(Y=y) + \sum_{x\in A}\sum_{y\in B} yp(X=x)p(Y=y), \tag{5}$$

$$= \textstyle\sum_{x\in A} xp(X=x)\sum_{y\in B} p(Y=y) + \sum_{x\in A} p(X=x)\sum_{y\in B} yp(Y=y), \tag{6}$$

$$= \mathbb{E}(X) + \mathbb{E}(Y). \tag{7}$$

De la meme manière

$$\mathbb{E}(aX+b) = a\mathbb{E}(X) + b.$$

car $\mathbb{E}(b) = b$. Exemple changement d'unites entre degres Celsius et Fahrenheit.Celsius = 1.8 Fahrenheit+32

Une variable dont l'espérance vaut 0 est dite *variable centrée*. D'après la propriété ci-dessus, si $Y = X - \mathbb{E}(X)$, $\mathbb{E}(Y) = 0$ et Y est centrée.

Proprietes de la variance

$$V(aX+b) = a^2 V(X)$$

$$V(aX) = \left(\textstyle\sum_{x\in\Omega}(ax)^2 p(X=x)\right) - \mathbb{E}^2(aX), \tag{8}$$

$$= \left(\textstyle\sum_{x\in\Omega} a^2x^2 p(X=x)\right) - a^2\mathbb{E}^2(X), \tag{9}$$

$$= a^2\left(\textstyle\sum_{x\in\Omega}(x)^2 p(X=x)\right) - \mathbb{E}^2(X), \tag{10}$$

$$= a^2\mathbb{V}(X). \tag{11}$$

Une autre propriété utile est :

$$\mathbb{V}(X) = 0 \text{ ssi } X = \mathbb{E}(X).$$

Une variable dont la variance vaut 1 est dite *variable réduite*. D'après la propriété ci-dessus, si $Y = \frac{X}{\sqrt{\mathbb{V}(X)}}$, $\mathbb{V}(Y) = 1$ et Y est réduite.

De plus, si X et Y sont deux variables aléatoires indépendantes ($p(X|Y) = p(X)$), alors on a en plus :

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \quad \text{et} \quad \mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y)$$

2.2.1 Bernouilli

La *Loi de Bernouilli* est la loi d'une v.a. correspondant à un tirage à 2 issues possibles 1 et 0, de probabilités respectives p et $1-p$.

Exemples : pile ou face, probabilité de rencontre d'un partenaire, par jour, lors de la période de reproduction, probabilité d'infection, par jour, par un virus...

$$p(X=0) = 1-p, \;\; p(X=1) = p \tag{12}$$
$$\mathbb{E}(X) = 0 \times (1-p) + 1 \times p = p \tag{13}$$
$$\mathbb{V}(X) = p(1-p). \tag{14}$$

DIAPO 8-9

2.2.2 Binomiale

La *Loi binomiale* est la loi d'une v.a. correspondant au nombre de succès lors du tirage de n variables de Bernouilli indépendantes. On la note souvent $\mathcal{B}(n,p)$.

Par exemple : nombre moyen de partenaires rencontrés lors de la période de reproduction.

$$p(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{15}$$
$$\mathbb{E}(X) = np \tag{16}$$
$$\mathbb{V}(X) = np(1-p). \tag{17}$$

DIAPO 10

2.2.3 Géométrique

La *Loi géométrique* est la loi d'une v.a. correspondant au nombre de tirages nécéssaires pour avoir le premier succès lors du tirage de variable de Bernouilli indépendantes.

Par exemple : nombre de jours avant d'etre infecté par un virus auquel vous êtes potentiellement exposé chaque jour.

$$p(X = n) = (1 - p)^{n-1}p \tag{18}$$

$$\mathbb{E}(X) = \frac{1}{p} \tag{19}$$

$$\mathbb{V}(X) = \frac{1 - p}{p^2}. \tag{20}$$

DIAPO 11

2.2.4 Binomiale négative

La *Loi binomiale négative* est la loi d'une v.a. correspondant au nombre de tirages nécéssaires pour avoir les k premiers succès lors du tirage de variable de Bernouilli indépendantes.

Exemple : Sur un génome, nombre de bases à parcourir pour rencontrer une répétition de k fois la même base.

$$p(X = n) = \binom{n-1}{k-1}(1 - p)^{n-k}p^k \tag{21}$$

$$\mathbb{E}(X) = \frac{k}{p} \tag{22}$$

$$\mathbb{V}(X) = \frac{k(1 - p)}{p^2}. \tag{23}$$

DIAPO 12

2.2.5 Poisson

La *loi de Poisson* est la loi d'une v.a. correspondant au nombre d'evenements indépendants qui se produisent dans un intervalle donné, si leur fréquence est constante et connue (on la note λ). On la note souvent $\mathcal{P}(\lambda)$.

