Eléments de statistiques pour les data sciences

Cours 12: inférence non paramétrique - test de Wilcoxon

Dr. Ph. Müllhaupt

IGM - EPFL

plan

1. rappel sur les tests paramétriques de centrage et d'étalement
2. statistique d'ordre de rang (ou statistique de rang)
3. propriété des stastistiques d'ordre
4. test de Wilcoxon

rappel sur les inférences paramétriques

centrage et étalement

- χ^2 pour l'adéquation statistique (Pearson)
- χ^2 pour $D = -2 \log r$ avec r le rapport de vraisemblance
- $Z \sim \mathcal{N}(0,1)$ pour le comportement asymptotique (grands échantillons)
- $Z \sim \mathcal{N}(0,1)$ pour comparer deux moyennes μ_0 et μ_1 lorsque les variance sont connues (avec changement de variables).
- la loi de Student t pour comparer deux moyennes lorsque la variance n'est pas connue et que l'on l'estime avec s.
- plusieurs cas divers en fonction du nombre de lots et variances des lots s_1 et s_2

Toutes nécessitent que les données proviennent de distributions connues, même en quelque sorte celle de Pearson car elle suppose de pouvoir attribuer des fréquences e_k attendues.

inférence non paramétrique

considérations générales

On aimerait avoir des tests qui suppose uniquement que les échantillons proviennent de population homogène, de distribution constante pour les échantillons mais dont la connaissance de la distribution n'est pas nécessaire.

Des estimations du centrage et l'étalement peuvent être, par exemple, obtenues par des statistiques d'ordre. Les deux statistiques d'ordre les plus élémentaire sont la valeur minimum de l'échantillon noté $\underline{X}$ et l'echantillon le plus grand $\overline{X}$

- centrage : la médiane
$$\frac{\underline{X} + \overline{X}}{2}$$
- étalement :
$$\overline{X} - \underline{X}$$

statististique de d'ordre

ou de rang

- Il est possible de classer les échantillons x_1, x_2, x_3 pour obtenir

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} < \dots$$

- Il est donc possible d'introduire les variables aléatoires (statistique d'ordre)

$$\begin{aligned} X_{(1)} &= \text{plus petite valeur des variables aléatoires } (X_1, X_2, \dots, X_n) \\ X_{(2)} &= \text{seconde plus petite valeur des variables aléatoires } (X_1, X_2, \dots, X_n) \\ \vdots &\quad \vdots \end{aligned}$$

statistique d'ordre

distribution exacte

On a les deux résultats dans le cas d'une distribution uniforme sur $[0;1]$ (on omet la démonstration)

- espérance mathématique :

$$\boxed{\mathbb{E}[X_{(r)}] = \frac{r}{n+1} \qquad r = 1, \ldots, n}$$

- variance :

$$\boxed{\text{var}[X_{(r)}] = \frac{r(n-r+1)}{(n+1)^2(n+2)} \qquad r = 1, \ldots, n}$$

- covariance :

$$\boxed{\text{cov}[X_{(r)}, X_{(s)}] = \frac{r(n-s+1)}{(n+1)^2(n+2)} \qquad 1 \leq r < s \leq n}$$

transformation intégrale de probabilité

théorème

Soit X une variable aléatoire de fonction de répartition F_X. Si F_X est continue, la variable aléatoire $Y = F_X(X)$ a la distribution continue uniforme sur l'intervalle $]0;1[$.

Comme $0 \leq F_X(x) \leq 1$ pour tout x, en laissant F_Y dénoter la fonction de répartition de Y, on a $F_Y(y) = 0$ pour $y \leq 0$ et $F_Y(y) = 1$ pour $y \geq 1$.

