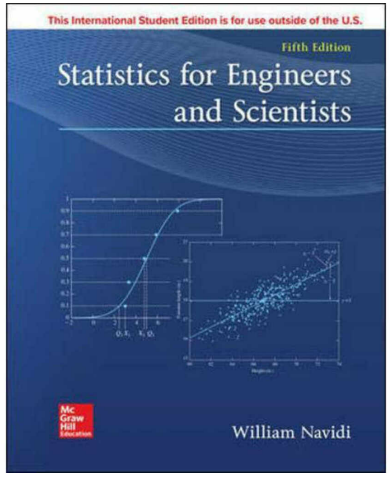
Introduction aux statistiques pour l'Ingénieur
Comparaison de deux échantillons

Stéphane Canu
asi.insa-rouen.fr/enseignants/~scanu
scanu@insa-rouen.fr



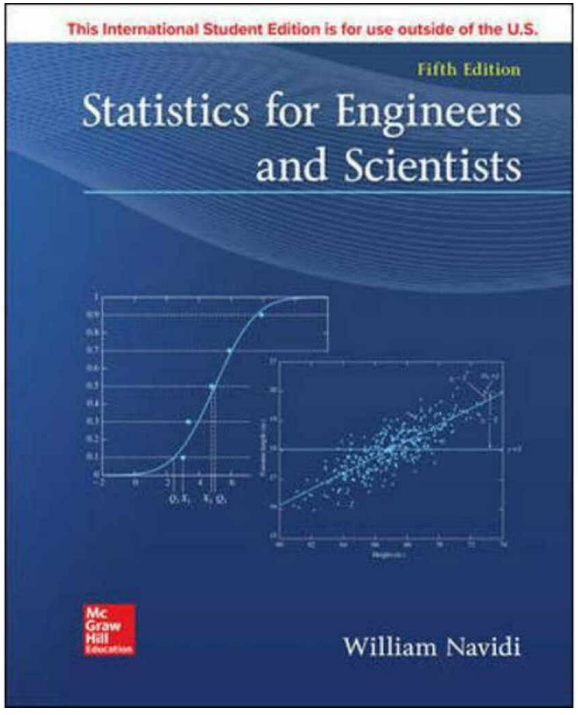
ITI 3, INSA Rouen Normandie, mai 2022

Lecture road map

1. **Comparaison de deux échantillons**

2. **Les 2 variables sont quantitatives)**
   - Les test de Student (cas gaussien)
   - Les tests de rang (distribution inconnue)

3. **Les test du chi 2**
   - La loi du χ^2
     - Théorème du χ^2 (Pearson)
     - Conditions d'utilisation du test du χ^2 d'indépe



https://moodle.insa-rouen.fr/course/view.php?id=93

Comparaison de deux échantillons

sur 140 thèses en médecine :

- 118 utilisent une méthode statistique
- 87 présentaient des tests statistiques.
  - 53 tests du χ^2
  - 33 tests de Student.

Le succès de ces 2 tests est dû à la question à laquelle ils répondent

il y a t'il oui ou non une relation entre ces deux variables ?

- deux variables qualitatives : test du χ^2
  - Les jurys de Rouen et de Lyon ont-il noté de la même façon ?
- une variable qualitative et l'autre quantitative : Analyse de la variance
  - ces deux variétés ont elles le même rendement ?
- deux variables quantitatives : test de Student
  - y a t'il une relation entre la concentration d'ozone et la température ?

Question et décisions

Questions ?

- Ce nouveau médicament est-il meilleur que celui que l'on utilise ?
  - test en double aveugle
  - test avant/après
- Les jurys de Rouen et de Lyon ont-il noté de la même façon ?