Vient de Siméon-Denis Poisson.
Exples : mutations, fréquence de passage d'un individu à un endroit précis.

$$p(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{24}$$
$$\mathbb{E}(X) = \lambda \tag{25}$$
$$\mathbb{V}(X) = \lambda. \tag{26}$$

DIAPO 13

2.3 Variables continues

$$f(x) = \frac{p(x)}{\Delta x},$$

avec Δx le pas que l'on voit.

DIAPOS 14-15 Passage continu

La loi de probabilité d'une v.a. continue est donnée par sa *densité de probabilité*. Comme vu avant pour une variable continue, $p(X = x) = 0$; on ne peut pas utiliser le formalisme du cas discret. La densité f associée à la variable aléatoire X est la probabilité de tirer une valeur dans un intervalle tout petit autour de x. On a toujours :

$$f(x) \geq 0 \quad \int_\Omega f(x) = 1.$$

On a toujours

$$P(a < X < b) = \int_a^b f(x)dx.$$

La fonction de répartition se calcule par intégration (dessin loi normale avec 10%, 50%...) :

DIAPO 16 Repartition loi continue

$$F(t) = P(x < t) = \int_{-\infty}^t f(x)dx$$

Noter le parallele entre les formules discretes et continues.

Dans le cas continu, on a :

$$\mathbb{E}(X) = \int_\Omega xp(x)dx, \tag{27}$$
$$\mathbb{V}(X) = \int_\Omega x^2 p(x)dx - \mathbb{E}^2(X), \tag{28}$$
$$\text{ou } \mathbb{V}(X) = \int_\Omega (x - \mathbb{E}(X))^2 p(x)dx, \tag{29}$$
$$\text{ou } \mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X). \tag{30}$$

2.3.1 Uniforme

La loi uniforme sur un intervalle $[a, b]$ est la loi d'un v.a. qui a autant de chance de prendre chaque valeur entre a et b.

Par exemple :

$$f(x) = \frac{1}{b-a} \tag{31}$$
$$\mathbb{E}(X) = \frac{a+b}{2} \tag{32}$$
$$\mathbb{V}(X) = \frac{(b-a)^2}{12}. \tag{33}$$

DIAPO 17

2.3.2 Loi normale

La loi normale ou de Gauss est la loi de probabilité des variables aléatoires continues dépendantes d'un grand nombre de causes indépendantes et additives. Elle se note $\mathcal{N}(\mu, \sigma)$ avec μ l'espérance de la loi et σ l'écart-type. Attention à la notation de l'écart-type.

Par exemple :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{34}$$
$$\mathbb{E}(X) = \mu \tag{35}$$
$$\mathbb{V}(X) = \sigma^2. \tag{36}$$

DIAPOS 18-19 avec table

2.3.3 Student

La loi de Student est une variante de la loi normale que l'on observe quand la variance de la variable étudiée est incconnue. Sa définition mathématique est le rapport entre une variable normale centrés réduite et la racine d'un χ^2.

Par exemple : loi rencontrée lors des tests de comparaison de moyenne.

$$t = \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_n}{n}}} \tag{37}$$

$$\mathbb{E}(X) = 0 \tag{38}$$

$$\mathbb{V}(X) = \frac{n}{n-2}. \tag{39}$$

DIAPO 20

3 Convergences

3.1 TCL et Convergences

Un énoncé du théorème central limite (TCL) est : *Toute somme de n variables aléatoires indépendantes converge vers une loi normale avec n.* On prend n v.a. indépendantes $X_1, ...X_n$ et on les somme. On sait que :

$$\mathbb{E}(Y = \sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathbb{E}(X_i), \tag{40}$$

$$\mathbb{V}(Y = \sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathbb{V}(X_i). \tag{41}$$

Le TCL dit que si n est assez grand on peut approximer Y par une $\mathcal{N}(\sum_{i=1}^{n} \mathbb{E}(X_i), \sum_{i=1}^{n} \mathbb{V}(X_i))$.

On en déduit que la loi de la moyenne d'un échantillon est une loi normale.

DIAPO 21 convergence moyenne

Deux autres convergence sont à connaitre : une variable binomiale peut (via le TCL) etre approximée par une loi normale de même moyenne et variance quand n devient grand (attention, discret vers continu). On note :

$$\mathcal{B}(n,p) \to \mathcal{N}(np, \sqrt{np(1-p)})$$

Si la moyenne de la loi binomiale est trop grande ou trop petite, la loi normale est une bonne approximation quand n devient extremement grand et donc la loi normale assez fine pour ne plus voir le bord ; si cette approximation ne fonctionne pas on utilise l'approximation de la binomiale par une loi de Poisson (discret vers discret) :

$$\mathcal{B}(n,p) \to \mathcal{P}(np)$$

DIAPO 22 convergence binomiale

4 Statistiques descriptives

4.1 Définitions

Une *population* est l'ensemble des individus auxquels on veut appliquer les résultats d'une étude. (orogine démographique du vocabulaire)

Un *Échantillon* est un sous-ensemble de la population, choisi par une méthode d'*échantillonage*.