Pour $0 < y < 1$ on définit u comme le plus grand nombre qui satisfait $F_X(u) = y$. Ainsi, $F_X(x) \leq y$ si et seulement si $X \leq u$ et on a

$$F_Y(y) = P[F_X(X) \leq y] = P(X \leq u) = F_X(u) = y \qquad \text{CQFD}$$

distribution asymptotique

des statistiques d'ordre $X_{(r)}$

théorème

Lorsque $r/n \to p$ lorsque $n \to \infty$ avec $0 < p < 1$ alors en dénotant $\theta \triangleq F^{-1}(p)$

$$\boxed{\sqrt{\frac{n}{p(1-p)}} f_X(\theta)[X_{(r)} - \theta] \to \mathcal{N}(0,1)}$$

statistique de la somme des rangs de Wilcoxon

introduction

- Lors de comparaison de deux populations, si une s'écarte de l'autre les rangs petits (resp. grands) auront tendance à se regrouper selon les populations.
- En effectuant la somme des rangs et en comparant avec les possiblités générées par le hasard, il est possible d'établir une valeur maximale de cette somme pour garantir un seuil de signification de 5 %, par exemple.
- On va présenter la comparaison de deux populations A et B que l'on suppose de distribution identique et de centrage identique sous l'hypothèse H_0.

Exemple

introductif et théorique

Pourcentage d'efficacité d'un insecticide (Wilcoxon "Individual Comparsions by Ranking Methods", Biometrics Bulletin, Vol. 1, No. 6, pp. 80-83, 1945)

% A	Rang A	% B	Rang B
68		60	
68		67	
59		61	
72		62	
64		67	
67		63	
70		56	
74		58	

Table – Comparaison de deux séries d'échantillons A et B, huit répliques, une associée à la préparation A et l'autre à la préparation B. Résultat en pourcentage d'efficacité.

Exemple

introductif et théorique

- Les résultats sont classés en attribuant 1 au moins bon et en classant successivement par ordre croissant d'efficacité.
- En examinant les données, il y a des ex-aequo, trois fois 67 et deux fois 68.
- On attribue le classement médian aux séries de mesures identiques.
- Comme 67 arrive en position 9, 10 et 11, on attribue le classement médian 10 aux trois 67.
- Le 68 arrive en position 12 et 13 (avant traitement des ex-aequo), on attribue donc 12.5 au deux 68.
- Une fois le classement effectué et les rangs attribués à chaque mesure individuelle, la somme des rangs par échantillon est effectuées. Tous les rangs sont additionnés pour A et tous les rangs pour B.

Exemple

introductif et théorique

% A	Rang A	% B	Rang B
68	12.5	60	4
68	12.5	67	10
59	3	61	5
72	15	62	6
64	8	67	10
67	10	63	7
70	14	56	1
74	16	58	2
542	91	494	45

Table – Comparaison de deux séries d'échantillons A et B, huit répliques, une associée à la préparation A et l'autre à la préparation B. Résultat en pourcentage d'efficacité, rangs associés et totaux.

table de probabilités

détermination du NS 5 % pour la différence entre deux expériences

nb. mesures/éch.	plus petit total des rangs	probabilité de ce total
5	16	0.016
5	18	0.055
6	23	0.0087
6	24	0.015
6	26	0.041
7	33	0.0105
7	34	0.017
7	36	0.038
8	44	0.0104
8	46	0.021
8	49	0.05

table de probabilités

détermination du NS 5 % pour la différence entre deux expériences

nb. mesures/éch.	plus petit total des rangs	probabilité de ce total
9	57	0.0104
9	59	0.019
9	63	0.05
10	72	0.0115
10	74	0.0185
10	79	0.052

détermination du niveau de signification (NS 5 %)

exemple introductif

- 91 est la somme des rangs pour les 8 mesures de l'échantillon A
- 45 est la somme des rangs pour les 8 mesures de l'échantillon B
- La plus petite somme des rangs, 45, est examinée dans la table des probabilités, associée à 8 mesures/éch.
- La table indique que la probabilité pour un total aussi petit que 45 se situe entre 0.0104 et 0.021.
- La conclusion est que les deux traitements diffèrent de manière significative (car < 5 % ; en effet, si les rangs étaient tiré au sord, au hasard, le total de 45 apparaitrait que $5\times$ sur 100 en moyenne).