Deux échantillons (éventuellement pas de même taille)

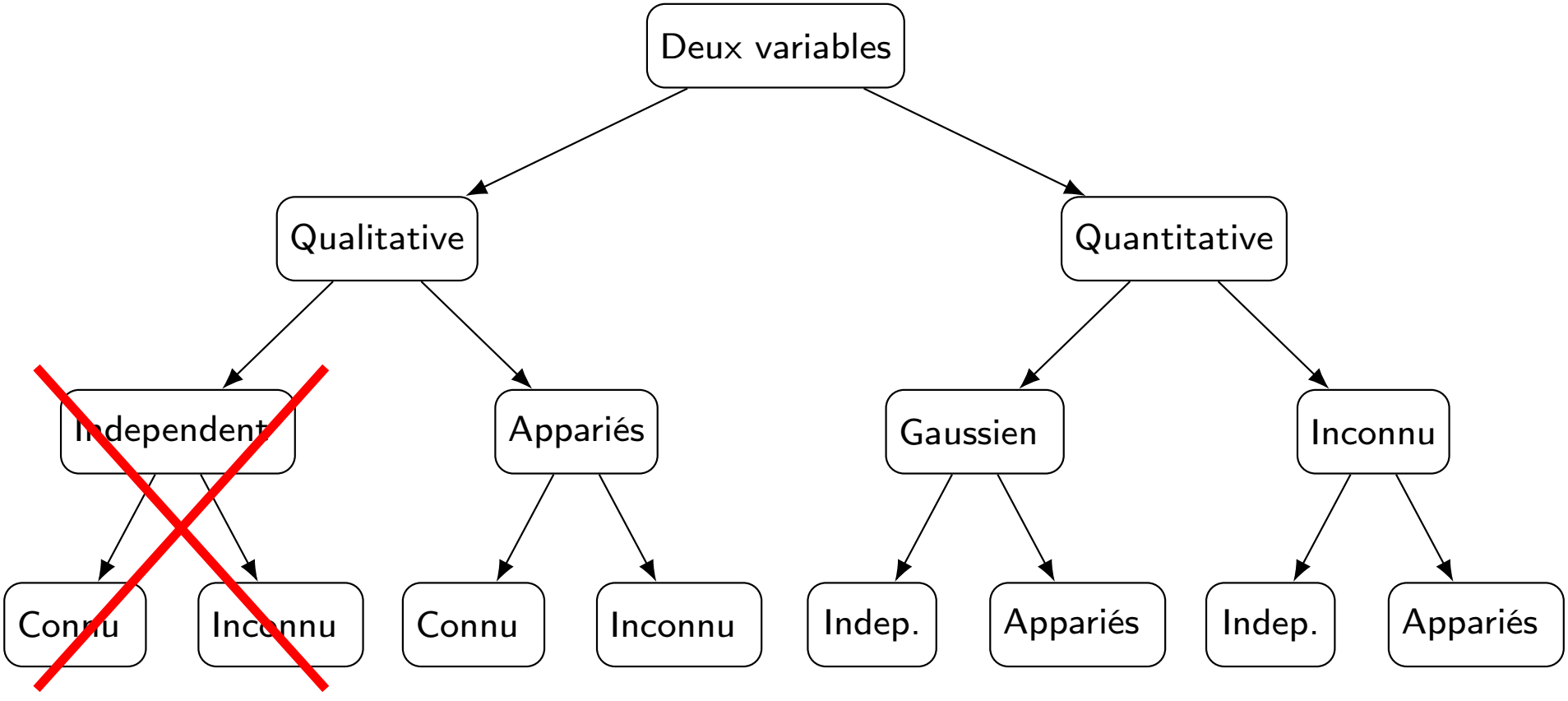
$$x_1, \quad \ldots, \quad x_n$$
$$y_1, \quad \ldots, \quad y_m$$

Décisions ?

