EXEMPLE INTRODUCTIF

Saporta

- Relevés réguliers sur plusieurs années $\Rightarrow$ niveau de pluie dans la Beauce suit une loi $N(600,40^2)$

- Mise au point d'un procédé pour augmenter de 50 mm le niveau moyen de pluie testé de 1951 à 1959

Année	1951	1952	1953	1954	1955	1956	1957	1958	1959
mm	**510**	**614**	**780**	**512**	**501**	**534**	**603**	**788**	**650**

Procédé sans effet

Procédé concluant

$$\begin{cases} H_0 \quad : \quad \mu = 600\,\text{mm} \\ H_1 \quad : \quad \mu = 650\,\text{mm} \end{cases}$$

où µ espérance de la variable aléatoire X égale au niveau annuel de pluie

Test de H_0 : Hypothèse nulle contre H_1 : Hypothèse alternative

EXEMPLE INTRODUCTIF (suite)

Comment choisir entre H_0 et H_1?

Pour tester m, on utilise son estimateur $\overline{X}$

$\overline{X}$ s'appelle la variable de décision

Si H_0 est vraie, alors comme n=9, $\overline{X}$ suit une loi $N(600, \frac{40^2}{9})$

En principe les grandes valeurs de $\overline{X}$ sont improbables d'où la règle de décision :

- Si $\overline{X}$ trop grand, *i.e* $P[\overline{X} \geq C] = \alpha$ alors H_1 est accepté avec α chance de se tromper
- Sinon on conserve H_0 faute de preuves suffisantes

EXEMPLE INTRODUCTIF (suite)

Comment calculer le seuil C ?

Supposons α=0.05 alors $P[\overline{X} \geq C] = \alpha$

$$\Leftrightarrow P[\frac{3(\overline{X}-600)}{40} \geq \frac{3(C-600)}{40}] = 0.05$$

$\Leftrightarrow P[Z \geq C'] = 0.05$ où Z suit une $N(0,1)$

$\overset{\text{Table}}{\Rightarrow}$ C'=1,64 $\Rightarrow$ C=600+40×1,64/3=622

La règle de décision devient donc

- Si $\overline{X}$>622 mm, repousser H_0 et accepter H_1
- Si $\overline{X}$<622 mm, conserver H_0

L'ensemble des évènements { $\overline{X}$>622} est appelé la **région critique**

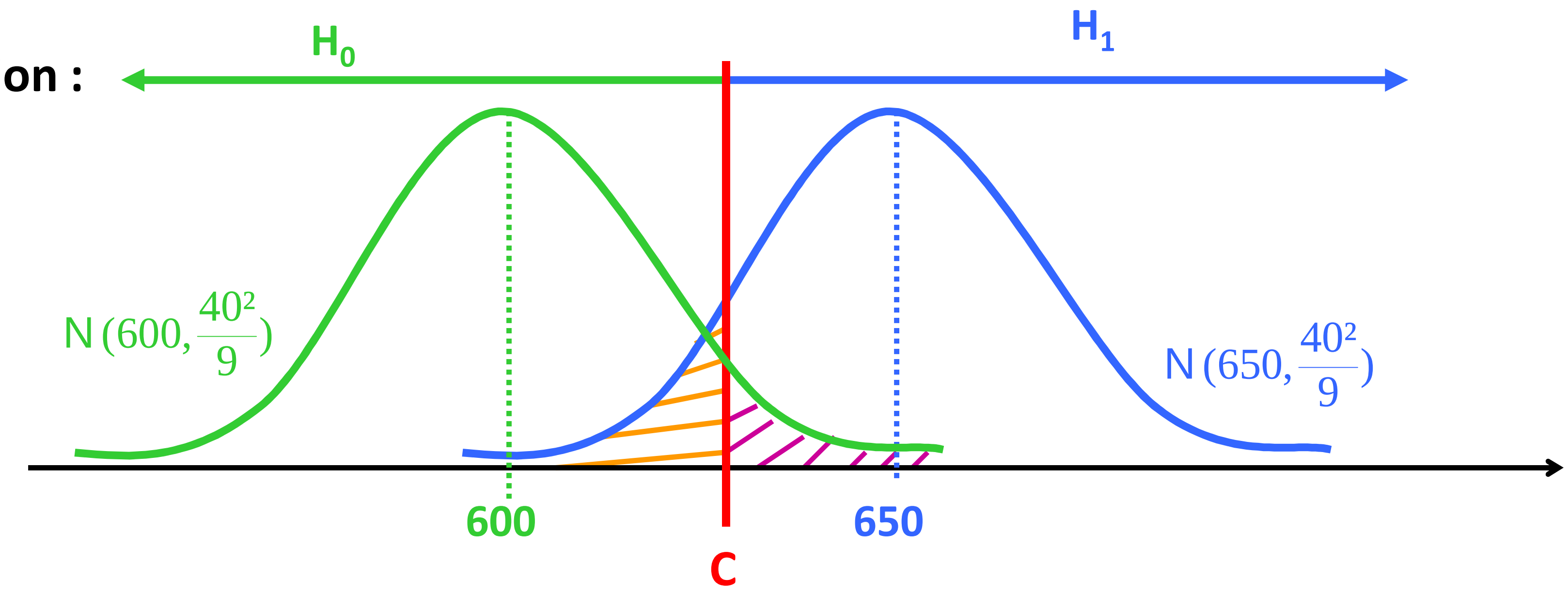
EXEMPLE INTRODUCTIF (suite)

