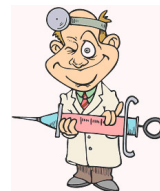


MAPSI — cours 5 : Tests d'indépendance

Christophe Gonzales

LIP6 – Université Paris 6, France

Motivations : monitoring de patients



37 variables aléatoires
 $> 10^{16}$ événements élémentaires !

BD suffisamment grande pour estimer les paramètres de la loi jointe par max de vraisemblance ?

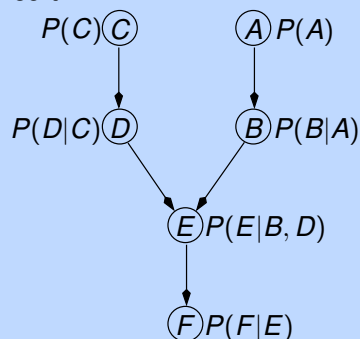
⚠ on peut encoder la loi jointe avec seulement 752 paramètres !

Solution : décomposer la loi jointe en produit de probas conditionnelles

Motivations : réseaux bayésiens

Définition d'un réseau bayésien

1 un graphe sans circuit :

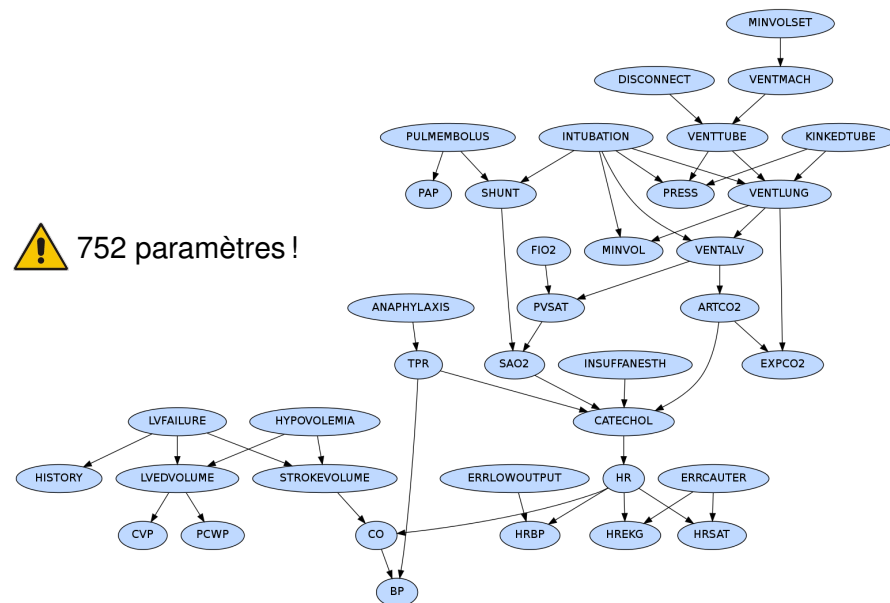


qui représente une décomposition de la loi jointe :
 $P(A, B, C, D, E, F) = P(F|E)P(E|B, D)P(D|C)P(C)P(B|A)P(A)$

2 À chaque noeud X du graphe est associé sa probabilité conditionnellement à ses parents.

Motivations : monitoring de patients

⚠ 752 paramètres !



Rappel sur l'indépendance conditionnelle

Indépendance conditionnelle de deux variables discrètes

X et Y sont *indépendantes* conditionnellement à Z si :
si $P(Y|Z) > 0$ alors $P(X|Y, Z) = P(X|Z)$

Interprétation

- Conditionnement = apport de connaissances
- Si l'on connaît la valeur de la variable Z , alors connaître celle de Y n'apporte rien sur la connaissance de X

Motivations : réseaux bayésiens

- n variables aléatoires X_1, \dots, X_n
- $P(X_n, \dots, X_1) = P(X_n|X_{n-1}, \dots, X_1)P(X_{n-1}, \dots, X_1)$