- Les hypothèses :
  - les deux échantillons sont issus de la même loi parente
  - les loi sont différentes
- H_0

Différents types d'échantillons : 3 distinctions

- Qu'observe t'on ? Variables qualitatives / Variables quantitatives
- Comment observe t'on ? Échantillons indépendant / appariés
- Quelles sont nos hypothèses à priori ? Modèle Gaussien / inconnu



Différents types d'échantillons : 3 distinctions

1. Ces deux variétés ont elles le même rendement, hypothèse normale
2. Les jurys de Rouen et de Lyon ont-il noté de la même façon ?
3. Taux de pollution à Paris avant et après l'installation d'un dispositif de dépollution, hypothèse normale
4. L'effet de la prise du médicament sur le taux de calcium dans le sang contre un placebo
5. Est-ce que ce médicament augment le taux de calcium dans le sang ?
6. Lequel de ces deux réseaux de neurones donne le taux de classification le plus faible ?

Quantitatif / indépendant / modèle Gaussien

La question : le taux de pollution à Paris est-il le même qu'à Berlin ?

Le modèle $\begin{cases} \text{Taux de pollution à Paris : } X \sim \mathcal{N}(\mu_X, \sigma_X^2) \\ \text{Taux de pollution à Berlin : } Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \end{cases}$

X = Paris	X_1	...	...	...	X_i	...	...	...	X_n	
Y = Berlin	Y_1	...	...	...	...	Y_i	...	...	...	Y_m

Analogue à une variable qualitative (ville) et une variable quantitative (taux)

$$\overline{X} \sim \mathcal{N}\left(\mu_X, \frac{\sigma_X^2}{n}\right) \quad \text{et} \quad \overline{Y} \sim \mathcal{N}\left(\mu_Y, \frac{\sigma_Y^2}{m}\right)$$

Les hypothèses

$$\begin{cases} \mathcal{H}_0 : \text{les sont est les mêmes :} & \mu_X = \mu_Y \text{ et } \sigma_X = \sigma_Y \\ \mathcal{H}_1 : \text{les taux sont différents :} & \mu_X \neq \mu_Y \end{cases}$$

La stratégie du test

1. On commence par vérifier l'hypothèse d'égalité des variances
2. Si l'hypothèse d'égalité des variances n'est pas rejetée, on peut tester l'égalité des espérances

Quantitatif / indépendant / modèle Gaussien

On commence par vérifier l'hypothèse d'égalité des variances

Test de Fisher : test de d'égalité des variances de 2 échantillons gaussiens

Les hypothèses (avec le modèle gaussien)

$$\begin{cases} \mathcal{H}_0 : \text{les sont est les mêmes :} & \sigma_X = \sigma_Y \quad \text{ou} \quad \frac{\sigma_X}{\sigma_Y} = 1 \\ \mathcal{H}_1 : \text{les taux sont différents :} & \sigma_X \neq \sigma_Y \quad \text{ou} \quad \frac{\sigma_X}{\sigma_Y} \neq 1 \end{cases}$$

La statistique Sous $\mathcal{H}_0$, la loi de Fisher(-Snedecor) ou loi F de Snedecor

$$\widehat{F} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}{\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2} \sim \mathcal{F}(n-1, m-1)$$

P-valeur (fractiles de la loi de Fisher si $\widehat{F} < 1$)

$$\text{p-val } = \mathbb{P}\left(1/\widehat{F} < \mathcal{F}(n-1, m-1)\right) + \mathbb{P}\left(\widehat{F} > \mathcal{F}(n-1, m-1)\right)$$

https://www.itl.nist.gov/div898/handbook/eda/section3/eda359.htm

Quantitatif / indépendant / modèle Gaussien

Si l'hypothèse d'égalité des variances n'est pas rejetée, on peut tester l'égalité des espérances

Test de Student : test de d'égalité des espérances de 2 échantillons gaussiens

Les hypothèses $\begin{cases} \mathcal{H}_0 : \text{les sont est les mêmes :} & \mu_X = \mu_Y \\ \mathcal{H}_1 : \text{les taux sont différents :} & \mu_X \neq \mu_Y \end{cases}$