Exemple

introductif et théorique

nb. mesures/éch.	plus petit total des rangs	probabilité de ce total
8	44	0.0104
8	46	0.021
8	49	0.05

principe des partitions

détermination de la table des probabilités

- Il y $2 \times 8 = 16$ nombres, les rangs, à répartir aux 16 mesures.
- Le total de 45 observé est une somme particulière des rangs

$$45 = 1 + 2 + 4 + 5 + 6 + 7 + 10 + 10$$

- Il existe d'autres façons de répartir 8 rangs (distincts si on ne compte pas les ex-aequos compris de 1 à 16) pour arriver au même total, par exemple

$$\begin{aligned} 45 &= 1 + 2 + 4 + 5 + 6 + 7 + 9 + 11 \\ 45 &= 1 + 2 + 3 + 5 + 9 + 10 + 12 + 13 \\ &\vdots \quad \vdots \quad \vdots \end{aligned}$$

- Il y a donc un lien étroit entre les partitions d'un entier, disons r (pour le rang), en $q = 8$ entiers (distincts compris entre deux entiers)

illustration des partitions

Pour simplifier l'exposé considérons

- $q = 5$ mesures différentes par échantillon
- Formons la somme des rangs la plus petit possible ($q = 5$ additions des rangs les plus petit)

$$1 + 2 + 3 + 4 + 5 = 15$$

- Comme $q = 5$ et que l'on compare deux échantillons le rang maximum est $2q = 10$. On doit donc assigner à chaque mesure un chiffre compris entre 1 et 10.

illustration des partitions

$q = 5$ et $r = 20$

- Supposons que la somme des rangs soit 20. Effectuons les partitions de 20 avec des entiers distincts avec au maximum 10 pour les entiers.

1	2	3	4	10	20
1	2	3	5	9	20
1	2	3	6	8	20
1	2	4	5	8	20
1	2	4	6	7	20
1	3	4	5	7	20
2	3	4	5	6	20

- Il y a 7 partitions de 20 en somme de 5 entiers différents compris entre 1 et 10.

illustration des partitions

relation avec les partitions d'entiers avec répétition

partition d'un entier avec des entiers distincts $\equiv$ partition avec répétition

				5
			1	4
			2	3
		1	1	3
		1	2	2
	1	1	1	2
1	1	1	1	1

Il y a 7 partitions de 5 avec répétition.

probabilités exactes

et construction des tables

- La suite de rangs la plus rare de 5 rangs est 1, 2, 3, 4, et 5, ce qui donne comme total 15.
- La somme des rangs commence donc avec 15 et augmente jusqu'à $10+9+8+7+6=40$.
- Si un total de 20 est observé, il s'agit de compter les partitions de $20-15=5$ avec répétition.

Dans la formule suivante, r est la somme des rangs observés, q est le nombre de mesures dans l'échantillon.

$$P = 2\left\{1+\sum_{i=1}^{r}\sum_{j=1}^{q}\sqcap_j^i - \sum_{n=1}^{r-q}\left[(r-q-n+1)\sqcap_{q-1}^{q-2+n}\right]\right\} / \frac{(2q)!}{q!\times q!}$$

avec $\sqcap_i^j$ désignant la partition de i en j parties (avec répétition).

probabilités exactes

illustration du décompte des partitions

- Remarque $(2q)!/(q! \times q!) = C_q^{2q}$, on a $C_5^{10} = 252$

r	différence	partitions	$2\times$ partitions$/C_q^{2q}$	cumulé $= P$
15	0	1	$\frac{2}{252}$	$\frac{2}{252} = 0.007936$
$\boxed{16}$	1	1	$\frac{2}{252}$	$\frac{4}{252} = \boxed{0.01587}$
17	2	2	$\frac{2\times 2}{252}$	$\frac{8}{252} = 0.031745$
$\boxed{18}$	3	3	$\frac{3\times 2}{252}$	$\frac{14}{252} = \boxed{0.0555}$

table moderne de Wilcoxon

Gibbons, Chakraborti

- Jean Dickson Gibbons, Subhabrata Chakraborti, Nonparametric Statistical Inference, CRC Press, 2021, pp. 617 et 618

	$m = 5$				$m = 5$		
$n = 5$	queue g.	P	queue dr.	$n = 5$	queue g.	P	queue dr.
	15	0.004	40		22	0.155	33
	16	0.008	39		23	0.210	32
	17	0.016	38		24	0.274	31
	18	0.028	37		25	0.345	30
	19	0.048	36		26	0.421	29
	20	0.075	35		27	0.500	28