Quelles erreurs commet-on ?

$\overline{x} = 610.2\,mm$

- Croire le procédé efficace alors qu'il n'est pour rien dans les résultats obtenu, *i.e* accepter H_1 alors que H_0 est vraie (probabilité α)
- Ne pas juger le procédé efficace alors que la méthode est bonne, *i.e* repousser H_1 alors que H_1 est vraie (probabilité β=?)

Décision :



EXEMPLE INTRODUCTIF (suite)

Calcul de β

Supposons H_1 est vraie, alors $\overline{X}$ suit une loi $N(650, \frac{40^2}{9})$

On commet une erreur à chaque fois que $\overline{X}$ prend une valeur inférieure à 622, d'où

$$\beta = P[\overline{X} < 622]$$

$$\Leftrightarrow \quad \beta = P[\frac{3(\overline{X} - 650)}{40} < \frac{3(622 - 650)}{40}]$$

$$\Leftrightarrow \quad \beta = P[Z < -2,1] = P[Z > 2,1] = 0,0179$$

L'erreur de juger le procédé inefficace alors qu'il est bénéfique est donc de 1,79%

EXEMPLE INTRODUCTIF (fin)

- α, la probabilité de choisir H_1 alors que H_0 est vraie, est appelée **le risque de première espèce**

- β, la probabilité de conserver H_0 alors que H_1 est vraie, est appelée **le risque de deuxième espèce**

Décision \ Vérité	H_0	H_1
H_0	$1-\alpha$	β
H_1	α	$1-\beta$

Dans notre exemple le risque de première espèce consiste à acheter un procédé qui ne vaut rien, le risque de deuxième espèce à laisser passer une occasion d'augmenter le niveau de pluie

METHODOLOGIE

Construction du test

- Choix de H_0 et H_1
- Déterminer la variable de décision
- Allure de la région critique en fonction de H_1
- Calcul de la région critique en fonction de α
- Calcul de la puissance $1-\beta$

Objectif :

- minimiser le risque de 1ère espèce α
- minimiser le risque de 2ème espèce β

problème insoluble, donc

on fixe α et on maximise la puissance $1-\beta$

Test UPP

Application du test

- Calcul de la valeur expérimentale de la variable de décision
- Conclusion : Rejet ou acceptation de H_0

CHOIX DE H_0 ET H_1

- H_0 est l'hypothèse privilégiée, celle que l'on garde quand les résultats ne sont pas clairs. Quand on accepte H_0 cela ne veut pas dire qu'elle est vraie mais cela veut dire qu'on n'a pas suffisamment de preuve pour la réfuter.
- H_1 est l'ensemble des hypothèses alternatives à H_0
- H_D est l'hypothèse à détecter (choisie parmi H_1)

Deux types d'hypothèses :

➢ Hypothèse simple du type H : $\theta=\theta_0$ où θ_0 est une valeur fixée du paramètre testé

➢ Hypothèse composite du type H : $\theta \in A$ où $A \subset \mathbb{R}$

En général : $\theta<\theta_0$ ou $\theta>\theta_0$ ou $\theta\neq\theta_0$

EXEMPLE

Un parti politique souhaite vérifier si la parité est respectée. Sur un échantillon de 100 adhérents, on compte 45 femmes et 55 hommes. Peut-on affirmer que la parité est respectée avec un risque de 5% de se tromper?