La statistique Sous $\mathcal{H}_0$,

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(0, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right) \quad \text{et} \quad \widehat{T} = \frac{\overline{X} - \overline{Y}}{\sqrt{\widehat{\sigma}^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim \mathcal{T}_{n+m-2}$$

avec $\widehat{\sigma}^2 = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{m}(Y_i - \overline{Y})^2\right)$

P-valeur (fractiles de la loi de student)

$$\text{p-val} = \mathbb{P}\left(|\widehat{T}| < \mathcal{T}_{n+m-2}\right) + \mathbb{P}\left(-|\widehat{T}| > \mathcal{T}_{n+m-2}\right)$$

https://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm

Quantitatif / indépendant / modèle Gaussien

Les observations

$X =$ Paris	5,33	6.13	5,66	4,50	5,35	6,32	4,24	5,83	6,27	
$Y =$ Berlin	5,32	6,00	5,64	4,59	5,19	6,17	4,11	5,86	6,13	4.68

Test de Fisher :

$$\widehat{F} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}{\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2} = \frac{0.90}{0.77} = 1.16$$

p-val $= \mathbb{P}\big(1/1.16 < \mathcal{F}(8,9)\big) + \mathbb{P}\big(1.16 > \mathcal{F}(8,9)\big) = 0.422 + 0.412 = 0.834$

Conclusion (décision) on garde $\mathcal{H}_0$, on peut faire le test de student

Test de student :

$\widehat{\sigma}^2 = 1/17 sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{m}(Y_i - \overline{Y})^2 = (8.10 + 7.74)/17 = 0.93$

$$\widehat{T} = \frac{\overline{X} - \overline{Y}}{\sqrt{\widehat{\sigma}^2\left(\frac{1}{n} + \frac{1}{m}\right)}} = 0,97$$

p-val $= \mathbb{P}\left(0,97 < \mathcal{T}_{17}\right) + \mathbb{P}\left(-0,97 > \mathcal{T}_{17}\right) = 0,17$

Conclusion on garde $\mathcal{H}_0$: les observations ne permettent pas de distinguer le taux de pollution de Paris de celui de Berlin

Quantitatif / appariés / modèle Gaussien

La question : le taux de pollution à Paris est-il le même après l'installation d'un dispositif de dépollution

Le modèle

$X =$ Avant	X_1	...	...	...	X_i	...	...	...	X_n
$Y =$ Après	Y_1	...	...	...	Y_i	...	...	...	Y_n
$D = X - Y$	D_1	...	...	...	D_i	...	...	...	D_n

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2), \quad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2), \quad D = X - Y \sim \mathcal{N}(\mu_X - \mu_Y, \sigma_D^2)$$

Les hypothèses

$$\begin{cases} \mathcal{H}_0 : \text{les taux sont est les mêmes :} & D \sim \mathcal{N}(0, \sigma_D^2) \\ \mathcal{H}_1 : \text{les taux sont différents :} & D \sim \mathcal{N}(\mu_D, \sigma_D^2) \text{ avec } \mu_D \neq 0 \end{cases}$$

La statistique : Sous $\mathcal{H}_0$,

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \qquad \frac{\overline{D}}{\sqrt{\frac{\widehat{\sigma}^2}{n}}} \sim \mathcal{T}_{n-1}$$

Individu	l	2	3	4	5	6	7	8	9	10
Avant	5,33	6.13	5,66	4,50	5,35	6,32	4,24	5,83	6,27	4,86
Après	5,32	6,00	5,64	4,59	5,19	6,17	4,11	5,86	6,13	4,68

Quantitatif / indépendant / modèle inconnu

Le médicament augmente le taux de calcium dans le sang : comparaison en double aveugle avec un placebo

Les hypothèses $\begin{cases} \mathcal{H}_0 : \text{le médicament n'a pas d'effet} \\ \mathcal{H}_1 : \text{le taux de calcium a augmenté} \end{cases}$