On peut employer différents types d'échantillonage. L'échantillonage *aléatoire simple* consiste à prendre au hasard et de manière indépendante des individus dans la population (ctre-exple : sondage internet). De nombreuses propriétés statistiques inférées sur la population dépendent du fait que l'échantillonage soit aléatoire.

Les échantillonages non aléatoires peuvent inclure par exemple l'échantillonage par boule de neige, ou un individu en recrute d'autre pour répondre. Ou les échantillonages de sondages téléphoniques, qui incluent différentes tranches d'age, de sexe, de manière fixée à l'avance.

Si les données ne proviennent pas d'un échantillonnage aléatoire simple, elle peuvent être autocorrélées (données temporelles).

La taille de l'échantillon est le nombre de données composant la série statistique.

4.2 Les données statistiques

4.2.1 Données qualitatives ou quantitatives discrètes

Rappel lien entre données stat et v.a en proba.

Si on a des données quali ou discrètes, on va les donner sous la forme d'une *table de contingence* qui compte le nombre de fois ou chaque modalité x_i apparait dans les données. On va avoir des données du type $n_i, i \in [1..k]$ avec $\sum_{i=1}^{k} n_i = N$ On va pouvoir calculer pour chaque modalité sa *fréquence* $f_i = \frac{n_i}{N}$.

Pour les caractères quanti discrets et quali ordinaux (avec un ordre) on peut calculer la fréquence cumulée, qui représente la fraction des données pour laquelle le caractère est inferieur ou égal à une valeur x_j :

$$f_i^{cum} = \sum_{j=1}^{i} f_j$$

DIAPOS 23-24 exple balanins

4.2.2 Données continues

Dans le cas de données continues, on peut garder les données x_i telles quelles. Il est parfois utile, pour résumer l'information, de les regrouper en classes, c-a-d de rassembler les données proches et de les considérer comme identiques.

DIAPOS 25 exple mangues

Chaque classe a une *amplitude* particulière que l'on peut choisir – on les prend souvent égales, par commodité et pour faciliter les comparaisons entre classes (12 personnes entre 1m60 et 1m70 et 17 personnes entre 1m90 et 2m30...). L'utilisateur peut choisir le nombre de classes : plus il y en a plus on est précis mais moins on résume l'information.

Une fois les données classées on peut calculer la fréquence et la fréquence cumulée de chaque classe de la même manière que pour les données discrètes.

Finalement, si on a des données quantitatives continues que l'on ne veut pas regrouper, on garde en l'état : c'est le cas le plus courant.

4.3 Représentations graphiques

On parle ici des représentations graphiques pour une ou deux variables jointes.

4.3.1 Variable unique

DIAPOS 26 Une seule variable discrete ou qualitative : diagramme en batons

DIAPOS 27 histogramme

4.3.2 Deux variables

DIAPOS 28 pirates

DIAPOS 29 boxplot quali+quanti

DIAPOS 30 mosaicplot

4.4 Indicateurs statistiques

Soit X la variable mesurée et x_i la mesure de X pour l'individu i de l'échantillon ($i = 1..n$). On appelle *série statistique* l'ensemble des valeurs :

$$x_1, x_2, x_3, \ldots, x_i, \ldots, x_{n-1}, x_n$$

On peut resumer cette série par des indices de :

position : quel est l'ordre de grandeur de la série dans son ensemble ?
moyenne, *médiane*

dispersion : quelle est la variabilité des valeurs de la série par rapport à l'indice de position ?
variance, *écart-type*

Données simples :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{42}$$

Données groupées :

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} n_j x_j^\star, \text{ avec } n = \sum_{j=1}^{k} n_j \tag{43}$$

DIAPO 31 mangue avec moyenne groupée

La *médiane* est la valeur de la série pour laquelle 50% des valeurs sont plus grandes, et 50% plus faibles.

Si n impair, $n = 2m + 1$ et la médiane est la valeur x_m

Si n pair, $n = 2m$ et la médiane est $\dfrac{x_m + x_{m+1}}{2}$

L'intérêt de la médiane est qu'elle peut donner une idée différente du "cas typique" qu'on s'attend à observer.

