

RFIDEC — cours 3 : Intervalles de confiance, tests d'hypothèses, loi du χ^2

Christophe Gonzales

LIP6 – Université Paris 6, France

- 1 Intervalles de confiance
- 2 Tests d'hypothèses
- 3 La loi du χ^2

Intervalles de confiance

Estimateur T d'un paramètre $\theta \implies$ valeur estimée $\hat{\theta}$

Problème : peut-on avoir confiance dans l'estimation ponctuelle ?

Intervalle de confiance

Un **intervalle de confiance** de niveau $1 - \alpha$ = intervalle $]a(T), b(T)[$ tel que :

$$\forall \theta \in \Theta, P_{\theta}(]a(T), b(T)[\ni \theta) = 1 - \alpha$$



$1 - \alpha$ = proba que l'intervalle contienne θ

Intervalles de confiance : exemple (1/2)

$$X \sim \mathcal{N}(\mu; \sigma^2)$$

échantillon de taille $n \implies \bar{X}$ = moyenne

théorème central-limite $\implies \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0; 1)$

$$\begin{cases} n \text{ très grand} \implies \text{la valeur observée } \bar{x} \text{ de } \bar{X} \approx \mu \\ n \text{ moins grand} \implies \bar{x} \not\approx \mu \end{cases}$$

\implies estimation par intervalle de confiance de niveau 95%

$$\text{loi normale} \implies P\left(-1,96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1,96\right) = 95\%$$

Intervalle de confiance : exemple (2/2)

$$P\left(-1,96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1,96\right) = 95\%$$

$$P\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 95\%$$

⇒ intervalle de confiance = $\left] \bar{x} - 1,96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1,96 \frac{\sigma}{\sqrt{n}} \right[$

⚠ seulement maintenant, on tire un échantillon de taille n

⇒ observation de \bar{x}

⇒ on peut calculer $\left] \bar{x} - 1,96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1,96 \frac{\sigma}{\sqrt{n}} \right[$

Intervalle de confiance : autre exemple (1/2)

Énoncé de l'exemple

- plusieurs centaines de candidats à un examen
- variance sur les notes obtenues ≈ 16
- correcteur \Rightarrow noté 100 copies, moyenne = 8,75
- Problème : moyenne sur toutes les copies de l'examen ?
- hypothèse : les notes suivent une loi normale $\mathcal{N}(\mu; 16)$

\bar{X} = variable aléatoire « moyenne des notes d'un correcteur »

théorème central-limite $\Rightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu}{4/10} \sim \mathcal{N}(0; 1)$

chercher dans la table de la loi normale $z_{\alpha/2}$ tel que :

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Intervalle de confiance : autre exemple (2/2)

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$1 - \alpha$	intervalle de confiance
50%	$[8,75 - 0,674 \times 0,4; 8,75 + 0,674 \times 0,4] = [8,48; 9,02]$
75%	$[8,75 - 1,15 \times 0,4; 8,75 + 1,15 \times 0,4] = [8,29; 9,21]$
80%	$[8,75 - 1,28 \times 0,4; 8,75 + 1,28 \times 0,4] = [8,24; 9,26]$
90%	$[8,75 - 1,645 \times 0,4; 8,75 + 1,645 \times 0,4] = [8,09; 9,41]$
95%	$[8,75 - 1,96 \times 0,4; 8,75 + 1,96 \times 0,4] = [7,96; 9,53]$
99%	$[8,75 - 2,575 \times 0,4; 8,75 + 2,575 \times 0,4] = [7,72; 9,78]$

Exemple : analyse des déchets (cf. cours 2)

- Grenelle de l'environnement
 - ⇒ réduction des déchets
 - ⇒ analyse des déchets
 - ⇒ échantillon de taille 100



- \bar{x} : moyenne de l'échantillon = 390 kg/an/habitant
- écart-type $\sigma = 20$ supposé connu
- $\bar{X} \sim \mathcal{N}(\mu, 4)$

$$\Rightarrow P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{2} \leq z_{\alpha/2}\right) = 1 - \alpha$$

⇒ estimation par intervalle de confiance de niveau $1 - \alpha$:

$$\left] \bar{x} - 2z_{\alpha/2}, \bar{x} + 2z_{\alpha/2} \right[$$

Exemple : analyse des déchets (suite)