- Par récurrence :

$$P(X_n, \dots, X_1) = P(X_1) \times \prod_{i=2}^n P(X_i|X_1, \dots, X_{i-1})$$

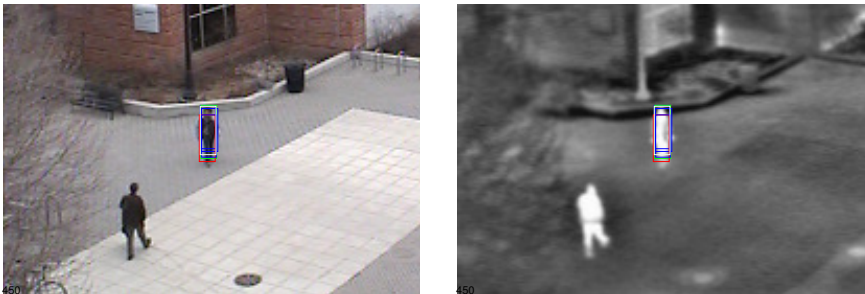
- $\forall i, \{X_1, \dots, X_{i-1}\} = L_i \cup K_i$, où $L_i \cap K_i = \emptyset$ et X_i indépendant de L_i conditionnellement à K_i

- Alors :

$$P(X_n, \dots, X_1) = P(X_1) \times \prod_{i=2}^n P(X_i|K_i)$$

- Tables de proba $P(X_i|K_i)$ plus petites que $P(X_i|X_1, \dots, X_{i-1})$

Motivations : tracking multimodal



hypothèse : observations indépendantes conditionnellement aux objets suivis

Problème : Peut-on tester la validité de cette hypothèse ?

Plan du cours n°5


- 1 Tests d'hypothèses
- 2 Loi du χ^2
- 3 Tests d'ajustement
- 4 Tests d'indépendance

Hypothèses

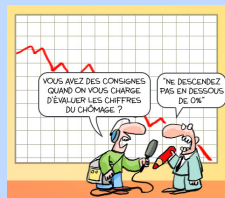
- Θ = ensemble des valeurs du paramètre θ
- Θ partitionné en Θ_0 et Θ_1
- **hypothèses** = assertions $H_0 = " \theta \in \Theta_0 "$ et $H_1 = " \theta \in \Theta_1 "$
- H_0 = **hypothèse nulle**, H_1 = **contre-hypothèse**
- hypothèse H_i est **simple** si Θ_i est un singleton ; sinon elle est **multiple**
- test **unilatéral** = valeurs dans Θ_1 toutes soit plus grandes, soit plus petites, que celles dans Θ_0 ; sinon test **bilatéral**

	hypothèse	test
$H_0 : \mu = 4$	simple	unilatéral
$H_1 : \mu = 6$	simple	
$H_0 : \mu = 4$	simple	test unilatéral
$H_1 : \mu > 4$	composée	
$H_0 : \mu = 4$	simple	test bilatéral
$H_1 : \mu \neq 4$	composée	
$H_0 : \mu = 4$	simple	formulation incorrecte : les hypothèses ne sont pas mutuellement exclusives
$H_1 : \mu > 3$	composée	

Exemples pratiques d'hypothèses

- association de consommateurs
 - échantillon de 100 bouteilles de Bordeaux
 - **Pb** : la quantité de vin est-elle bien égale à 75cl ? 
-
- paramètre θ étudié = $\mu = E(X)$
 - X = quantité de vin dans les bouteilles
 - rôle de l'association $\implies H_0 : \mu = 75\text{cl}$ et $H_1 : \mu < 75\text{cl}$

- le mois dernier, taux de chômage = 10%
 - échantillon : 400 individus de la pop. active
 - **Pb** : le taux de chômage a-t-il été modifié ?
-
- paramètre étudié = $p = \%$ de chômeurs
 - $H_0 : p = 10\%$ et $H_1 : p \neq 10\%$



Tests d'hypothèse

Définition du test

- test entre deux hypothèses H_0 et H_1 = **règle de décision** δ
- règle fondée sur les observations
- ensemble des décisions possibles = $\mathcal{D} = \{d_0, d_1\}$
- d_0 = "accepter H_0 "
- d_1 = "accepter H_1 " = "rejeter H_0 "

région critique

- échantillon $\implies n$ -uplet (x_1, \dots, x_n) de valeurs (dans \mathbb{R})
- δ = fonction $\mathbb{R}^n \mapsto \mathcal{D}$
- **région critique** : $W = \{n\text{-uplets } \mathbf{x} \in \mathbb{R}^n : \delta(\mathbf{x}) = d_1\}$
- région critique = **région de rejet**
- **région d'acceptation** = $A = \{\mathbf{x} \in \mathbb{R}^n : \delta(\mathbf{x}) = d_0\}$

