REMPLISSAGE DES POTS DE YAOURT

vanne étiquette conditionnement

rentabilité de
l'entreprise

législation → supermarché

balance

Législation Rentabilité Limites techniques

Échantillons + réglage de vanne

EXEMPLE

X = variable aléatoire : poids au temps t

Y = variable aléatoire : poids au temps $t + 5\,\mathrm{h}$

Deux 10-échantillons :

- au temps t :

  125,1 124,8 126,1 125,6 125,8
  124,2 125,3 125,0 125,1 124,2 $\rightarrow$ m-échantillon

- au temps $t + 5\,\mathrm{h}$:

  128,2 127,9 129,4 128,8 128,2
  128,0 128,4 128,1 129,1 127,9 $\rightarrow$ n-échantillon

Faut-il modifier le temps d'ouverture de la vanne ?

HYPOTHÈSES :

$H_0 : \mu_X = \mu_Y$ (pas besoin de recalage) contre
$H_1 : \mu_X \neq \mu_Y$ (action nécessaire)

POSTULATS sur les distributions théoriques de X et Y :

- lois normales

$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ $\qquad$ $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$

- $\sigma_X = \sigma_Y = \sigma$ (estimés par $\widehat{\sigma}_X$ et $\widehat{\sigma}_Y$)
- variables X_i et Y_j indépendantes

DISTRIBUTIONS
DE X ET DE Y

$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ $\qquad\qquad$ $\downarrow$ $\qquad\qquad$ $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$

DISTRIBUTIONS
DE $\overline{X}$ ET DE $\overline{Y}$

$$\overline{X} \sim \mathcal{N}(\mu_X, \underbrace{\frac{\sigma_X^2}{m}}_{\sigma_{\overline{X}}^2}) \qquad\qquad \overline{Y} \sim \mathcal{N}(\mu_Y, \underbrace{\frac{\sigma_Y^2}{n}}_{\sigma_{\overline{Y}}^2})$$

$$\overline{X} = \frac{1}{m} \sum_{i=1}^{m} X_i \qquad\qquad \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$$\widehat{\sigma}_{\overline{X}} = \frac{\widehat{\sigma}_X}{\sqrt{m}} \qquad\qquad \widehat{\sigma}_{\overline{Y}} = \frac{\widehat{\sigma}_Y}{\sqrt{n}}$$

DISTRIBUTION DE $\overline{X}$ ET DE $\overline{Y}$

$$\overline{X} \sim \mathcal{N}(\mu_X, \tfrac{\sigma_X^2}{m}) \qquad \overline{Y} \sim \mathcal{N}(\mu_Y, \tfrac{\sigma_Y^2}{n})$$

$$(\overline{X} - \overline{Y}) \sim \mathcal{N}([\mu_X - \mu_Y], \sigma^2[\tfrac{1}{m} + \tfrac{1}{n}])$$

$$\widehat{\sigma}^2 = \frac{\Sigma(X_i - \overline{X})^2 + \Sigma(Y_i - \overline{Y})^2}{(m + n - 2)}$$

Statistique :

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\widehat{\sigma}\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

Sous H_0, $\mu_X = \mu_Y \Longrightarrow T_0 \sim t_{m+n-2}$,

loi de Student à $(m+n-2)$ degrés de liberté.

Loi connue, tabulée.

Sous H_1, T_0 suit une loi inconnue.

Rejet de H_0 si $|T_0| > t_{m+n-2}$ $(5\,\%)$.

Lecture de la table.

Avec les données des yaourts on a donc :

ddl = 18

$\alpha = 5\,\%$ réparti sur les 2
queues de la distribution

$t_{\alpha/2} = 2{,}101$

rejet si $|T_0|$ trop grand
(test bilatéral)