	$m = 6$		
$n = 6$	queue gauche	P	queue droite
	21	0.001	57
	22	0.002	56
	23	0.004	55
	24	0.008	54
	25	0.013	53
	26	0.021	52
	27	0.031	51

correspondances entre les tables

- La probabilité de la table de l'article de Wilcoxon est $2x$ celle de la probabilité de la table de Gibbons, Chakraborti.
- Cela rend possible de tester des hypothèses composites telle que $\theta < 0$ au lieu de $\theta \neq 0$.
- Dans la table de Gibbons, Chakraborti, possibilité d'avoir $m \neq n$.

application

$H_0 : F_Y(x) = F_X(x), \forall x,\ H_1 : F_Y(x) + F_X(x-\theta)\ \forall x,\ \exists\theta$

alternative	région de rejet	P
$\theta < 0 \quad (Y <^{ST} X)$	$W_N \geq w_a$	$P(W_N \geq w_0)$
$\theta > 0 \quad (Y >^{ST} X)$	$W_N \leq w'_a$	$P(W_N \leq w_0)$
$\theta \neq 0$	$W_N \geq w_{a/2}$ ou $W_N \leq w'_{a/2}$	$2\times$ la plus petite des deux

Table – Test d'hypothèse composite et simple. Le dernier cas donne l'hypothèse H_0 simple pour laquelle la probabilité est doublée, ce qui correspondait à la première table (celle de l'article de Wilcoxon) et qui donne également le test de signification.

valeurs asymptotiques

alternative	région de rejet	P
$\theta < 0$	$W_N \geq \frac{m(N+1)}{2} + 0.5 + z_\alpha \sqrt{\frac{mn(N+1)}{12}}$	$1 - \Phi\left(\frac{w_0 - 0.5 - m(N+1)/2}{\sqrt{mn(N+1)/12}}\right)$
$\theta > 0$	$W_N \leq \frac{m(N+1)}{2} - 0.5 - z_\alpha \sqrt{\frac{mn(N+1)}{12}}$	$\Phi\left(\frac{w_0 + 0.5 - m(N+1)/2}{\sqrt{mn(N+1)/12}}\right)$
$\theta \neq 0$	les deux du dessus avec z_α remplacé par $z_{\alpha/2}$	$2\times$ le plus petit des deux

Table – distribution asymptotique des résultats de la statistique de Wilcoxon W_N, seuil w_0 et signification P.

exemple

- Le tableau suivant indique le temps de fabrication d'une pièce, opération qui a lieu soit le matin, soit l'après-midi. On suspecte une différence entre le matin et l'après-midi.

matin		après-midi	
12.6	11.2	16.4	15.4
11.4	9.4	14.1	14.0
13.2	12.0	13.4	11.3

- $m = n = 6$
- X : matin, Y : après-midi
- $H_1 : \theta = \mu_Y - \mu_X > 0$ (alternative)
- P est la queue de gauche pour W_N

- Il faut arranger par colonne et trier chaque colonne par ordre croissant (c'est un peu plus systématique comme cela).

matin	après-midi
9.4	11.3
11.2	13.4
11.4	14.0
12.0	14.1
12.6	15.4
13.2	16.4

- On effectue un tri par insertion :

$$\underline{9.4} < \underline{11.2} < 11.3 < \underline{11.4} < \underline{12.0} < \underline{12.6} < \underline{13.2} < 13.4 < 14.0 < 14.1 < \dots$$

- On assigne les rangs aux éléments sous-lignés

$$1 < 2 < 4 < 5 < 6 < 7$$

- On calcule le total

$$r = W_N = 1 + 2 + 4 + 5 + 6 + 7 = 25$$

- On lit P dans la table $m = 6$, $n = 6$

$$P(W_N \leq 25) = 0.013$$

- H_0 est rejetée en faveur de l'alternative $H_1 : \theta > 0$ à tout niveau de signification (NS) $\alpha \geq 0.013$

	$m = 6$		
$n = 6$	queue gauche	P	queue droite
	21	0.001	57
	22	0.002	56
	23	0.004	55
	24	0.008	54
	25	0.013	53
	26	0.021	52
	27	0.031	51