Hypothèses	Règle de décision
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	« rejeter H_0 si $\bar{x} > c$ », où c est un nombre plus grand que μ_0
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	« rejeter H_0 si $\bar{x} < c$ », où c est un nombre plus petit que μ_0
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	« rejeter H_0 si $\bar{x} < c_1$ ou $c_2 < \bar{x}$ », où c_1 et c_2 sont des nombres respectivement plus petit et plus grand que μ_0 , et également éloignés de celui-ci

Problème : erreurs dans les décisions prises

Décision prise \ Réalité	H_0 est vraie	H_1 est vraie
	H_0 est rejetée	mauvaise décision : erreur de type I
H_0 n'est pas rejetée	bonne décision	mauvaise décision : erreur de type II

α = risque de première espèce

= probabilité de réaliser une erreur de type I

= probabilité de rejeter H_0 sachant que H_0 est vraie

= $P(\text{rejeter } H_0 | H_0 \text{ est vraie})$,

β = risque de deuxième espèce

= probabilité de réaliser une erreur de type II

= probabilité de rejeter H_1 sachant que H_1 est vraie

= $P(\text{rejeter } H_1 | H_1 \text{ est vraie})$.

Exemple de calcul de α (1/2)

Exemple

- échantillon de taille 25
- paramètre estimé : μ d'une variable $X \sim \mathcal{N}(\mu; 100)$
- hypothèses : $H_0 : \mu = 10$ $H_1 : \mu > 10$

Sous H_0 : $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 10}{10/5} = \frac{\bar{X} - 10}{2} \sim \mathcal{N}(0; 1)$

Sous H_0 : peu probable que \bar{X} éloignée de plus de 2 écarts-types de μ (4,56% de chance)

⇒ peu probable que $\bar{X} < 6$ ou $\bar{X} > 14$

⇒ région critique pourrait être « rejeter H_0 si $\bar{x} > 14$ »

Exemple de calcul de α (2/2)

- échantillon de taille 25
- paramètre estimé : μ d'une variable $X \sim \mathcal{N}(\mu; 100)$
- hypothèses : $H_0 : \mu = 10$ $H_1 : \mu > 10$
- région critique : « rejeter H_0 si $\bar{x} > 14$ »

$\alpha = P(\text{rejeter } H_0 | H_0 \text{ est vraie})$

= $P(\bar{X} > 14 | \mu = 10)$

= $P\left(\frac{\bar{X} - 10}{2} > \frac{14 - 10}{2} \mid \mu = 10\right)$

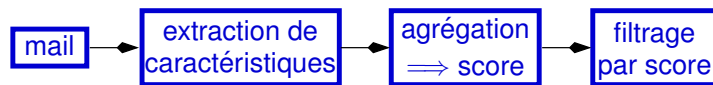
= $P\left(\frac{\bar{X} - 10}{2} > 2\right) = 0,0228$



en principe α est fixé et on cherche la région critique

Exemple de test d'hypothèses (1/2)

- filtre de mails sur un serveur mail :



- $X = \text{score} \geq 18000 \implies \text{spam}$; historiques des mails $\implies \sigma_X = 5000$
- le serveur reçoit un envoi en masse de $n = 400$ mails de $xx@yy.fr$
- **Problème** : $xx@yy.fr$ est-il un spammeur ?
- H_0 : $xx@yy.fr$ est « spammeur » v.s. H_1 : $xx@yy.fr \neq$ « spammeur »
- test : $H_0 : \mu = 18000$ v.s. $H_1 : \mu < 18000$ où $\mu = E(X)$
- règle : si $\bar{x} < c$ alors rejeter H_0
- 400 mails \implies théorème central limite \implies sous H_0 :

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 18000}{5000/\sqrt{400}} = \frac{\bar{X} - 18000}{250} \sim \mathcal{N}(0; 1)$$

Exemple de test d'hypothèses (2/2)

