MAPSI — cours 5 :
Tests d'indépendance

Pierre-Henri Wuillemin & Nicolas Thome

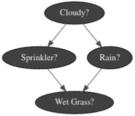
LIP6 / ISIR – Sorbonne Université, France

Cloudy?	
0	1
$\frac{1}{2}$	$\frac{1}{2}$

$P(Cloudy)$

	Sprinkler?	
Cloudy?	0	1
0	$\frac{1}{2}$	$\frac{1}{2}$
1	$\frac{9}{10}$	$\frac{1}{10}$

$P(Sprinkler|Cloudy)$



	Rain?	
Cloudy?	0	1
0	$\frac{4}{5}$	$\frac{1}{5}$
1	$\frac{1}{5}$	$\frac{4}{5}$

$P(WetGrass|Sprinkler, Rain)$

		Wet Grass?	
Rain?	Sprinkler?	0	1
0	0	$\frac{1}{1}$	$\frac{0}{1}$
	1	$\frac{1}{10}$	$\frac{9}{10}$
1	0	$\frac{1}{10}$	$\frac{9}{10}$
	1	$\frac{1}{100}$	$\frac{99}{100}$

$P(WetGrass|Sprinkler, Rain)$

Définition d'un réseau bayésien

1. un graphe sans circuit qui représente une décomposition de la loi jointe :
   $P(C, S, R, W) = P(C)P(S|C)P(R|C)P(W|R, S)$
2. À chaque noeud X du graphe est associé sa probabilité conditionnellement à ses parents.

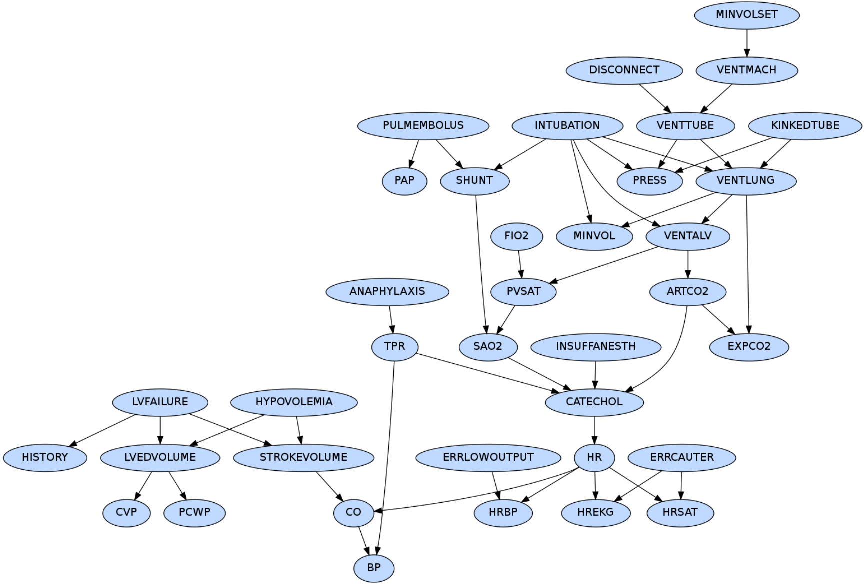
Motivations : réseaux bayésiens

- 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 :

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

- $\forall i$, $\{X_1, \dots, X_{i-1}\} = NP_i \cup P_i$, où
  - $NP_i \cap P_i = \emptyset$
  - X_i indépendant de NP_i conditionnellement à P_i
- Alors $P(X_n, \dots, X_1) = \prod_{i=1}^{n} P(X_i | NP_i, P_i) = \prod_{i=1}^{n} P(X_i | P_i)$
- Tables de proba $P(X_i | P_i)$ plus petites que $P(X_i | X_1, \dots, X_{i-1})$

De 10^{16} à 752 paramètres !