DIAPOS 32-32 Salaire moyen France 1500/2400

Le *mode* d'une série statistique ne peut se calculer que si les données sont regroupées en classe. Il s'agit de la valeur moyenne de la classe de plus grand effectif, et cette valeur dépend donc des classes qui ont été choisies.

DIAPO 33 Mangue mode

La *variance observée* s^2 d'une série statistique se calcule ainsi :

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 , \tag{44}$$

ou, de manière équivalente,

$$s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 \tag{45}$$

L'*écart-type* s se calcule à partie de la variance et vaut :

$$s = \sqrt{s^2} \tag{46}$$

La variance sur des données groupées se calcule ainsi :

$$s^2 = \frac{1}{n}\sum_{j=1}^{k} n_j \left(x_j^\star - \bar{x}\right)^2, \tag{47}$$

ou encore

$$s^2 = \frac{1}{n}\left(\sum_{j=1}^{k} n_j \left(x_j^\star\right)^2\right) - \bar{x}^2. \tag{48}$$

DIAPO 34 mangue avec variance

5 Estimation et intervalles de confiance

On rappelle qu'un échantillon est une sous-partie de la population étudiée.

L'objectif de l'*inférence* statistique consiste à trouver les valeurs de certaines caractéristiques de la population, à partir de celles observées dans l'échantillon.

Quand on veut la valeur numérique d'un paramètre, on parle d'*estimation.*

Une remarque importante est que l'inférence statistique ne dit pas si les choses sont ou ne sont pas dans la population, mais elle donne une probabilité à différents évenements, ou une probabilité à la valeur de certains paramètres.

5.1 Estimation ponctuelle

Dans ce cours, on peut vouloir estimer 3 paramètres dans une pop : μ, σ^2 et p.

On veut mesurer l'épaisseur moyenne de la couche de gaisse, notée μ chez les ours blancs du Groenland. On prend un échantillon de n ours blancs pour lesquels on veut mesurer l'épaisseur moyenne de la couche de graisse. On a une série statistique $x_1, ...x_n$. On peut calculer la moyenne de cette série $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

On dit que $\bar{x}$ est un estimateur de μ.

On peut montrer que $\bar{x} \to \mu$ quand $n \to \infty$: on dit que la moyenne empirique (observée) $\bar{x}$ est un estimateur non biaisé de μ. La meilleure estimation ponctuelle de μ que l'on puisse faire à partir des $x_1...x_n$ est $\hat{\mu} = \bar{x}$.

De la même manière, si on veut estimer la fréquence d'occurence d'un caractère comme un allèle particulier, on va compter, sur n ours, combien ont cet allèle. On note ce nombre k. Dans la population, la vraie probabilité d'avoir l'allèle en question est p ; on peut montrer que $f = \frac{k}{n}$ est un estimateur non biaisé de p. On note $\hat{p} = f = \frac{k}{n}$

En ce qui concerne l'estimation de la variance σ^2 de l'épaisseur de la graisse des ours, un léger problème se pose. L'estimateur naturel serait $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$; mais cet estimateur est biaisé et sous-estime la variance globale dans la population (car on rate forcement les individus les plus extremes si on en prend peu). Il faut corriger cet estimateur ; un estimateur non biaisé de la variance de la population est $\hat{\sigma^2} = \frac{n}{n-1}s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Bien faire la différence entre σ^2, variance de la pop, son estimateur $\hat{\sigma^2}$ et la variance observée s^2.

5.2 Distribution d'échantillonage

A partir d'une population, on prend généralement un échantillon aléatoire. On pourrait en prendre plusieurs ; ils seraient différents, et les valeurs des variables mesurées dans chaque échantillon ne seront pas les même, et ne seront pas strictement identiques à celles de la population (sauf constance). On parle de *distribution d'échantillonage* d'une variable. Grace aux probabilités, on peut calculer cette distribution.

DIAPOS 35-36 distro moyenne avec n variable mais tjs grand

Que vaut cette distribution d'échantillonage ? Prenons le cas de la moyenne d'un grand échantillon. On a vu avec le TCL que la somme d'un grand nombre de v.a quelconques suit une loi normale. En particulier, si chacun des x_i est indépendant et vient d'une population qui suit une loi d'espérance μ et de variance σ^2, la moyenne observée est une v.a. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et aura :

- une espérance de $\mathbb{E}(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X) = \mu$
- une variance de $\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X) = \frac{\sigma^2}{n}$
- suivra une loi normale, car elle est la somme d'une très grand nombre de variables indépendantes.

DIAPO 37 distro échantillonage

DIAPOS 38-39-40-41 repartition des valeurs autour de mu et sigma : 95%, 90%, 99%

Donc la notion de taille d'intervalle pour un risque donne de se planter.