TABLE 4.1 : **Distribution du t de Student**

Bilatéral							
γ	0,5	0,8	0,9	0,95	0,98	0,99	
$\alpha/2$	0,25	0,1	0,05	0,025	0,01	0,005	
$\alpha = 1-\gamma$	0,5	0,2	0,1	**0,05**	0,02	0,01	Unilatéral
Degrés de	0,75	0,90	0,95	0,975	0,99	0,995	$\gamma' = 1-\alpha'$
liberté :	0,25	0,10	0,05	0,025	0,01	0,005	$\alpha' = \alpha/2$
1	1,000	3,078	6,314	12,706	31,821	63,657	
2	0,816	1,886	2,920	4,303	6,965	9,925	
3	0,765	1,638	2,353	3,182	4,541	5,841	
4	0,741	1,533	2 132	2,776	3,747	4,604	
5	0,727	1,476	2,015	2,571	3,365	4,032	
6	0,718	1,440	1,943	2,447	3,143	3,707	
7	0,711	1,415	1,895	2,365	2,998	3,499	
8	0,706	1,397	1,860	2,306	2,896	3,355	
9	0,703	1,383	1,833	2,262	2,821	3,250	
10	0,700	1,372	1,812	2,228	2,764	3,169	
11	0,697	1,363	1,796	2,201	2,718	3,106	
12	0,695	1,356	1,782	2,179	2,681	3,055	
13	0,694	1,350	1,771	2,160	2,650	3,012	
14	0,692	1,345	1,761	2,145	2,624	2,977	
15	0,691	1,341	1,753	2,131	2,602	2,947	
16	0,690	1,337	1,746	2,120	2,583	2,921	
17	0,689	1,333	1,740	2,110	2,567	2,898	
18	0,688	1,330	1,734	**2,101**	2,552	2,878	
19	0,688	1,328	1,729	2,093	2,539	2,861	
20	0,687	1,325	1,725	2,086	2,528	2,845	
21	0,686	1,323	1,721	2,080	2,518	2,831	
22	0,686	1,321	1,717	2,074	2,508	2,819	
23	0,685	1,319	1,714	2,069	2,500	2,807	
24	0,685	1,318	1,711	2,064	2,492	2,797	
25	0,684	1,316	1,708	2,060	2,485	2,787	
26	0,684	1,315	1,706	2,056	2,479	2,779	
27	0,684	1,314	1,703	2,052	2,473	2,771	
28	0,683	1,313	1,701	2,048	2,467	2,763	
29	0,683	1,311	1,699	2,045	2,462	2,756	
30	0,683	1,310	1,697	2,042	2,457	2,750	
40	0,681	1,303	1,694	2,021	2,423	2,704	
60	0,679	1,296	1,671	2,000	2,390	2,660	
120	0,677	1,289	1,658	1,980	2,358	2,617	
∞	0,674	1,282	1,645	1,960	2,326	2,576	

Coefficient de confiance γ : surface entre $-t_{\alpha/2}$ et $t_{\alpha/2}$.

PROGRAMME *SAS*

(correspondant au test t de Student)

```
exercice poids de yaourts';
data yaourts;
input poids heure;
cards;
125.1 0
124.8 0
126.1 0
125.6 0
125.8 0
124.2 0
125.3 0
125.0 0
125.1 0
124.2 0
128.2 5
127.9 5
129.4 5
128.8 5
128.2 5
128.0 5
128.4 5
128.1 5
129.1 5
127.9 5
RUN;
```

```
proc sort data=yaourts;
by heure;
run;

proc ttest data=yaourts;
class heure;
var poids;
run;
```

SORTIES DU PROGRAMME *SAS*

Standard Deviation : $\widehat{\sigma}_X = \sqrt{\Sigma(X_i - \overline{X})^2/(N-1)}$

Standard Error : $\widehat{\sigma}_{\overline{X}} = \widehat{\sigma}_X/\sqrt{N}$

exercice poids de yaourts

TTEST PROCEDURE

Variable : POIDS

HEURE	N	Mean	Std Dev	Std Error	Minimum	Maximum
0	10	125.1200000	0.62325311	0.19708994	124.2000000	126.1000000
5	10	128.4000000	0.52493386	0.16599866	127.9000000	129.4000000

Variances	T	DF	Prob>|T|
unequal	-12.7289	17.5	0.0001
equal	-12.7289	18.0	0.0000

for H0: Variances are equal, F' = 1.41 DF = (9,9) Prob>F' = 0.6172

T_0

test bilatéral : $P =$ Prob>|T| = P-variable

correspond à nos hypothèses : $\sigma_X = \sigma_Y = \sigma$.

On calcule donc $|T_O| = +12,7289 > t_{\alpha/2} = 2{,}101$ ($\alpha = 5\,\%$).

Conclusion : $\mu_x \neq \mu_y$, il faut donc recaler la machine.

LES LIMITES DU TEST

• Importance relative de postulats faux :

- indépendance : indispensable,

- normalité : pas trop d'importance (si grands échantillons),

- inégalité des variances : ne pose pas trop problème
(si échantillons de même taille).

• Comment détecter des postulats faux :

	Graphique	Test
Normalité	*QQ-plot*/normal (droite de Henry)	divers : Lilliefors, χ^2 d'ajustement, etc.
Indépendance	Y_i *versus* X_i	runs, Spearman, etc.
Variances	*boxplot*	Fisher (pas terrible)

EXERCICE

2 régimes alimentaires sont essayés sur 2 lots de jeunes bovins.