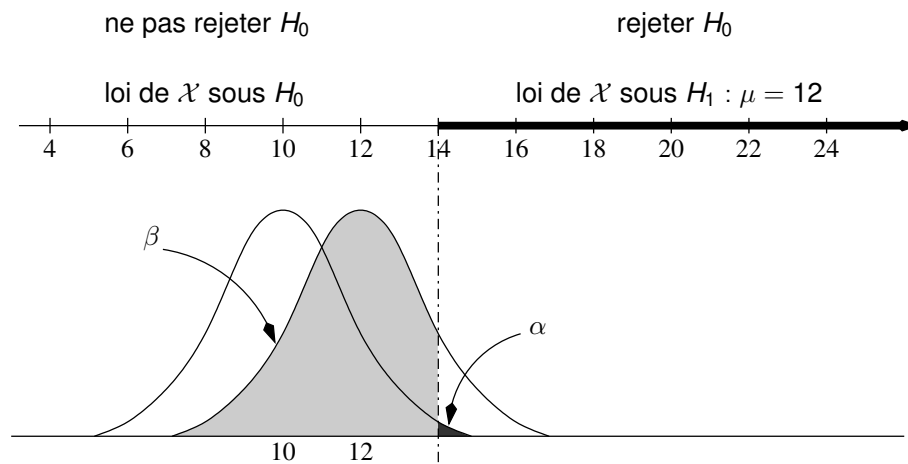
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 18000}{5000/\sqrt{400}} = \frac{\bar{X} - 18000}{250} \sim \mathcal{N}(0; 1)$$

- choix du risque de première espèce : $\alpha = 0,01$

$$\begin{aligned} \alpha = 0,01 &= P(\bar{X} < c | \mu = 18000) \\ &= P\left(\frac{\bar{X} - 18000}{250} < \frac{c - 18000}{250} \mid \mu = 18000\right) \\ &= P\left(Z < \frac{c - 18000}{250}\right) \\ &= P(Z < -2,326) \\ \implies \frac{c - 18000}{250} &= -2,326 \implies c = 17418,5 \end{aligned}$$

règle de décision : si $\bar{x} < 17418,5$, rejeter $H_0 \implies$ non spam

Interprétation de α et β



Puissance du test

$$\alpha = P(\text{rejeter } H_0 | H_0 \text{ est vraie})$$

$$\beta = P(\text{rejeter } H_1 | H_1 \text{ est vraie})$$

α et β varient en sens inverse l'un de l'autre

\implies test = compromis entre les deux risques

H_0 = hypothèse privilégiée, vérifiée jusqu'à présent et que l'on n'aimerait pas abandonner à tort

\implies on fixe un *seuil* α_0 :

- $\alpha \leq \alpha_0$
- test minimisant β sous cette contrainte
- $\min \beta = \max 1 - \beta$

$$1 - \beta = \text{puissance du test}$$

Exemple de calcul de β (1/2)

- échantillon de taille 25
- paramètre estimé : μ d'une variable $X \sim \mathcal{N}(\mu; 100)$
- hypothèses : $H_0 : \mu = 10$ $H_1 : \mu > 10$
- région critique : « rejeter H_0 si $\bar{x} > 14$ »

sous H_1 : plusieurs valeurs de μ sont possibles

\Rightarrow courbe de puissance du test en fonction de μ

Supposons que $\mu = 11$:

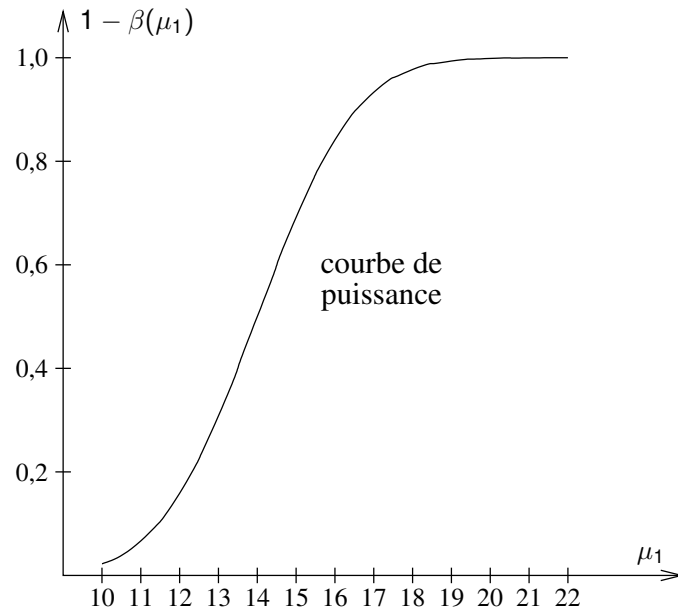
$$\mu = 11 \Rightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 11}{2} \sim \mathcal{N}(0; 1)$$