Plan du cours n°5

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

Tests d'hypothèses en statistique classique (1/2)

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*

Si $\Theta \in \mathbb{R}$,

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

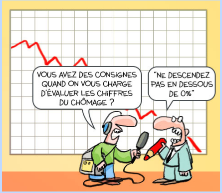
Exemples pratiques d'hypothèses

- 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 $\Longrightarrow H_0 : \mu = 75\text{cl}$ 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 $H_1 : p \neq 10\%$



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 $\Longrightarrow$ n-uplet $(x_1, \ldots, x_n)$ de valeurs (dans $\mathbb{R}$)
- Fonction de décision $\delta : \mathbb{R}^n \mapsto \mathcal{D}$
- *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* = $A = \{\mathbf{x} \in \mathbb{R}^n : \delta(\mathbf{x}) = d_0\}$

Régions critiques

Hypothèses	Règle de décision
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	« rejeter H_0 si $\overline{x} > c$ », où c est un nombre plus grand que μ_0
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	« rejeter H_0 si $\overline{x} < c$ », 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 $\overline{x} < c_1$ ou $c_2 < \overline{x}$ », 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

Erreurs dans les décisions

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

$$\begin{aligned}\alpha &= \text{risque de première espèce}\\ &= \text{probabilité de réaliser une erreur de type I}\\ &= \text{probabilité de rejeter } H_0 \text{ sachant que } H_0 \text{ est vraie}\\ &= P(\text{rejeter } H_0 | H_0 \text{ est vraie}),\end{aligned}$$

$$\begin{aligned}\beta &= \text{risque de deuxième espèce}\\ &= \text{probabilité de réaliser une erreur de type II}\\ &= \text{probabilité de rejeter } H_1 \text{ sachant que } H_1 \text{ est vraie}\\ &= P(\text{rejeter } H_1 | H_1 \text{ est vraie}).\end{aligned}$$

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 : $H_0 : \mu = 10 \quad H_1 : \mu > 10$

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

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

$\Longrightarrow$ peu probable que $\overline{X} < 6$ ou $\overline{X} > 14$

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

Exemple de calcul de α (2/2)

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

$$
\begin{aligned}
\alpha &= P(\text{rejeter } H_0 | H_0 \text{ est vraie}) \\
&= P(\overline{X} > 14 | \mu = 10) \\
&= P\left(\frac{\overline{X} - 10}{2} > \frac{14 - 10}{2} \middle| \mu = 10\right) \\
&= P\left(\frac{\overline{X} - 10}{2} > 2\right) = 0,0228
\end{aligned}
$$

⚠ 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 :

mail → extraction de caractéristiques → agrégation $\Longrightarrow$ score → filtrage par score

- X = score $\geq 18000 \Longrightarrow$ spam ; historiques des mails $\Longrightarrow \sigma_X = 5000$
- 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 = « 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 $\overline{x} < c$ alors rejeter H_0
- 400 mails $\Longrightarrow$ théorème central limite $\Longrightarrow$ sous H_0 :

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

Exemple de test d'hypothèses (2/2)

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

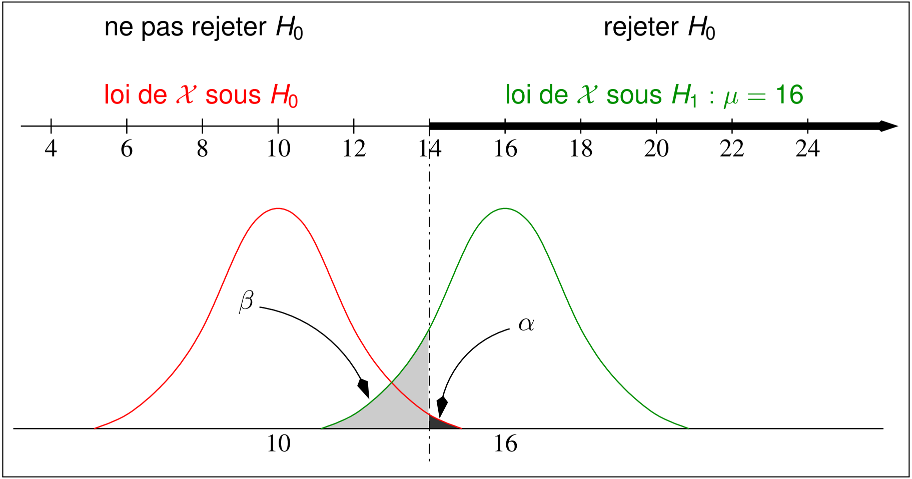
- choix du risque de première espèce : $\alpha = 0,01$
- $$\begin{aligned} \alpha = 0,01 &= P(\overline{X} < c|\mu = 18000) \\ &= P\left(\frac{\overline{X} - 18000}{250} < \frac{c - 18000}{250}|\mu = 18000\right) \\ &= P\left(Z < \frac{c - 18000}{250}\right) \\ &= P(Z < -2,326) \end{aligned}$$

$$\Longrightarrow \frac{c - 18000}{250} = -2,326 \Longrightarrow c = 17418,5$$

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

Interprétation de α et β

- hypothèses : H_0 : $\mu = 10$ H_1 : $\mu > 10$
- région critique : « rejeter H_0 si $\overline{x} > 14$ »