Mathématiquement, on écrit, que, si on prend un risque α de se tromper :

$$P(\mu - C_\alpha < \bar{x} < \mu + C_\alpha) = 1 - \alpha$$
$$P(\mu - \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}} < \bar{x} < \mu + \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}}) = 1 - \alpha$$
$$P(\epsilon_\alpha < \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} < \epsilon_\alpha) = 1 - \alpha$$

Or $\frac{\bar{x}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$ est une v.a. centrée réduite qui suit une loi normale comme $\bar{x}$, car c'est un transformation linéaire d'une v.a. normale ; ϵ_α est donc indépendant de μ et σ, et dépend uniquement de α :

$$P(\epsilon_\alpha < \mathcal{N}(0,1) < \epsilon_\alpha) = 1 - \alpha$$

DIAPO 42 loi normale centrée réduite

Ces valeurs seront les memes pour tous les problemes ou on se ramenera a une loi normale centree reduite, et donc tous les problemes ou on aura les memes hypotheses au depart. On les lit dans des tables (voir TD). Les plus couramment utilisees sont $\epsilon_{0.05} = 1.96$.

On a donc au final :

$$P(\mu - \epsilon_\alpha \sqrt{\frac{\sigma^2}{n}} < \bar{x} < \mu + \epsilon_\alpha \sqrt{\frac{\sigma^2}{n}}) = 1 - \alpha.$$

5.3 Estimation par intervalle de confiance

L'idée de l'estimation par intervalle de confiance est d'associer à l'estimation ponctuelle la connaissance que l'on a sur la distribution d'échantillonnage. En fonction des situations et des hyppotheses, on peut avoir une idee plus ou moins precise de la distribution d'échantillonnage, et donc un intervalle de confiance plus ou moins précis.

On construit l'IC au risque α de se tromper en regardant l'intervalle qu'on s'autorise à avoir dans la distro d'échantillonnage au risque α, et en appliquant cet intervalle autour de la valeur estimée. Dans le cas précédent :

Hyp : moyenne d'une v.a. quelconque, $n > 30$ grand, l'IC va etre de la forme :

$$IC : [\bar{x} - \epsilon_\alpha \sqrt{\frac{\sigma^2}{n}}, \bar{x} + \epsilon_\alpha \sqrt{\frac{\sigma^2}{n}}]$$

On peut construire les IC pour d'autres hypthèses.

Hyp : Si X suit une loi normale de variance connue, alors $\forall n$ la moyenne observée suivra aussi une loi normale, même si n est petit dans ce cas : c'est la normalité de X qui se conserve :

$$IC : [\bar{x} - \epsilon_\alpha \sqrt{\frac{\sigma^2}{n}}, \bar{x} + \epsilon_\alpha \sqrt{\frac{\sigma^2}{n}}]$$

Hyp : Si X suit une loi normale de variance inconnue – bien définir une variance inconnue – et que $n > 30$ environ, on fait une petit erreur car on

doit estimer la variance ; la loi sous-jacente n'est plus une loi nomrale mais une loi de Student, et l'IC devient :

$$IC : [\bar{x} - t_{\alpha,n-1}\sqrt{\frac{\hat{\sigma^2}}{n}}, \bar{x} + t_{\alpha,n-1}\sqrt{\frac{\hat{\sigma^2}}{n}}]$$

Hyp : Finalement si X suit une loi inconnue, et l'échantillon est petit, on ne connait pas bien la distribution d'échantillonage, et on ne peut pas faire d'IC avec les lois connues en première année – on travaillera avec d'autres lois plus complexes mais plus générales, les statistiques non paramétriques.

Pour ce qui est de l'estimation de la variance on ne donne en général pas d'IC, même s'il est en théorie possible d'en construire un.

6 Tests

6.1 Raisonnement général des tests statistiques

A cause des variations d'échantillonage, plusieurs échantillons pris dans la même population n'auront pas les mêmes caractéristiques. Taille moyenne des femmes en France, entre 20 et 29 ans : 164.6 cm

Si on prend différents échantillons de taille $n = 40$ dans cette population, on peut trouver une moyenne de :

— 178.4 cm à l'ASVEL Basket ($n = 40$)
— 168.6 cm dans une agence de communication ($n = 40$)
— 164.9 cm à l'université de Lyon
— 164.1 cm en prenant 40 femmes complètement au hasard en France
— 162.7 cm à l'université de Nice

Idée générique : si l'écart observé entre $\bar{x}$ et μ_0, moyenne de la population, est petit, on va dire que l'erreur est due au hasard ; si l'écart est grand on va dire que le hasard ne suffit pas. Cet écart va être calculé sous al formce de ce qu'on appelle la *statistique* du test.