Exemple de calcul de β (2/2)

$$\begin{aligned} 1 - \beta(11) &= P(\text{rejeter } H_0 | H_1 : \mu = 11 \text{ est vraie}) \\ &= P(\bar{X} > 14 | \mu = 11) \\ &= P\left(\frac{\bar{X} - 11}{2} > \frac{14 - 11}{2} | \mu = 11\right) \\ &= P\left(\frac{\bar{X} - 11}{2} > 1,5\right) = 0,0668 \end{aligned}$$

μ_1	$z_1 = \frac{14 - \mu_1}{2}$	$1 - \beta(\mu_1) = P(Z > z_1)$	$\beta(\mu_1)$
10	2,0	0,0228	0,9772
11	1,5	0,0668	0,9332
12	1,0	0,1587	0,8413
13	0,5	0,3085	0,6915
14	0,0	0,5000	0,5000
15	-0,5	0,6915	0,3085
16	-1,0	0,8413	0,1587
17	-1,5	0,9332	0,0668

Courbe de puissance du test



Exemple : notes d'examen de MAPSI (1/3)

- les années précédentes, notes d'examen $\sim \mathcal{N}(14, 6^2)$
- cette année, correction d'un échantillon de 9 copies :

10	8	13	20	12	14	9	7	15
----	---	----	----	----	----	---	---	----

Les notes sont-elles en baisse cette année ?

- hypothèse H_0 = « la moyenne est égale à 14 »
 - hypothèse H_1 = « la moyenne a baissé, i.e., elle est ≤ 14 »
 - test d'hypothèse de niveau de confiance $1 - \alpha = 95\%$
- \Rightarrow déterminer seuil c tel que $\bar{x} < c \Rightarrow H_1$ plus probable que H_0

Exemple : notes d'examen de MAPSI (2/3)

10 | 8 | 13 | 20 | 12 | 14 | 9 | 7 | 15 $H_0 : \mu = 14, \sigma = 6$

• sous hypothèse H_0 , on sait que $\frac{\bar{X} - 14}{\sigma/\sqrt{n}} = \frac{\bar{X} - 14}{2} \sim \mathcal{N}(0; 1)$

• calcul du seuil c (région de rejet) :

$$P\left(\frac{\bar{X} - 14}{2} < \frac{c - 14}{2} \mid \frac{\bar{X} - 14}{2} \sim \mathcal{N}(0; 1)\right) = 0,05$$

• Table de la loi normale : $\frac{c-14}{2} \approx -1,645 \implies c = 10,71$

• **Règle de décision** : rejeter H_0 si $\bar{x} < 10,71$

• tableau $\implies \bar{x} = 12$

\implies on ne peut déduire que la moyenne a diminué

Exemple : notes d'examen de MAPSI (3/3)

Problème : le risque de 2ème espèce est-il élevé ?

Puissance du test pour une moyenne de 12

• H_1 : la moyenne est égale à 12

• Puissance du test = $1 - \beta(12)$

$$= P(\text{rejeter } H_0 | H_1)$$

$$= P\left(\bar{X} < 10,71 \mid \frac{\bar{X} - 12}{2} \sim \mathcal{N}(0; 1)\right)$$

$$= P\left(\frac{\bar{X} - 12}{2} < -0,645 \mid \frac{\bar{X} - 12}{2} \sim \mathcal{N}(0; 1)\right)$$

$$\approx 25,95\%$$

Loi du χ^2 (1/3)

• population \implies répartie en k classes

p_1	p_2	p_3		p_k
-------	-------	-------	--	-------

• hypothèse : répartition dans les classes connues

$\implies p_r =$ proba qu'un individu appartienne à la classe c_r

• échantillon de n individus

• $N_r =$ variable aléatoire « nombre d'individus tirés de classe c_r »

• Chaque individu $\implies p_r$ chances d'appartenir à la classe c_r

$\implies X_i^r =$ v.a. succès si l'individu i appartient à la classe c_r

$\implies X_i^r \sim \mathcal{B}(1, p_r)$

$\implies N_r \sim \mathcal{B}(n, p_r)$

$\implies N_r \sim$ loi normale quand n grand

Loi du χ^2 (2/3)