Test de rang : Wilcoxon Rank Sum test ou encore Wilcoxon two-sample test. Le test de Mann-Whitney est légèrement différent

Les observations :

Médicament :	8.50	9.48	8.65	8.16	7.76	8.25			
Placebo :	8.27	8.20	8.63	8.14	9.00	8.10	7.20	8.32	7.70

On mélange et on trie

Rang :	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Obs. :	7.2	7.7	7.76	8.1	8.14	8.16	8.2	8.25	8.27	8.32	8.5	8.63	8.65	9.0	9.48

La somme des rang des observations avec le médicament est de

$$w_a = 3 + 6 + 8 + 11 + 13 + 15 = 56$$

Quantitatif / indépendant / modèle inconnu

La somme des rang des observations avec le médicament est de

$$w_a = 3 + 6 + 8 + 11 + 13 + 15 = 56$$

Le modèle Sous H_0, la somme des rangs $\mathbb{E}(W_a) = \mu_W$ et $Var(W_a) = \sigma_W^2$ avec $\mu_W = \frac{1}{2} n_1(n_1 + n_2 + 1) = 48$ et $\sigma_W^2 = \frac{1}{12} n_1 n_2 (n_1 + n_2 + 1) = 72$

La statistique : Sous $\mathcal{H}_0$, si n est grand...

$$\frac{W_a - \mu_W}{\sqrt{\sigma_W^2}} \sim \mathcal{N}(0, 1)$$

La p-valeur :

$$\text{P-val} = \mathbb{P}\left(Z > \frac{w_a - \mu_W}{\sqrt{\sigma_W^2}} = 0.943 \right) = 0.173 \qquad \text{avec } Z \sim \mathcal{N}(0, 1)$$

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ranksums.html

Wilcoxon distribution

1-tail / 2-tail		$\alpha = 0.025$ / $\alpha = 0.05$			$\alpha = 0.05$ / $\alpha = 0.10$			1-tail / 2-tail		$\alpha = 0.025$ / $\alpha = 0.05$			$\alpha = 0.05$ / $\alpha = 0.10$		
m	n	W	d	P	W	d	P	m	n	W	d	P	W	d	P
3	3				6 15	1	.0500	5	10	23 57	9	.0200	26 54	12	.0496
3	4				6 18	1	.0286	5	11	24 61	10	.0190	27 58	13	.0449
3	5	6 21	1	.0179	7 20	2	.0357	5	12	26 64	12	.0242	28 62	14	.0409
3	6	7 23	2	.0238	8 22	3	.0476	5	13	27 68	13	.0230	30 65	16	.0473
3	7	7 26	2	.0167	8 25	3	.0333	5	14	28 72	14	.0218	31 69	17	.0435
3	8	8 28	3	.0242	9 27	4	.0424	5	15	29 76	15	.0209	33 72	19	.0491
3	9	8 31	3	.0182	10 29	5	.0500	5	16	30 80	16	.0201	34 76	20	.0455
3	10	9 33	4	.0245	10 32	5	.0385	5	17	32 83	18	.0238	35 80	21	.0425
3	11	9 36	4	.0192	11 34	6	.0440	5	18	33 87	19	.0229	37 83	23	.0472
3	12	10 38	5	.0242	11 37	6	.0352	5	19	34 91	20	.0220	38 87	24	.0442
3	13	10 41	5	.0196	12 39	7	.0411	5	20	35 95	21	.0212	40 90	26	.0485
3	14	11 43	6	.0235	13 41	8	.0456	5	21	37 98	23	.0243	41 94	27	.0457
3	15	11 46	6	.0196	13 44	8	.0380	5	22	38 102	24	.0234	43 97	29	.0496
3	16	12 48	7	.0237	14 46	9	.0423	5	23	39 106	25	.0226	44 101	30	.0469
3	17	12 51	7	.0202	15 48	10	.0465	5	24	40 110	26	.0219	45 105	31	.0445
3	18	13 53	8	.0233	15 51	10	.0398	5	25	42 113	28	.0246	47 108	33	.0480
3	19	13 56	8	.0201	16 53	11	.0435	6	6	26 52	6	.0206	28 50	8	.0465
3	20	14 58	9	.0232	17 55	12	.0469	6	7	27 57	7	.0175	29 55	9	.0367
3	21	14 61	9	.0203	17 58	12	.0410	6	8	29 61	9	.0213	31 59	11	.0406
3	22	15 63	10	.0230	18 60	13	.0443	6	9	31 65	11	.0248	33 63	13	.0440
3	23	15 66	10	.0204	19 62	14	.0473	6	10	32 70	12	.0210	35 67	15	.0467
3	24	16 68	11	.0229	19 65	14	.0421	6	11	34 74	14	.0238	37 71	17	.0491
3	25	16 71	11	.0205	20 67	15	.0449	6	12	35 79	15	.0207	38 76	18	.0415
4	4	10 26	1	.0143	11 25	2	.0286	6	13	37 83	17	.0231	40 80	20	.0437
4	5	11 29	2	.0159	12 28	3	.0317	6	14	38 88	18	.0204	42 84	22	.0457