Les gains de poids sont :

1^{er} lot :	+5	+2	–0.7	+9	+8	+4
2^{e} lot :	+12	+15	+13	+9	+11	+14

Y a-t-il une différence entre les 2 régimes ?

BIBLIOTHÈQUE		
Romans	Livres de statistiques	Autres
Population X	Population Y	

↓ 16 romans

↓ 12 livres de statistiques

Nombre de pages :
29, 39, 60, 78, 82, 112, 125, 170, 192, 224, 263, 275, 276, 286, 369, 756.

126, 142, 156, 228, 245, 246, 370, 419, 433, 454, 478, 503.

Échantillon $(X_1, \ldots, X_i, \ldots, X_{16})$

Échantillon $(Y_1, \ldots, Y_j, \ldots, Y_{12})$

Question :

Le nombre médian de pages diffère-t-il entre X et Y ?

TEST DE MANN-WHITNEY-WILCOXON

Tester la différence entre les médianes
m_X et m_Y de deux populations X et Y,

à partir de deux échantillons :
$(X_1, \ldots, X_m)$ et $(Y_1, \ldots, Y_n)$.

Hypothèses :

$H_0 : m_X = m_Y$

$H_1 : m_X \neq m_Y \rightarrow$ test bilatéral

$H_1' : m_X > m_Y \searrow$
$H_1'' : m_X < m_Y \nearrow$ tests unilatéraux

POSTULATS

– les deux échantillons sont indépendants (important),

– les X_i ont tous une même distribution (important),

– les Y_j ont tous une même distribution (important),

– ces distributions sont continues (problème des ex æquo).

Mais :

– pas de postulat Gaussien,

– pas de postulat de symétrie des distributions.

STATISTIQUE

U_X = nombre de couples (X_i, Y_j) où $X_i > Y_j$.

U_Y = nombre de couples (X_i, Y_j) où $X_i < Y_j$.

U = la plus petite des 2 valeurs U_X et U_Y.

- Si H_0 vraie : la distribution de U est symétrique
  (autour de sa médiane $mn/2$).

- Si H_1 vraie : la distribution de U est asymétrique
  (médiane de U plus proche de 0 ou mn).

Si U est trop loin de $mn/2$, rejet de H_0.

EXEMPLE THÉORIQUE : $m = 3, \quad n = 2.$

U_X = nombre de couples où $X_i > Y_j$:

1	YYXXX	$\rightarrow$	6
2	YXYXX	$\rightarrow$	5
3	YXXYX	$\rightarrow$	4
4	XYYXX	$\rightarrow$	4
5	YXXXY	$\rightarrow$	3
6	XYXYX	$\rightarrow$	3
7	XYXXY	$\rightarrow$	2
8	XXYYX	$\rightarrow$	2
9	XXYXY	$\rightarrow$	1
10	XXXYY	$\rightarrow$	0

$\Rightarrow$

u	$\Pr\{U_X = u\}$
0	1/10
1	1/10
2	2/10
3	2/10
4	2/10
5	1/10
6	1/10

Loi de U_X sous H_0 : symétrique :

Région de rejet = ensemble des valeurs de U qui sont :
peu probables si H_0 vraie,
plus probables si H_1 vraie.

Exemple ici : $\{0, 6\}$, ou $\{0, 1, 5, 6\}$.

CALCUL PRATIQUE DE U

• Ordonner toutes les valeurs de X_i et Y_j ensemble.

• Attribuer un rang (par ordre croissant) à chacune.

S_X = somme des rangs des X_i.
S_Y = somme des rangs des Y_j.

On a :

$U_X = S_X - m(m+1)/2$
$U_Y = S_Y - n(n+1)/2$

• Remarque : s'il y a des observations de même valeur, donner un rang moyen à chacune.

EXEMPLE THÉORIQUE : $m = 3, \quad n = 2.$

S_X = somme des rangs des X_i

			S_X		$U_X = S_X - m(m+1)/2$
1	YYXXX	$\rightarrow$	12	$\rightarrow$	6
2	YXYXX	$\rightarrow$	11	$\rightarrow$	5
3	YXXYX	$\rightarrow$	10	$\rightarrow$	4
4	XYYXX	$\rightarrow$	10	$\rightarrow$	4
5	YXXXY	$\rightarrow$	9	$\rightarrow$	3
6	XYXYX	$\rightarrow$	9	$\rightarrow$	3
7	XYXXY	$\rightarrow$	8	$\rightarrow$	2
8	XXYYX	$\rightarrow$	8	$\rightarrow$	2
9	XXYXY	$\rightarrow$	7	$\rightarrow$	1
10	XXXYY	$\rightarrow$	6	$\rightarrow$	0

$$U_X + U_Y = mn$$

Cas des grands échantillons (m et $n > 20$)

On utilise $Z = (U - mn/2)/\sqrt{[mn(m+n+1)/12]}$, qui est asymptotiquement distribuée comme une variable gaussienne $\mathcal{N}(0,1)$.