La p-valeur

Les tests statistiques usuels sont implémentés dans les divers logiciels d'analyse de données (Excel, SAS, R-cran project,…). La plupart de ces logiciels retournent la p-valeur du test statistique et

Valeur de la variable de décision sur l'échantillon → p-valeur = valeur minimum du risque α à partir de laquelle l'hypothèse H_0 est rejetée

C'est à partir de cette p-valeur que l'utilisateur prend une décision.

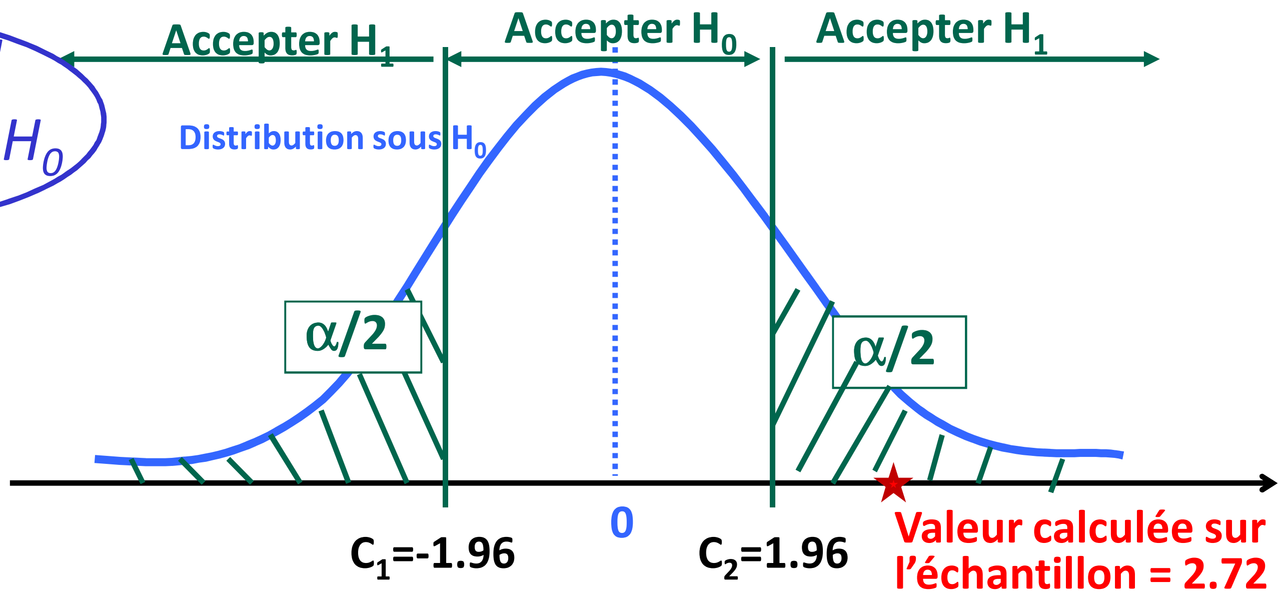
- **Si la p-valeur < risque autorisé (α = 1%, 5%,…) alors on rejette H_0**
- **Sinon on garde H_0**

Remarque : Le logiciel calcule toujours une p-valeur même si les conditions d'application du test ne sont pas vérifiées.