Elte University (2016) Table of critical values for the Wilcoxon rank-sum test
http://plantsys.elte.hu/oktatas/Biometria/tablazatok/Wilcoxon_table_ketmintas_probahoz.pdf

Quantitatif / appariés / modèle inconnu

La prise du médicament fait diminuer le taux de potassium dans le sang

Test du signe

Individu	I	2	3	4	5	6	7	8	9	10
Avant	5,33	6.13	5,66	4,50	5,35	6,32	4,24	5,83	6,27	4,86
Après	5,32	6,00	5,64	4,59	5,19	6,17	4,11	5,86	6,13	4,68
Signe	-	-	-	+	-	-	-	+	-	-

La statistique et le modèle Sous H_0, N le nombre de signes + suit une loi binomiale $N \sim \mathcal{B}(n = 10, p = 1/2)$

La p-valeur :

$$\text{P-val} = \mathbb{P}\left(N \leq n = 2\right) = 0.05469 \qquad \text{avec } N \sim \mathcal{B}(n = 10, p = 1/2)$$

Conclusion on garde $\mathcal{H}_0$: le médicament n'a pas d'effet (il faudrait tester plus d'individus, augmenter n).

Quantitatif / appariés / modèle inconnu

La prise du médicament fait diminuer le taux de potassium dans le sang

Test de Wicoxon signé

Individu	I	2	3	4	5	6	7	8	9	10
Avant	5,33	6.13	5,66	4,50	5,35	6,32	4,24	5,83	6,27	4,86
Après	5,32	6,00	5,64	4,59	5,19	6,17	4,11	5,86	6,13	4,68
Signe Diff.	0,01	0,13	0,02	-0,09	0,16	0,15	0,13	-0,03	0,14	0,18
Trié $\lvert$ diff.$\rvert$	0,01	0,02	0,03	0,09	0,13	0,13	0,14	0,15	0,16	0,18
rang	1	2	3	4	5	6	7	8	9	10
signe	+	+	-	-	+	+	+	+	+	+

La statistique et le modèle Sous H_0, la somme des rangs de signes + suit la loi de Wilcoxon signée, asymptotiquement gaussienne $W_A \sim \mathcal{N}(\mu_A, \sigma_A^2)$

avec $\mu_A = \dfrac{n(n+1)}{4} = 27,5$ et $\sigma_A^2 = \dfrac{n(n+1)(2n+1)}{24} = 96,25$

La p-valeur : on observe $w_A = 1+2+5+6+7+8+9+10 = 48$

$$\text{P-val} = \mathbb{P}\left(W_A \leq w_A = 48\right) \approx 0,04$$

Conclusion on rejette $\mathcal{H}_0$: le médicament a bien un effet.