• population \implies répartie en k classes

p_1	p_2	p_3		p_k
-------	-------	-------	--	-------

• $p_r =$ proba qu'un individu appartienne à la classe c_r

• échantillon de n individus

• $N_r =$ v.a. « nb d'individus tirés de classe c_r » \sim loi normale

$$D_{(n)}^2 = \sum_{r=1}^k \frac{(N_r - n.p_r)^2}{n.p_r}$$

$\implies D_{(n)}^2 =$ somme des carrés de k v.a. \sim lois normales

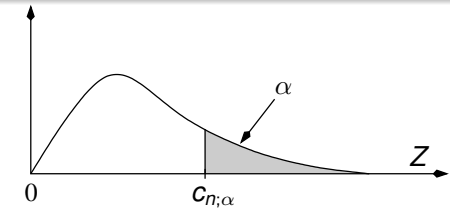
• $D_{(n)}^2 =$ écart entre théorie et observation

• $D_{(n)}^2$ tend en loi, lorsque $n \rightarrow \infty$, vers une loi du χ_{k-1}^2

Loi du χ^2

- loi du χ_r^2 = la loi de la somme des carrés de r variables indépendantes et de même loi $\mathcal{N}(0, 1)$
- espérance = r
- variance = $2r$

valeurs dans le tableau ci-dessous : les $c_{n,\alpha}$ tels que $P(Z > c_{n,\alpha}) = \alpha$



$n \setminus \alpha$	0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,005
1	0,00004	0,0002	0,001	0,0039	0,0158	2,71	3,84	5,02	6,63	7,88
2	0,0100	0,0201	0,0506	0,103	0,211	4,61	5,99	7,38	9,21	10,6
3	0,0717	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,3	12,8
4	0,207	0,297	0,484	0,711	1,06	7,78	9,49	11,1	13,3	14,9
5	0,412	0,554	0,831	1,15	1,61	9,24	11,1	12,8	15,1	16,7
6	0,676	0,872	1,24	1,64	2,20	10,6	12,6	14,4	16,8	18,5
7	0,989	1,24	1,69	2,17	2,83	12,0	14,1	16,0	18,5	20,3
8	1,34	1,65	2,18	2,73	3,49	13,4	15,5	17,5	20,1	22,0
9	1,73	2,09	2,70	3,33	4,17	14,7	16,9	19,0	21,7	23,6
10	2,16	2,56	3,25	3,94	4,87	16,0	18,3	20,5	23,2	25,2

Tests d'ajustement

Définition

- test d'ajustement = test \implies 2 issues possibles :
 - acceptation de l'hypothèse que l'échantillon observé est tiré selon une certaine loi
 - rejet de l'hypothèse ①
- contre-hypothèse : ne précise pas de quelle autre loi l'échantillon aurait pu être tiré

Tests d'ajustement II : le retour du χ^2

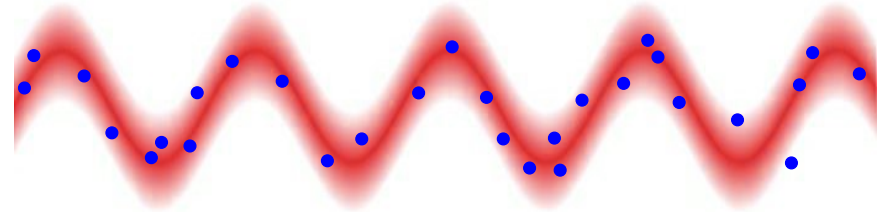
- population répartie en k classes
- échantillon de taille $n \implies$ répartition = (n_1, \dots, n_k)
- supposons l'échantillon tiré selon la loi discrète (p_1, \dots, p_k)
 $\implies (n_1, \dots, n_k) \approx (n.p_1, \dots, n.p_k)$

$$\text{Rappel : } D_{(n)}^2 = \sum_{r=1}^k \frac{(N_r - n.p_r)^2}{n.p_r} \sim \chi_{k-1}^2$$

- d^2 valeur prise par $D_{(n)}^2$
 \implies si échantillon tiré selon (p_1, \dots, p_k) alors d^2 petit
- table de la loi du $\chi^2 \implies d_\alpha^2$ tel que $P(\chi_{k-1}^2 > d_\alpha^2) = \alpha$
 \implies règle de décision : si $d^2 < d_\alpha^2$ alors OK