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

$\Longrightarrow$ test = compromis entre les deux risques

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

$\Longrightarrow$ on fixe un *seuil* α_0 :

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

$1 - \beta$ = 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 : $H_0 : \mu = 10 \qquad H_1 : \mu > 10$
- région critique : « rejeter H_0 si $\overline{x} > 14$ »

sous H_1 : plusieurs valeurs de μ sont possibles

$\Longrightarrow$ courbe de puissance du test en fonction de μ

Supposons que $\mu = 11$:

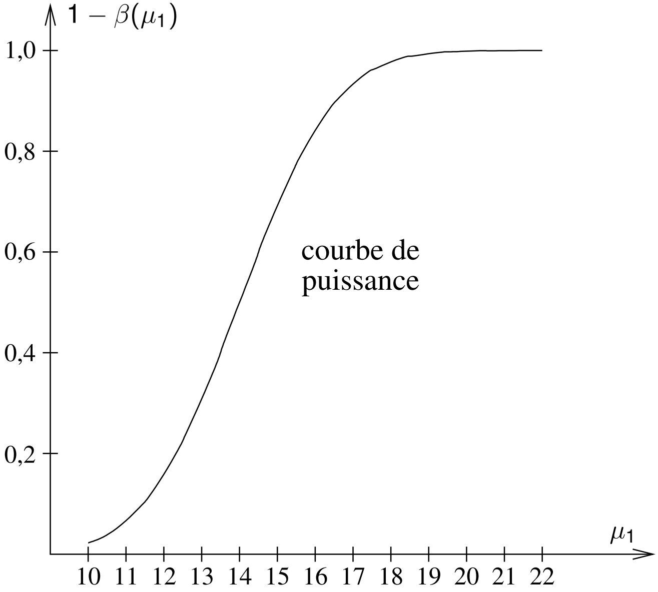
$$\mu = 11 \Longrightarrow \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - 11}{2} \sim \mathcal{N}(0; 1)$$

Exemple de calcul de β (2/2)

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

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 $1 - \alpha = 95\%$

$\Longrightarrow$ déterminer seuil c tel que $\overline{x} < c \Longrightarrow H_1$ 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 $\dfrac{\overline{X} - 14}{\sigma/\sqrt{n}} = \dfrac{\overline{X} - 14}{2} \sim \mathcal{N}(0; 1)$
- calcul du seuil c (région de rejet) :

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

- Table de la loi normale : $\frac{c-14}{2} \approx -1,645 \Longrightarrow c = 10,71$
- Règle de décision : rejeter H_0 si $\overline{x} < 10,71$
- tableau $\Longrightarrow \overline{x} = 12$

  $\Longrightarrow$ 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 $= 1 - \beta(12)$

$$
\begin{aligned}
&= P(\text{rejeter } H_0 | H_1) \\
&= P\left(\overline{X} < 10,71 \,\middle|\, \frac{\overline{X}-12}{2} \sim \mathcal{N}(0;1)\right) \\
&= P\left(\frac{\overline{X}-12}{2} < -0,645 \,\middle|\, \frac{\overline{X}-12}{2} \sim \mathcal{N}(0;1)\right) \\
&\approx 25,95\%.
\end{aligned}
$$

Tests d'ajustement

Définition

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

- population $\Longrightarrow$ répartie en k classes

p_1	p_2	p_3		p_k

- hypothèse : répartition dans les classes connues

  $\Longrightarrow p_r =$ 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 $\Longrightarrow p_r$ chances d'appartenir à la classe c_r

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

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

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

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

Loi du χ^2

- population $\Longrightarrow$ 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

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

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

- $D^2_{(n)}$ = écart entre théorie et observation
- $D^2_{(n)}$ tend en loi, lorsque $n \to \infty$, vers une loi du χ^2_{k-1}

Loi du χ^2 (3/3)