Qualitatif / appariés / modèle inconnu

Le test du chi 2

La question : est-ce que les deux jury ont noté de la même façon ?

- Hypothèses : choix de l'état de référence $\mathcal{H}_0$

$$\begin{cases} \mathcal{H}_0 : \text{les jurys sont identiques} \\ \mathcal{H}_1 : \text{les jurys sont différents} \end{cases}$$

- Modèle : si les jurys sont identiques alors les variables sont indépendantes $\mathbb{P}(\text{reçu par le jury 1}) = \mathbb{P}(\text{reçu})\mathbb{P}(\text{passer par le jury 1})$
- Décision :
  - calcul de la $p - \textit{valeur}$. Si l'on admet l'équivalence des jurys, quelle est la probabilité d'observer un tableau encore plus « différent » que celui qu'on a.
  - Prise de décision : si cette probabilité est faible (typiquement inférieure à $0,05$), on rejette l'hypothèse $\mathcal{H}_0$

Le tableau de référence

Comment calculer « la probabilité d'observer un tableau » ?
Pour répondre à cette question, on peut construire un tableau théorique,

O	reçu	refusé	
Jury 1	50	5	55
Jury 2	47	14	61
Jury 3	56	8	64
	153	27	180

T	reçu	refusé	
Jury 1	46,75	8,25	30,56%
Jury 2	51,85	9,15	33,89 %
Jury 3	54,40	9,60	35,56 %
	85 %	15 %	180

Table: Données marginalisées (à gauche : les effectif marginaux sont en gris) et effectifs théoriques sous $\mathcal{H}_0$ (à droite en bleu)

$$\frac{N_{\bullet j}}{n} = \widehat{p}_{\bullet j} \to \frac{153}{180} = 0,85 \qquad \underbrace{0,85 \times 0,30}_{\widehat{p}_{ij}=\widehat{p}_{i\bullet}\widehat{p}_{\bullet j}} \times 180 = \frac{153 \times 55}{180} = 46,75$$

Comment mesurer la distance

$D(O, T)$ entre le tableau observé (O) et le tableau théorique (T)

Comment mesurer la distance entre les tableaux

Definition (Distance du χ^2)

Soit O un tableau de contingence de I lignes et J colonnes d'effectif total n. Soit T un tableau de probabilité de même dimension. On appelle distance du χ^2 entre les tableaux O et T la quantité

$$D(O,T) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - T_{ij})^2}{T_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(N_{ij} - n\,\widehat{p}_{ij})^2}{n\,\widehat{p}_{ij}}$$

$O_{ij} = N_{ij}$ les effectifs observés

$T_{ij} = n\,\widehat{p}_{ij}$ les effectifs théoriques sous hypothèse d'indépendance

n l'effectif total

$\widehat{p}_{ij}$ la probabilité estimée sous hypothèse d'indépendance

Exemple :

$$\begin{aligned} D(O,T) &= \frac{(50-46{,}75)^2}{46{,}75} + \frac{(5-8{,}25)^2}{8{,}25} + \frac{(47-51{,}85)^2}{51{,}85} + \frac{(14-9{,}15)^2}{9{,}15} + \frac{(56-54{,}40)^2}{54{,}40} + \frac{(8-9{,}60)^2}{9{,}60} \\ &= 0{,}2259 + 1{,}2803 + 0{,}4537 + 2{,}5708 + 0{,}0471 + 0{,}2667 \\ &= 4{,}84 \end{aligned}$$

Distance du χ^2 est-elle grande ?