Mise en place d'un test d'ajustement

- 1 population répartie en k classes
- 2 échantillon de taille $n \Rightarrow$ répartition = (n_1, \dots, n_k)
- 3 on vérifie si l'échantillon tiré selon la loi (p_1, \dots, p_k) :
 - A choix du risque de première espèce α
 - B calcul de $d^2 = \sum_{r=1}^k \frac{(n_r - n.p_r)^2}{n.p_r}$
 - C lecture dans une table de d_α^2 tel que $P(\chi_{k-1}^2 > d_\alpha^2) = \alpha$
 - D si $d^2 < d_\alpha^2$ alors règle de décision :
 (p_1, \dots, p_k) est la loi selon laquelle est tiré l'échantillon
 sinon l'échantillon est tiré selon une autre loi



- observations = ● = $\{(x_i, y_i)\}$
- **Problème** : les ● proviennent-ils de points situés sur la courbe $y = \sin(x)$ mais observés avec un bruit gaussien ?

\Rightarrow problème : $T_i = Y_i - \sin(x_i) \sim \mathcal{N}(0, 1)$?

observations des t_i , réparties en 8 classes :

t_i	$] -\infty; -3[$	$[-3; -2[$	$[-2; -1[$	$[-1; 0[$	$[0; 1[$	$[1; 2[$	$[2; 3[$	$[3; +\infty[$
N_r	1	2	13	35	30	15	3	1

Exemple de test d'ajustement (2/3)

Rappel : $T_i \sim \mathcal{N}(0, 1)$

t_i	$] -\infty; -3[$	$[-3; -2[$	$[-2; -1[$	$[-1; 0[$	$[0; 1[$	$[1; 2[$	$[2; 3[$	$[3; +\infty[$
N_r	1	2	13	35	30	15	3	1
$n.p_r$	0.14	2.14	13.59	34.13	34.13	13.59	2.14	0.14

$$\Rightarrow d^2 = \sum_{r=1}^8 \frac{(n_r - n.p_r)^2}{n.p_r} \approx 11.61$$

pour $\alpha = 0.05$, $P(\chi_7^2 > d_\alpha^2) = \alpha \Rightarrow d_\alpha^2 = 14.1$

$\Rightarrow d^2 < d_\alpha^2 \Rightarrow$ règle de décision :

l'échantillon est bien tiré selon $\sin(x)$ + un bruit gaussien

Exemple de test d'ajustement (3/3)

Nouvel échantillon :

t_i	$] -\infty; -3[$	$[-3; -2[$	$[-2; -1[$	$[-1; 0[$	$[0; 1[$	$[1; 2[$	$[2; 3[$	$[3; +\infty[$
N_r	2	2	12	35	30	15	3	1
$n.p_r$	0.14	2.14	13.59	34.13	34.13	13.59	2.14	0.14

$$\Rightarrow d^2 = \sum_{r=1}^8 \frac{(n_r - n.p_r)^2}{n.p_r} \approx 31.20$$

pour $\alpha = 0.05$, $P(\chi_7^2 > d_\alpha^2) = \alpha \Rightarrow d_\alpha^2 = 14.1$

$\Rightarrow d^2 > d_\alpha^2 \Rightarrow$ règle de décision :

l'échantillon n'est pas tiré selon $\sin(x)$ + un bruit gaussien

Exemple de test d'ajustement (1/2)

- péage d'autoroute : 10 cabines
- nombre de clients / cabine sur une heure :



N° cabine	1	2	3	4	5	6	7	8	9	10
Nb clients	24	14	18	20	23	13	23	24	23	18

Clients distribués uniformément sur l'ensemble des cabines ?

⇒ test d'ajustement, niveau de confiance : $1 - \alpha = 95\%$

- H_0 = « la répartition des clients est uniforme »
- H_1 = « la répartition n'est pas uniforme »
- $H_0 \Rightarrow 20$ clients / cabine (uniforme)

Exemple de test d'ajustement (2/2)

- X_i : variable « effectif » recensé pour la i ème cabine
- Statistique d'ajustement : $D^2 = \sum_{i=1}^{10} \frac{(X_i - 20)^2}{20}$
- $D^2 \sim \chi_9^2$
- $\alpha = 0,05 = P(\text{rejeter } H_0 | H_0 \text{ est vraie})$
 $= P(D^2 > d_\alpha | D^2 \sim \chi_9^2)$
 $\Rightarrow d_\alpha = 16,9$