Estimation par intervalle de confiance de niveau $1 - \alpha$:

$$\left] \bar{x} - 2z_{\alpha/2}, \bar{x} + 2z_{\alpha/2} \right[$$

$1 - \alpha$	intervalle de confiance
50%	$[390 - 0,674 \times 2; 390 + 0,674 \times 2] = [388,65; 391,35]$
75%	$[390 - 1,15 \times 2; 390 + 1,15 \times 2] = [387,70; 392,30]$
80%	$[390 - 1,28 \times 2; 390 + 1,28 \times 2] = [387,44; 392,56]$
90%	$[390 - 1,645 \times 2; 390 + 1,645 \times 2] = [386,71; 393,29]$
95%	$[390 - 1,96 \times 2; 390 + 1,96 \times 2] = [386,08; 393,92]$
99%	$[390 - 2,575 \times 2; 390 + 2,575 \times 2] = [384,85; 395,15]$

Exemple du réchauffement climatique (cf. cours 2)

- opinion des gens sur le réchauffement climatique
- 1000 personnes de 15 ans et + interrogées
- 790 pensent qu'il y a un changement climatique
- 210 ne le pensent pas
- \bar{P} : proportion de succès moyenne de l'échantillon
- p : proportion de personnes pensant qu'il y a dérèglement climatique dans la population française



$$\frac{\bar{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0; 1)$$

\Rightarrow estimation par intervalle de confiance de niveau $1 - \alpha$:

$$\left] \bar{p} - \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2}; \bar{p} + \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2} \right[\approx \left] \bar{p} - \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} z_{\alpha/2}; \bar{p} + \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} z_{\alpha/2} \right[$$

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**

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

	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})
- $\delta =$ 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** = $A = \{\mathbf{x} \in \mathbb{R}^n : \delta(\mathbf{x}) = d_0\}$

Régions critiques

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

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
H_0 n'est pas rejetée	bonne décision	mauvaise décision : erreur de type II

$\alpha =$ risque de première espèce

= probabilité de réaliser une erreur de type I

= probabilité de rejeter H_0 sachant que H_0 est vraie

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

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

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

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

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

\Rightarrow 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 $\bar{x} > 14$ »

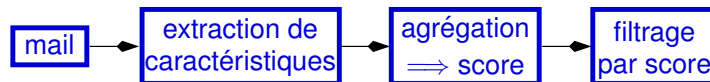
$$\begin{aligned}\alpha &= P(\text{rejeter } H_0 | H_0 \text{ est vraie}) \\ &= P(\bar{X} > 14 | \mu = 10) \\ &= P\left(\frac{\bar{X} - 10}{2} > \frac{14 - 10}{2} \mid \mu = 10\right) \\ &= P\left(\frac{\bar{X} - 10}{2} > 2\right) = 0,0228\end{aligned}$$



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 :



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

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

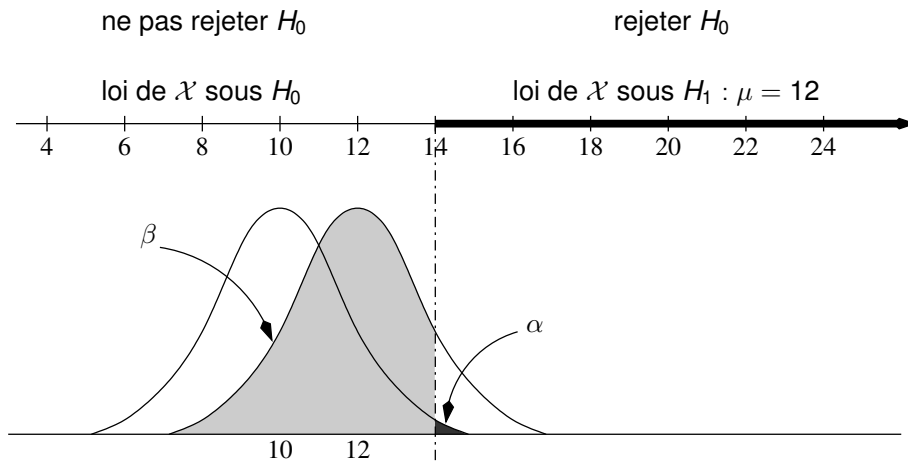
Exemple de test d'hypothèses (2/2)

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

- choix du risque de première espèce : $\alpha = 0,01$
- $\alpha = 0,01 = P(\bar{X} < c | \mu = 18000)$
$$= P\left(\frac{\bar{X} - 18000}{250} < \frac{c - 18000}{250} \mid \mu = 18000\right)$$
$$= P\left(Z < \frac{c - 18000}{250}\right)$$
$$= P(Z < -2,326)$$
$$\Rightarrow \frac{c - 18000}{250} = -2,326 \Rightarrow c = 17418,5$$

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

⇒ test = compromis entre les deux risques

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

⇒ 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

⇒ courbe de puissance du test en fonction de μ

Supposons que $\mu = 11$:

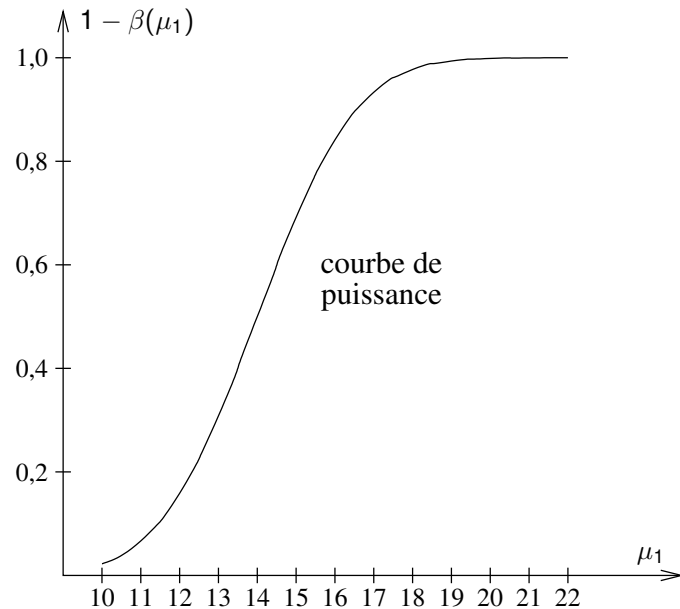
$$\mu = 11 \Rightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 11}{2} \sim \mathcal{N}(0; 1)$$

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 RFIDEC (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 $H_0 = \ll \text{la moyenne est égale à } 14 \gg$
hypothèse $H_1 = \ll \text{la moyenne a baissé, i.e., elle est } \leq 14 \gg$
test d'hypothèse de niveau de confiance $1 - \alpha = 95\%$
 \implies déterminer seuil c tel que $\bar{x} < c \implies H_1$ plus probable que H_0

Exemple : notes d'examen de RFIDEC (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 RFIDEC (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 = $1 - \beta(12)$
= $P(\text{rejeter } H_0 | H_1)$
= $P(\bar{X} < 10,71 \mid \frac{\bar{X} - 12}{2} \sim \mathcal{N}(0; 1))$
= $P(\frac{\bar{X} - 12}{2} < -0,645 \mid \frac{\bar{X} - 12}{2} \sim \mathcal{N}(0; 1))$
 $\approx 25,95\%$.

Lemme de Neyman-Pearson (1/2)

$$\text{Cas : } \Theta_0 = \{\theta_0\} \quad \Theta_1 = \{\theta_1\}$$

- Échantillon (x_1, \dots, x_n) de taille n
- Échantillon \implies les x_i = réalisations de variables aléatoires X_i
- Échantillon i.i.d. \implies les X_i sont mutuellement indépendants
 $\implies P(X_1 = x_1, \dots, X_n = x_n | \theta = \theta_k) = \prod_{i=1}^n P(X_i = x_i | \theta = \theta_k)$

Vraisemblance d'un échantillon

- $x = (x_1, \dots, x_n)$: échantillon de taille n
- $L(x, \theta_k)$ = Vraisemblance de l'échantillon
- $L(x, \theta_k)$ = proba d'obtenir **cet** échantillon sachant que $\theta = \theta_k$

$$L(x, \theta_k) = P(x_1, \dots, x_n | \theta = \theta_k) = \prod_{i=1}^n P(x_i | \theta = \theta_k)$$

Lemme de Neyman-Pearson (2/2)

$$\text{Cas : } \Theta_0 = \{\theta_0\} \quad \Theta_1 = \{\theta_1\}$$

Lemme de Neyman-Pearson

- il existe toujours un test (aléatoire) le plus puissant de seuil donné α_0
- c'est un test du rapport de vraisemblance :

$$\frac{L(x, \theta_0)}{L(x, \theta_1)} > k \implies x \in A \text{ (accepter } H_0)$$

$$\frac{L(x, \theta_0)}{L(x, \theta_1)} < k \implies x \in W \text{ (rejeter } H_0)$$

$$\frac{L(x, \theta_0)}{L(x, \theta_1)} = k \implies \delta(x) = \rho \text{ (accepter } H_0 \text{ avec proba } 1 - \rho$$

H_1 avec proba ρ)

- k et ρ déterminés de façon unique par $\alpha = \alpha_0$

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

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

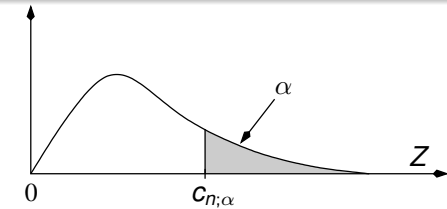
$\implies D_{(n)}^2$ = somme des carrés de k v.a. \sim lois normales

- $D_{(n)}^2$ = 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