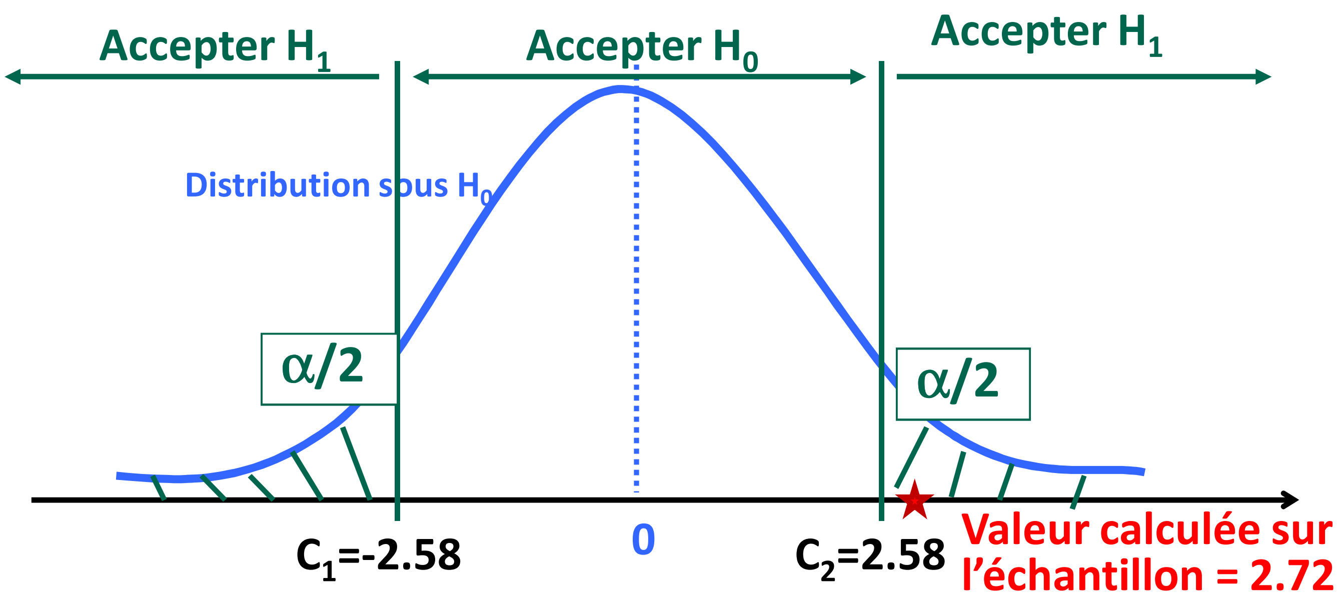
A vous de faire attention !!!