$$D(O,T) = 4,84$$

Est-ce Grand ? p-valeur $= \mathbb{P}\big(D(O,T) \geq 4,84\big)$

O	reçu	refusé	
Jury 1	50	5	55
Jury 2	47	14	61
Jury 3	56	8	64
	153	27	180

T	reçu	refusé	
Jury 1	46,75	8,25	30,56%
Jury 2	51,85	9,15	33,89 %
Jury 3	54,40	9,60	35,56 %
	85 %	15 %	180

sous l'hypothèse $\mathcal{H}_0$ D est distribué suivant une loi du χ^2 à $(3-1)(2-1) = 2$ degrés de libertés :

$$\boxed{p\text{-valeur} = \mathbb{P}(D \geq 4,84) = 0,0887}$$

On estime à $8,9\%$ la probabilité d'observer un tableau encore plus différent que celui que l'on a. On conclue que la distance n'est pas très grande et que l'on ne peut pas rejeter l'hypothèse d'indépendance des jurys.

Il est très peu vraisemblable que $D(O,T) = 0$. Autrement dit, des résultats trop loin des prévisions sont mauvais, mais des résultats trop bons ont vraisemblablement été falsifiés...

Theorem (Théorème du χ^2 (Pearson))

$$X_i = \frac{N_i - n\,\widehat{p}_i}{\sqrt{n\,\widehat{p}_i}} \qquad \sum_{i=1}^{I} X_i^2 \longrightarrow \chi^2_{I-1}$$

$$X_{ij} = \frac{N_{ij} - n\,\widehat{p}_{ij}}{\sqrt{n\,\widehat{p}_{ij}}} \qquad \sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}^2 \longrightarrow \chi^2_{(I-1)(J-1)}$$

Eléments de preuve dans le cas d'une variable à $I = 2$ modalités.

dans ce cas on a $n = N_1 + N_2$ et $p_1 + p_2 = 1$ et $N_1 \sim \mathcal{B}(n, p_1)$. Pour des échantillons assez grand on peut accepter une estimation gaussienne

$$\frac{N_1 - np_1}{\sqrt{np_1(1-p_1)}} \sim \mathcal{N}(0,1) \qquad \Longrightarrow \qquad \frac{(N_1 - np_1)^2}{np_1(1-p_1)} \sim \chi^2_1$$

et

$$\begin{aligned}
\tfrac{(N_1-np_1)^2}{np_1(1-p_1)} &= \tfrac{(N_1-np_1)^2}{np_1(1-p_1)}(1-p_1+p_1)\\
&= \tfrac{(N_1-np_1)^2}{np_1} + \tfrac{(N_1-np_1)^2}{n(1-p_1)}\\
&= \tfrac{(N_1-np_1)^2}{np_1} + \tfrac{(N_1-np_1-n+n)^2}{n(1-p_1)}\\
&= \tfrac{(N_1-np_1)^2}{np_1} + \tfrac{(N_2-np_2)^2}{n(1-p_1)} \qquad = \sum_{i=1}^{2} \frac{(N_i - np_i)^2}{np_i}
\end{aligned}$$

□

Mise en œuvre du test du χ^2

1. on construit un tableau de contingence O des observations (2 variables qualitatives de respectivement I et J modalités)
2. on calcule les marginales $p_i = \frac{1}{n}\sum_{j=1}^{J} O_{ij}$
3. on calcule pour chaque case du tableau des effectifs théoriques $T_{ij} = np_ip_j$ (en supposant l'indépendance)
4. on calcule la distance du χ^2

$$D(O,T) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$

5. on calcule le nombre de degrés de liberté du χ^2 : $d = (I-1)(J-1)$
6. on regarde dans les tables d'une variable aléatoire Z distribué suivant une loi χ^2 à d degrés de liberté la p-valeur de $D(O,T)$

$$pval = \mathbb{P}\big(Z \geq D(O,T)\big)$$

7. on décide qu'on ne peut pas conclure à la dépendance si la p-valeur est supérieure à 0,05, si $pval \geq 0,05$

Conditions d'utilisation du test du χ^2

- des observations tirées au hasard
- des observations indépendantes
- n suffisamment grand
- des Effectifs > 5 pour chaque élément du tableau

Conclusion

La réalité est bien plus complexe...