Peut-on en conclure à chaque fois que les différents échantillons sont *significativement différents* de la moyenne francaise ? L'objectif d'un test statistique est de répondre à cette question, en distinguant entre 2 hypothèses :

H_0 l'hypothèse nulle : la sous-population de laquelle provienne ces femmes a la même taille moyenne que la population francaise : la différence observée entre $\bar{x}$ et μ vient uniquement de la variabilité de la distribution d'échantillonage, donc du hasard. Mathématiquement on a $\mu = \mu_0$. Attention, $\bar{x} = \mu_0$ ne veut rien dire !

H_1 l'hypothèse alternative : la sous-population de laquelle provienne ces femmes a une taille moyenne différente de celle de la population francaise : il existe une différence réelle entre μ_0, la moyenne globale de la population, et μ, la moyenne de la sous-population de laquelle provient l'échantillon. Mathématiquement on a $\mu \neq \mu_0$. Attention, $\bar{x} \neq \mu_0$ ne veut rien dire !

Il faut remarquer que H_0 est structurellement plus simple que H_1, puisque H_0 implique qu'un seul paramètre décrit la population, alors que H_1 implique l'existance d'un deuxième paramètre. On dit que H_0 est l'hypothèse nulle parce que c'est celle que l'on va priviliégier (la plus simple) sauf si les données disent le contraire.

Notion Rasoir d'Occam.

Logique des tests : les tests fonctionnent au rejet. Il faut se rappeler que $A \rightarrow B$ est équivalent à $nonB \rightarrow nonA$, mais pas du tout à $B \rightarrow A$. Exemple avec B : je mange toujours des céréales au petit déjeuner, et A : je suis un poulet. On a tjs $A \rightarrow B$. Si on ne mange pas de céréales ($nonB$), on peut en conclure que l'on n'est pas un poulet (si on en était un il faudrait qu'on mange des céréales). Mais on ne peut pas en conclure que manger toujours des céréales au petit dejeuner implique que vous etes un poulet, puisque d'autres choses que les poulets peuvent manger la meme chose qu'eux.

Pour chacun des échantillons ci-dessus, je peux réaliser un test statistique, qui va se baser sur l'assertion logique suivante : $H_0 \rightarrow$ statistique $\in []$.

On va donc calculer la statistique.

1. Si elle est dans l'intervalle, on a $nonB \rightarrow nonA$ et H_0 est fausse : on rejette H_0 si la statistique est forte.
2. Si la stat est hors de l'intervalle, on ne peut pas en conclure logiquement que H_0 est vraie ; on va l'accepter par défaut, et parce que c'est l'hypothèse *la plus simple*.

La difficulté réside dans le fait qu' il n'y a pas de limite précise à l'intervalle qui nous intéresse, à cause des propriétés des lois statistiques qui

nous intéressent : une loi normale peut donner n'importe quelle valeur entre $-\infty$ et ∞. Prenons le cas précédent, et l'échantillon des joueuses de l'ASVEL. On sait que la taille des femmes dans la pop. a une moyenne de μ =164.6cm, et un écart-type de σ =5cm. On peut donc calculer la distribution d'échantillonnage de la moyenne attendue de $n = 40$ individus dans cette pop ; ce sera une $\mathcal{N}(164.6, \sqrt{\frac{5^2}{40}})$. La question est ensuite : ou se trouve la valeur $\bar{x}$ sur ce graphe ? Quelle est la probabilité que $\bar{x}$ soit aussi éloigné de μ ?

DIAPO 43-44 comparaison xbarre et seuils

On voit qu'en fonction du seuil de précision que l'on choisit, $\bar{x}$ est d'un cote ou de l'autre. Si on veut quelque chose de strictement exact, on peut dire que :

- Toute valeur de $\bar{x}$ est possible si la distribution d'échantillonnage est une loi normale
- Il est très rare d'obtenir "exactement" μ quand on calcule $\bar{x}$ sur un échantillon.

Comme précedemment, on va se ramener à une loi normale centré réduite. On a :

DIAPO 45 equivalence seuils

$$P(\bar{x}|\mathcal{N}(\mu, \sqrt{\frac{\sigma^2}{n}})) = P(\frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}|\mathcal{N}(0, 1))$$

La question qui se pose est donc : si je prends comme hypothèse qu'un résultat au hasard doit tomber dans les $1-\alpha$ pour cents des résultats les plus probables, $\bar{x}$ est-il dans cet intervalle ? On pourrait répondre à cette question en passant par des IC, et en se demandant si μ est dans l'IC autour de $\bar{x}$. On peut aisément retourner le raisonnement, et mathématiquement, cela revient à demander si :

$$-\epsilon_\alpha \leq \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \leq \epsilon_\alpha \text{ ou bien } \left|\frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}\right| \leq \epsilon_\alpha$$