Test bilatéral
Loi N(0,1) sous H_0

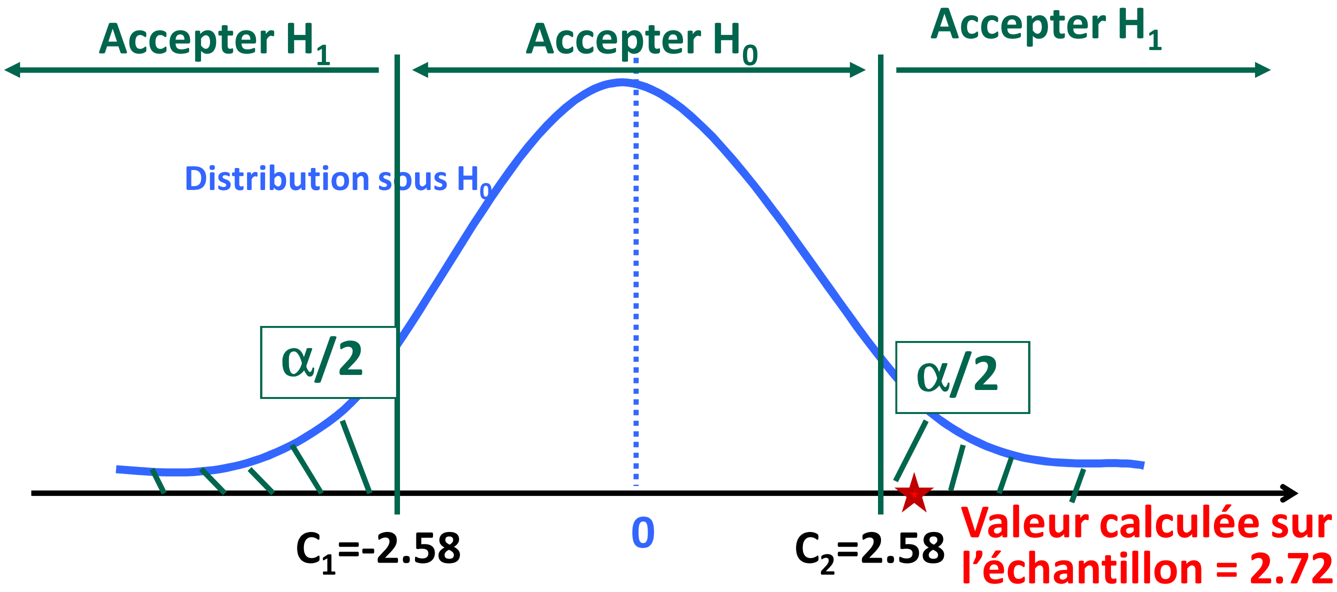
Si α=5%
H_0 est rejetée



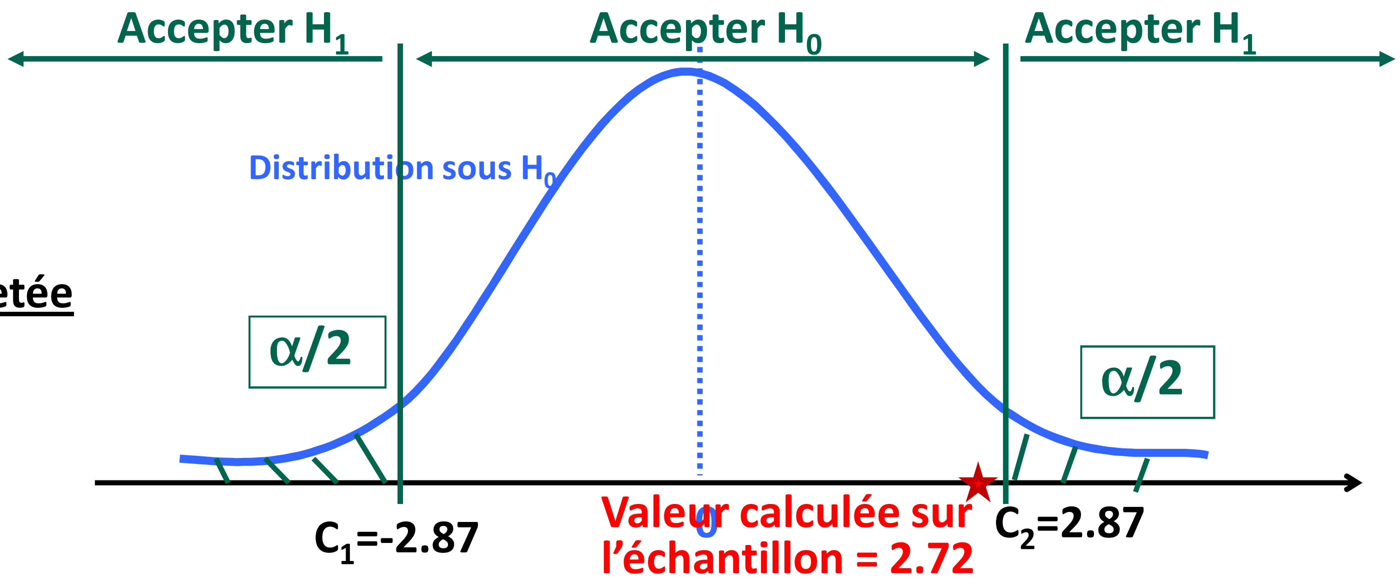
Si α=1%
H_0 est rejetée



Si α=1%
H_0 est rejetée



Si α=0.5%
<u>H0 n'est pas rejetée</u>

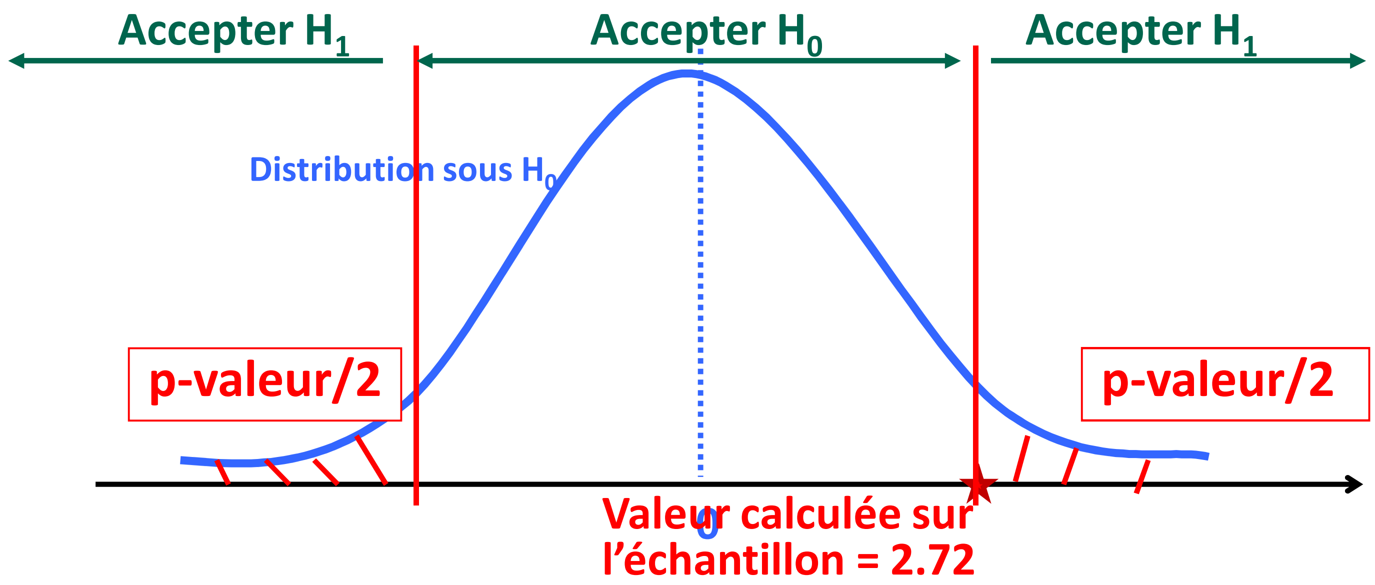


La valeur minimale de α pour laquelle H_0 est rejetée se situe donc entre 0.5% et 1%.

0.5%<p-valeur<1%

Elle est en fait de 0.65%

p-valeur=0.65%



Tests de comparaison d'échantillons

Il s'agit ici de répondre à la question :

« Deux échantillons ont-ils été prélevés dans la même population? »

ou encore

« L'écart entre deux échantillons est-il dû au hasard ou est-il significatif? »

En pratique, on se contente de comparer les échantillons au travers de valeurs remarquables telles que la moyenne, variance, ….

➢ Le cas usuel consiste à comparer des moyennes (ou proportions) sur de grands échantillons. Il suffit alors de faire la différence entre les deux moyennes (proportions) et de tester si la différence est nulle (cf. TD)

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \Longleftrightarrow \begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

➢ Les autres cas (échantillons de petites tailles, tests d'égalité de variance, ...) sont à étudier au cas par cas : Test de Student, Test de Fisher, Test de Wilcoxon-Mann-Whitney,… (cf. Fiches).