- calcul de la valeur de d observée sur l'échantillon :

$$d^2 = \frac{1}{20} [(14 - 20)^2 + (24 - 20)^2 + (18 - 20)^2 + (20 - 20)^2 + (23 - 20)^2 + (13 - 20)^2 + (23 - 20)^2 + (18 - 20)^2 + (24 - 20)^2 + (23 - 20)^2] = 7,6.$$

⇒ estimation : répartition uniforme

Tests d'indépendance (1/3)

- 2 caractères X et Y
- classes de X : A_1, A_2, \dots, A_I
- classes de Y : B_1, B_2, \dots, B_J
- échantillon de taille n
- tableau de contingence :

$X \setminus Y$	B_1	B_2	...	B_j	...	B_J
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1J}
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2J}
⋮	⋮	⋮		⋮		⋮
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iJ}
⋮	⋮	⋮		⋮		⋮
A_I	n_{I1}	n_{I2}	...	n_{Ij}	...	n_{IJ}

Tests d'indépendance (2/3)

$X \setminus Y$	B_1	B_2	...	B_j	...	B_J	total
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1J}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2J}	$n_{2.}$
⋮	⋮	⋮		⋮		⋮	⋮
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iJ}	$n_{i.}$
⋮	⋮	⋮		⋮		⋮	⋮
A_I	n_{I1}	n_{I2}	...	n_{Ij}	...	n_{IJ}	$n_{I.}$
total	$n_{.1}$	$n_{.2}$...	$n_{.j}$...	$n_{.J}$	n

$$\frac{n_{ij}}{n} = P(X \in A_i, Y \in B_j)$$

$$P(X \in A_i) = \frac{n_{i.}}{n} = \frac{\sum_{j=1}^J n_{ij}}{n} \quad \text{et} \quad P(Y \in B_j) = \frac{n_{.j}}{n} = \frac{\sum_{i=1}^I n_{ij}}{n}$$

X et Y indépendants $\Rightarrow P(X \in A_i, Y \in B_j) = P(X \in A_i) \times P(Y \in B_j)$

Tests d'indépendance (3/3)

$X \setminus Y$	B_1	B_2	\dots	B_j	\dots	B_J	total
A_1	n_{11}	n_{12}	\dots	n_{1j}	\dots	n_{1J}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	\dots	n_{2j}	\dots	n_{2J}	$n_{2\cdot}$
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots
A_i	n_{i1}	n_{i2}	\dots	n_{ij}	\dots	n_{iJ}	$n_{i\cdot}$
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots
A_l	n_{l1}	n_{l2}	\dots	n_{lj}	\dots	n_{lJ}	$n_{l\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 2}$	\dots	$n_{\cdot j}$	\dots	$n_{\cdot J}$	n

$$X \text{ et } Y \text{ indépendants} \implies \frac{n_{ij}}{n} = \frac{n_{i\cdot}}{n} \times \frac{n_{\cdot j}}{n} \implies n_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$$

$$\chi_{(l-1) \times (j-1)}^2 = \sum_{i=1}^l \sum_{j=1}^J \frac{(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n})^2}{\frac{n_{i\cdot} n_{\cdot j}}{n}}$$

Exemple de test d'indépendance (1/2)

- notes d'examen de MAPSI \implies 3 classes :

c_1	c_2	c_3
note < 8	note $\in [8, 12[$	note ≥ 12



- X : variable aléatoire « note 1ère session »
- Y : variable aléatoire « note 2ème session »

X et Y sont-elles des variables aléatoires indépendantes ?

- sélection d'un échantillon de 100 notes :

$X \setminus Y$	c_1	c_2	c_3
c_1	2	13	6
c_2	11	27	13
c_3	3	17	8

Exemple de test d'indépendance (2/2)

Test d'indépendance de niveau de confiance 90%

- calcul des marginales :

$X \setminus Y$	c_1	c_2	c_3	total
c_1	2	13	6	21
c_2	11	27	13	51
c_3	3	17	8	28
total	16	57	27	

- tableau obtenu si X et Y sont indépendants :

$X \setminus Y$	c_1	c_2	c_3
c_1	3.36	11.97	5.67
c_2	8.16	29.07	13.77
c_3	4.48	15.96	7.56

- calcul de la statistique d^2 : $d^2 = 2,42$
- $D^2 \sim \chi_4^2 \implies d_\alpha^2 = 7,78 \implies d^2 < d_\alpha^2 \implies$ indépendance