Le fait qu'on aie ici un ϵ_α vient du fait qu'on est dans un cas ou la distro d'échantillonnage est une loi normale ; si ce n'était pas le cas, on aurait un autre seuil, et on va les détailler dans la suite. La limite exacte dépend de la distribution d'échantillonnage et donc des hypothèses que l'on peut faire

sur les données. Dans ce cas précis, la statistique, notée ϵ_{obs}, va etre $\left|\frac{\bar{x}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}}\right|$ et le seuil ϵ_α : si $\epsilon_{obs} > \epsilon_\alpha$, on va rejeter H_0 et accepter H_1 ; si $\epsilon_{obs} < \epsilon_\alpha$, on conservera H_0.

Quand on effectue un test statistique avec un seuil choisi au risque α, on dit en pratique que si $\bar{x}$ appartient aux α pour cents de la distribution d'échantillonage les plus rares, on va rejeter H_0 comme étant fausse. Cette assertion est par définition fausse dans α pour cents des cas. On prend donc un risque α de se tromper, dit *risque de première espèce*. si on veut réduire ce risque, on va augmenter ϵ_α jusqu'à ce qu'on finisse toujours par conclure que H_0 est vraie par défaut : la seule chose dont on peut etre certain si on ne veut pas se tromper, c'est l'hypothèse qu'on formule au départ. Dans la pratique, si on rejette H_0 avec un risque 1/1000 000 de se tromper, tout se passe bien. On prend en général comme seuil $\alpha = 0.05$.

Il existe un autre risque de se tromper : c'est celui ou on conserve H_0 par défaut alors que H_0 était fausse. C'est le cas ou H_1 est vraie, mais peut-etre pas tres différente de H_0, et donc on ne voit pas bien la différence. On note ce risque de *deuxième espèce* β.

DIAPOS 46-47-48-49 exemple beta

Tableau recap risques

Réalité	H_0	H_1
Choix		
H_0	1-α	β
H_1	α	$1-\beta$

Exemple : si on choisit tjs H_0 sans regarder les données, alors $\alpha = 0$, mais $\beta = 1$. ã l'inverse, si on choisit tjs H_1, alors $\beta = 0$ mais $\alpha = 1$. On sait que pour chaque test, $\exists c$ t.q $\alpha + \beta \geq c$.

Un mot sur la latéralité : si je m'intéresse à une hypothèse biologique unilatérale (un médicament par exemple), je vaid changer mon seuil à l'avance, et décider que je ne considererai que les effets par exemple positifs. Dans ce cas, pour conserver le meme risque, il faut que je prenne un ϵ_α différent ; vu que la loi est symétrique, il faut que je prenne pour un test unilatéral un seuil $\epsilon_{2\alpha}$.

DIAPOS 50 tests unilatéraux

6.2 Test de comparaison de moyennes

6.2.1 Comparaison à une moyenne théorique

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Hyp : Variance connue, X suit une loi normale : c'est le cas précédent :

$$\epsilon_{obs} = \left| \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \right| \text{ avec un seuil } \epsilon_\alpha$$

Hyp : X normale, variance estimée $\hat{\sigma}^2 = \frac{n}{n-1} s^2$: l'erreur sur la variance nous ferait faire une erreur si on employait la formule précédente. On va utiliser une loi de Student :

$$t_{obs} = \left| \frac{\bar{x} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \right| \text{ avec un seuil } t_{\alpha,n-1}$$

En pratique, si n grand, on peut remarquer dans les tables stats que $t_{\alpha,n-1}$ tend vers ϵ_{alpha}. Donc quand n est grand, même si la variance est estimée, on pourra employer la table ϵ (mais on ne fait pas d'erreur si on emploie la table t : les valeurs sont les mêmes !).

Hyp : X loi inconnue, mais n grand : le TCL nous dit que $\bar{x}$ suit une loi normale, donc on peut utiliser le paragraphe précédent, et comme n grand, on a la convergence de t vers ϵ.

Hyp : Variable X non normale, n petit : comme pour l'IC on ne peut rien faire, il faudrait employer des statitstiques non paramétriques

DIAPOS 51 récap tests conformité

6.2.2 Comparaison de moyennes observées

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Hyp : 2 grands échantillons : on va employer le TCL pour dire que les deux moyennes suivent des lois normales, et on calcule :

$$\epsilon_{obs} = \left| \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n_X} + \frac{\hat{\sigma}_Y^2}{n_Y}}} \right| \text{ avec un seuil } \epsilon_\alpha$$

Hyp : Au moins un des 2 échantillon est petit, les 2 variances sont inconnues. les variables sont normales. En théorie, on ne peut comparer ces moyennes que si les variances sont égales. Ne le sachant pas (et ne sachant pas comment faire le test correspondant en 1ère année), on va le supposer et calculer l'estimateur de la variance commune :

$$\hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}.$$

Ensuite, on va dire que X étant normale, la statistique va presque l'être ; la variance étant (mal) estimée, on va faire une erreur et la statistique ne suivra pas une loi normale mais une loi de Student (on voit que le dénominateur et le numérateur changent quand on change l'échantillon). La statistique est :

$$t_{obs} = \left| \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \right|$$

Et on la compare à t_{α, n_1+n_2-2} comme valeur seuil.