Tests d'adéquation

Il s'agit ici de répondre à la question

« Est-ce que l'échantillon suit une loi usuelle (donnée)? »

Echantillon gaussien? Loi uniforme? Loi de Poisson? Etc…

Les tests d'adéquation mesure la différence entre la distribution de la loi usuelle, F, et la distribution de l'échantillon, F_n. Notons $d(F,F_n)$ l'écart entre F et F_n. On fait le test :

$$\begin{cases} H_0 : d(F,F_n)=0 \\ H_1 : d(F,F_n)>0 \end{cases}$$

La région critique : $W=\{d(F,F_n)>C\}$

La règle de décision :

- **si $d(F,F_n)<C$ alors je considère que l'échantillon suit la loi usuelle**
- **si $d(F,F_n)>C$ alors j'en déduis qu'il ne suit pas la loi usuelle avec un risque α de me tromper**

Pour construire complètement ce test, il faut déterminer

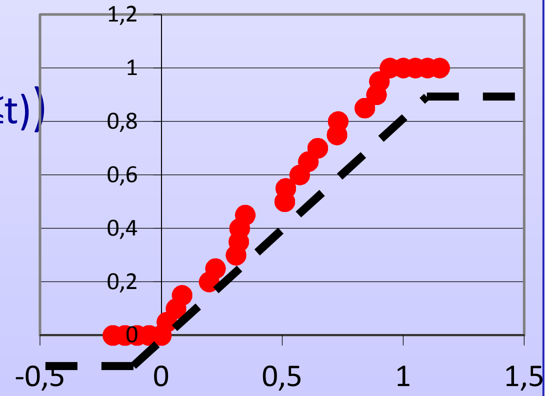
- **Ce que représentent F et F_n**
- **Comment on calcule l'écart $d(F,F_n)$**
- **La loi de probabilité de $d(F,F_n)$**

→

- **Test du chi-deux**
- **Test de Kolmogorov-Smirnov**
- **etc …**

Tests d'adéquation de Kolmogorov-Smirnov (1/2)

lois continues



- F = fonction de répartition de la loi usuelle ($t\in\mathbb{R}$, $F(t)=P(X\leq t)$)
- F_n = fonction de répartition empirique de l'échantillon,

$$t \mapsto F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \leq t\}}$$

où 1_A est la fonction indicatrice de A.