Loi du χ^2 de degré de liberté r

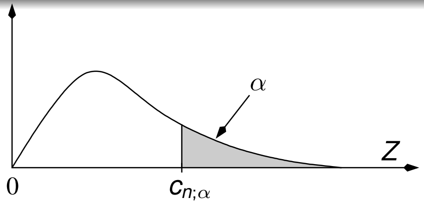
- loi du χ^2_r = la loi de la somme des carrés de r variables Z_i iid, $\forall i, Z_i \sim \mathcal{N}(0,1)$

$$\chi^2_r = \sum_{i=1}^{r} Z_i^2$$

- espérance = r
- variance = $2r$
- si $r > 100$ alors $\chi^2_r \approx \mathcal{N}(r; 2r)$

Table de la loi du χ^2

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

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

  $\Longrightarrow (n_1, \ldots, n_k) \approx (n.p_1, \ldots, n.p_k)$

Rappel : $D^2_{(n)} = \sum_{r=1}^{k} \frac{(N_r - n.p_r)^2}{n.p_r} \sim \chi^2_{k-1}$

- d^2 valeur prise par $D^2_{(n)}$

  $\Longrightarrow$ si échantillon tiré selon $(p_1, \ldots, p_k)$ alors d^2 petit
- table de la loi du $\chi^2 \Longrightarrow d^2_\alpha$ tel que $P(\chi^2_{k-1} > d^2_\alpha) = \alpha$

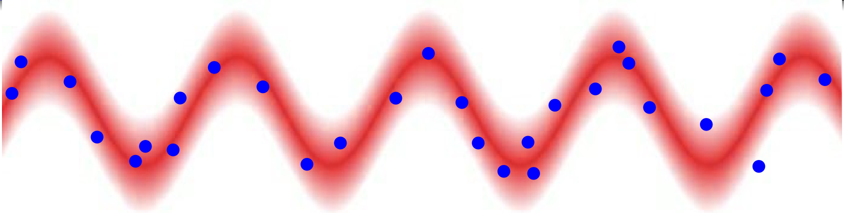
  $\Longrightarrow$ règle de décision : si $d^2 < d^2_\alpha$ alors OK

Mise en place d'un test d'ajustement

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

     $(p_1, \ldots, p_k)$ 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 ?

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

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

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

pour $\alpha = 0.05$, $P(\chi^2_7 > d^2_\alpha) = \alpha \Longrightarrow d^2_\alpha = 14.1$

$\Longrightarrow d^2 < d^2_\alpha \Longrightarrow$ 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

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

pour $\alpha = 0.05$, $P(\chi^2_7 > d^2_\alpha) = \alpha \Longrightarrow d^2_\alpha = 14.1$

$\Longrightarrow d^2 > d^2_\alpha \Longrightarrow$ 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 ?

$\Longrightarrow$ 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 »

- $H_0 \Longrightarrow$ 20 clients / cabine (uniforme)

Exemple de test d'ajustement (2/2)

- X_i : variable « effectif » recensé pour la $i$ème cabine
- Statistique d'ajustement : $D^2 = \sum_{i=1}^{10} \frac{(X_i - 20)^2}{20}$
- $D^2 \sim \chi^2_9$
- $\alpha = 0,05 = P(\text{rejeter } H_0 | H_0 \text{ est vraie})$

  $= P\left(D^2 > d_\alpha \mid D^2 \sim \chi^2_9\right)$

  $\Longrightarrow d_\alpha = 16,9$
- calcul de la valeur de d observée sur l'échantillon :

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

$\Longrightarrow$ estimation : répartition uniforme

Tests d'indépendance (1/3)

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

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

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

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

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

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

Tests d'indépendance (3/3)

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

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

$$\chi^2_{(I-1)\times(J-1)} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

Exemple de test d'indépendance (1/2)

- notes d'examen de MAPSI $\Longrightarrow$ 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 :

$X \backslash Y$	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%

1. calcul des marginales :

$X \backslash Y$	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	

2. tableau obtenu si X et Y sont indépendants :

$X \backslash Y$	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

3. calcul de la statistique d^2 : $d^2 = 2,42$
4. $D^2 \sim \chi^2_4 \Longrightarrow d^2_\alpha = 7,78 \Longrightarrow d^2 < d^2_\alpha \Longrightarrow$ indépendance