Hyp : Au moins un des 2 échantillon est petit, les 2 variances sont inconnues. Les variables ne sont pas normales, ou bien les variances sont différentes. On doit faire du non paramétrique car oin ne peut pas calculer la distribution de la statistique.

DIAPOS 52 récap tests homogénéité
DIAPOS 53-54 marmotte

6.3 Test du χ^2

6.3.1 χ^2 d'ajustement

On a une table de contingence pour une variable X. On se demande si les comptes observés pour chaque intervalle ou valeur de X suivent un loi donnée $p(X)$.

$$H_0 : X \text{ suit la loi } p(X)$$
$$H_1 : X \text{ ne suit pas la loi } p(X)$$

Les étapes consistent en :
- Calculer les effectifs théoriques attendus si H_0 est vraie
- Regrouper les catégories pour que les effectifs théoriques vaillent au moins 5 (en réalité ne soient pas trop petits, on utilise une convergence vers la normale et pas la Poisson)
- Calculer la statistique qui mesure la différence entre effectifs théoriques et effetifs observés
- Conclure en comparant à la valeur théorique que l'on a pu tabuler si H_0 était vraie.

X	X_1	X_2	$\dots$	X_k	Total
Effectifs observés O_i	n_1	n_2	$\dots$	n_k	n
Effectifs théoriques T_i	$np(X = X_1)$	$np(X = X_2)$	$\dots$	$np(X = X_k)$	n

La statistique est ensuite la suivante :

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(O_i - T_i)^2}{T_i}$$

Cette statistique s'appelle χ^2 car chaque terme dans la somme est le carré d'une loi normale centrée réduite si H_0 est vraie, et la somme de k carrés de loi normales centrées réduites suit un χ^2_k (ici un χ^2_{k-1} car la donnée des $k-1$ premieres valeurs détermine la dernière).

On va comparer cette statistique à une valeur seuil χ^2_{k-1-c} à $k-1-c$ ddl, où c est le nombre de paramètres estimés. En effet chaque paramètre estimé à partir des données donne artificiellement un meilleur ajustement, on enlève donc un ddl pour compenser ce biais.

Ajustement à une loi de Poisson (données discrètes) Pour cet ajustement, la probabilité $p(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$. Pour trouver λ on va l'estimer à partir des données ; λ étant l'espérance de la loi de Poisson, on va calculer $\bar{x}$ et dire que $\lambda = \bar{x}$.

DIAPOS 55-56-57 balanins

Ajustement à une loi uniforme (données continues) Il n'est pas nécessaire d'estimer de paramètre ici ; si les données sont réparties entre x_{min} et x_{max}, la probabilité d'être dans un intervalle donné $[x_a, x_b]$ est $p = \frac{x_b - x_a}{x_{max} - x_{min}}$.

Ajustement à une loi normale (données continues) Pour cet ajustement, on va commencer par estimer à partir des données les 2 paramètres $\hat{\mu} = \bar{x}$ et $\hat{\sigma}^2 = \frac{n}{n-1}s^2$. Puis on va calculer la probabilité pour chaque valeur d'être dans chaque intervalle $[x_a, x_b]$, comme suit :

$$\begin{aligned}
P(x_a < \mathcal{N}(\mu, \sigma) < x_b) =& P(x_a - \mu < \mathcal{N}(0, \sigma) < x_b - \mu) \\
=& P(\frac{x_a - \mu}{\sigma} < \mathcal{N}(0, 1) < \frac{x_b - \mu}{\sigma}) \\
=& P\left(\mathcal{N}(0, 1) < \frac{x_b - \mu}{\sigma}\right) - P\left(\frac{x_a - \mu}{\sigma} < \mathcal{N}(0, 1)\right) \\
=& P\left(\mathcal{N}(0, 1) < \frac{x_b - \mu}{\sigma}\right) - \left(1 - P\left(\mathcal{N}(0, 1) < \frac{x_a - \mu}{\sigma}\right)\right)
\end{aligned}$$

Les valeurs sur la dernière ligne sont celles qui sont tabulées, et on peut donc faire le calcul pour chaque intervalle.