- Ecart :

$$d(F,F_n) = \sup_{x\in\mathbb{R}} \left| F(x) - F_n(x) \right| = K_n$$

Sup sur $\mathbb{R}$ impossible à calculer

Méthodologie :

1) Ordonner l'échantillon : $X_{(1)}<X_{(2)}<\ldots<X_{(n)}$.
2) Calculer

$$K_n = \max_{i=1,\ldots,n} \left\{ \left| F\left(X_{(i)}\right) - \frac{i}{n} \right|, \left| F\left(X_{(i)}\right) - \frac{i-1}{n} \right| \right\}$$

N.B.

$$F_n(x) = \begin{cases} 0 & \text{si} \quad x < X_{(1)} \\ \frac{i}{n} & \text{si} \quad X_{(i)} \leq x < X_{(i+1)} \\ 1 & \text{si} \quad x \geq X_{(n)} \end{cases}$$

Sous l'hypothèse H_0, la loi de $\sqrt{n}\,K_n$ est connue

Tests d'adéquation de Kolmogorov-Smirnov (2/2)

lois continues

- soit on calcule le seuil C à partir de la table de loi : $\alpha=P(K_n>C \mid H_0 \text{ vraie})$

Table des valeurs de C en fonction de n et α

n	α=0,10	α=0,05	α=0,01
5	0,509	0,563	0,669
10	0,369	0,409	0,486
15	0,304	0,338	0,404
20	0,265	0,294	0,352
25	0,238	0,264	0,317
30	0,218	0,242	0,290
40	0,189	0,210	0,252
n>40	1,22/√n	1,36/√n	1,63/√n

- soit on calcule la p-valeur du test,

$$p(t) \approx 2\sum_{j=1}^{N}(-1)^{j+1} e^{-2j^2t^2}$$

où t= $\sqrt{n}\, k_n$ est la valeur de la variable de décision prise sur l 'échantillon (N~3)

Remarque : Dans le cas d'une loi normale, il existe un test plus puissant que celui de Kolmogorov, le **test de Shapiro-Wilk**. On peut aussi avoir une représentation graphique permettant de juger si l'échantillon suit une loi normale, la **droite de Henry** (ou **qq-plot**).

EXEMPLE

Comment vérifier si un échantillon de 15 individus est gaussien de loi N(0,1)?

D'après la table (slide 17) du test, pour n=15 et α=5%, on accepte que l'échantillon suit une loi N(0,1) si $k_n<0{,}338$.

Calcul de k_n

- Ordonner la série
- Calculer $F(x_i)$ où F fonction de répartition de N(0,1)
- Calculer

$$\left|F\left(X_{(i)}\right)-\frac{i}{n}\right| \quad \text{et} \quad \left|F\left(X_{(i)}\right)-\frac{i-1}{n}\right|$$

- Prendre le max

ici $k_n=0{,}2520<0{,}338$

i	x_i	$F(x_i)$	$\lvert F(x_i)-i/n \rvert$	$\lvert F(x_i)-(i-1)/n \rvert$
1	-7,64	0	0,0667	0
2	-5,23	0	0,1333	0,0667
3	-1,84	0,0329	0,1671	0,1004
4	-1,25	0,1056	0,1610	0,0944
5	-1,02	0,1539	0,1795	0,1128
6	-0,65	0,2578	0,1422	0,0755
7	-0,01	0,4960	0,0293	0,0960
8	0,06	0,5239	0,0094	0,0573
9	0,12	0,5478	0,0522	0,0144
10	0,24	0,5948	0,0718	0,0052
11	0,74	0,7704	0,0370	0,1037
12	2,18	0,9854	0,1854	0,2520
13	2,85	0,9978	0,1311	0,1978
14	6,42	1	0,0667	0,1333
15	28,54	1	0	0,0667

Série ordonnée

Calcul de la p-valeur

$$t=\sqrt{15}\times k_n=0{,}976$$

$$p(0{,}976)\approx 2\sum_{j=1}^{3}(-1)^{j+1}e^{-2j^2 0{,}976^2} \approx 0{,}3$$

signifie qu'il faudrait α=30% pour rejeter H_0 à tord

Tests d'adéquation du chi-deux

lois discrètes

Loi discrète de modalités x_i avec la probabilité p_i, i=1,…,k

- F = Effectif théorique : $n \times p_i$
- F_n = Effectif observé : N_i
- Ecart :

$$d(F,F_n) = D_n = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

Nb modalités -1

Sous l'hypothèse H_0, D_n suit une loi du chi-deux à (k-1) d.d.l. , d'où

$\alpha = P(D_n > C | H_0 \text{ vraie})$ $\Rightarrow$ valeur de C dans la table du chi-deux

- **Validité du test : effectif théorique $np_i \geq 5$ sinon regroupement de modalités**
- Si la loi usuelle dépend de paramètres (Poisson, …) et que ces paramètres sont estimés sur l'échantillon alors on enlève autant de d.d.l. à la loi du chi-deux que de paramètres estimés.
- Possibilité de faire le test d'adéquation du chi-deux à loi continue avec un regroupement en classe (mais perte d'information)

EXEMPLE

Comment vérifier si un dé est truqué à partir du résultat de 60 lancers?

L'hypothèse à tester est celle d'une loi uniforme discrète sur les 6 faces chacune de probabilité $p_i=1/6$.

Sous l'hypothèse H_0, D_n suit une loi du chi-deux à 6-1=5 d.d.l. D'où pour $\alpha=5\%$

$$0{,}05=P(D_n>C) \overset{\text{Table loi}}{\Longrightarrow} C=11{,}07$$

- Si $d_n>11{,}07$ alors on admet que le dé est truqué avec un risque de 5% d'erreur
- Si $d_n<11{,}07$ alors on admet que le dé est équilibré

<u>Calcul de d_n</u>

Effectif théorique : $np_i=60\times(1/6)=10$ **≥ 5 (condition d'application du test validée)**

Modalité	x_i	1	2	3	4	5	6
Effectif théorique	N_i	11	8	9	12	7	13
Effectif observé	np_i	10	10	10	10	10	10
	$\frac{(N_i-np_i)^2}{np_i}$	0,1	0,4	0,1	0,4	0,9	0,9

$\Rightarrow d_n=2{,}8<11{,}07$

Test d'indépendance du chi-deux

Il s'agit ici de répondre à la question :

«Est-ce que deux variables qualitatives sont indépendantes »

X variable qualitative à *r* modalités

Y variable qualitative à *s* modalités

On fait le test

$$\begin{cases} H_0 : & p_{ij}=p_{i.}\times p_{.j} \\ H_1 : & p_{ij}\neq p_{i.}\times p_{.j} \end{cases}$$

X et Y sont indépendantes si
loi conjointe (p_{ij}=) = produit des lois marginales
($p_{i.}$= loi de X / $p_{.j}$ = loi de Y)

On compare alors le tableau de contingence des effectifs observés ($n_{i,j}$) avec le tableau des effectifs théoriques ($t_{i,j}$) s'il y a indépendance (cf stat. descriptive ING1)

On calcule l'écart entre les deux tableaux avec la « distance » du chi-deux

$$d\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{i,j}-t_{i,j}\right)^2}{t_{i,j}}$$

La région critique est de la forme $W=\{d\chi^2>C\}$.
Sous l'hypothèse H_0, on peut approcher la loi de $d\chi^2$ par une loi du chi-deux à (r-1)(s-1) d.d.l. , d'où

$$\alpha=P(d\chi^2>C|H_0 \text{ vraie}) \quad \Rightarrow \quad \text{valeur de C dans la table du chi-deux}$$

EXEMPLE

Le proviseur d'un lycée désire comparer le taux de réussite au baccalauréat des trois sections générales.

	L	*ES*	*S*
Réussite	41	59	54
Echec	21	36	75

Sous l'hypothèse H_0, $d\chi^2$ suit une loi du chi-deux à (3-1)(2-1)=2 d.d.l. D'où pour α=5%

$$0{,}05=P(D_n>C) \overset{\text{Table loi}}{\Longrightarrow} C=5{,}991$$

- Si $d\chi^2$ >5,991 alors on admet que le taux de réussite est différent suivant la section avec un risque de 5% d'erreur
- Si $d\chi^2$ <5,991 alors on admet que la section n'a pas d'impact sur le taux de réussite

Effectifs observés

	L	ES	S	**Total**
Réussite	41	59	54	**154**
Echec	21	36	75	**132**
Total	**62**	**95**	**129**	**286**

/286

Fréquences observées

	L	ES	S	**Total**
Réussite	0,14	0,21	0,19	**0,54**
Echec	0,07	0,13	0,26	**0,46**
Total	**0,22**	**0,33**	**0,45**	**1,00**

$f_{ij} = f_{i.} \times f_{.j}$

Fréquences théoriques

	L	ES	S	**Total**
Réussite	0,12	0,18	0,24	**0,54**
Echec	0,10	0,15	0,21	**0,46**
Total	**0,22**	**0,33**	**0,45**	**1,00**

×286

Effectifs théoriques

	L	ES	S	**Total**
Réussite	33,38	51,15	69,46	**154**
Echec	28,62	43,85	59,54	**132**
Total	**62**	**95**	**129**	**286**

$d\chi^2 = 13{,}83 > 5{,}991$