Efficacité du test

L'efficacité asymptotique relative de ce test par rapport au t de Student est de 95,5 %.

$$\Rightarrow \text{Peu de perte de puissance.}$$

(puissance = probabilité de choisir H_1 à bon escient).

DISTRIBUTION DE LA STATISTIQUE U

(test de Mann-Whitney-Wilcoxon)

Test bilatéral (valeurs critiques)

n	5	6	7	8	9	10	11	12	13	14	15	**16**	17	18	19	20
m																
5		3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	1		6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	1	3		10	12	14	16	18	20	22	24	26	28	30	32	34
8	2	4	6		15	17	19	22	24	26	29	31	34	36	38	41
9	3	5	7	9		20	23	26	28	31	34	37	39	42	45	48
10	4	6	9	11	13		26	29	33	36	39	42	45	48	52	55
11	5	7	10	13	16	18		33	37	40	44	47	51	55	58	62
12	6	9	12	15	18	21	24		41	45	49	**53**	57	61	65	69
13	7	10	13	17	20	24	27	31		50	54	59	63	67	72	76
14	7	11	15	18	22	26	30	34	38		59	64	69	74	78	83
15	8	12	16	20	24	29	33	37	42	46		70	75	80	85	90
16	9	13	18	22	27	31	36	41	45	50	55		81	86	92	98
17	10	15	19	24	29	34	39	44	49	54	60	65		93	99	105
18	11	16	21	26	31	37	42	47	53	58	64	70	75		106	112
19	12	17	22	28	33	39	45	51	57	63	69	74	81	87		119
20	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	

Les valeurs situées à droite et au-dessus de la diagonale sont les valeurs critiques au risque $\alpha = 5\,\%$.

Les valeurs situées à gauche et en dessous de la diagonale sont les valeurs critiques au risque $\alpha = 1\,\%$.

EXERCICE D'APPLICATION

Deux échantillons aléatoires sont tirés de façon indépendante dans deux populations inconnues d'où sont prélevées respectivement $m = 8$ et $n = 7$ unités indépendantes.

Echantillon (X)	14	25	30	32
$(m = 8)$	40	41	43	45

Echantillon (Y)	12	16	19	20
$(n = 7)$	24	27	35	

Faire un test bilatéral de Mann-Whitney-Wilcoxon d'identité des moyennes en adoptant le risque de 1re espèce $\alpha = 0{,}05$.

EXERCICE D'APPLICATION AVEC APPROXIMATION GAUSSIENNE

(Exercice n° 5.5, Sprent, 1989, p. 108)

Énoncé :

Un psychologue veut savoir s'il existe une différence entre hommes et femmes quant au niveau d'anxiété enregistrée avant une opération chirurgicale.

Un indice d'anxiété est mesuré sur 17 hommes et 23 femmes une semaine avant leur admission à l'hopital.

Ces personnes sont rangées de 1 à 40 sur une échelle d'anxiété croissante.

La somme des rangs pour les 17 hommes est de 428.

Problème : est-ce que l'anxiété dépend du sexe ?
(faire un test au risque 5 %).
Si oui, quel sexe présente la plus grande anxiété ?

TABLE 4.3 : **Distribution normale centrée réduite**

Soit la variable $X \sim \mathcal{N}(0,1)$, à valeurs réelles, de densité ϕ.
Si u est un nombre positif, et si on pose $\alpha = \Pr\{|X| \geq u) = 2[1 - \Phi(u)]$, la table donne u pour différentes valeurs de α.
Exemple : $\alpha = \Pr\{|X| \geq 1,200) \approx 0,23$.
Rejet de H_0 au niveau 5 % si $|X| \geq 1,96$.
$\Pr\{|X| \geq 1,96) \approx 0,05$.

α	+0,00	+0,01	+0,02	**+0,03**	+0,04	**+0,05**	+0,06	+0,07	+0,08	+0,09
0,0	$-\infty$	2,576	2,326	2,170	2,054	**1,960**	1,881	1,812	1,751	1,695
0,1	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311
0,2	1,282	1,254	1,227	**1,200**	1,175	1,150	1,125	1,103	1,080	1,058
0,3	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,860
0,4	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,690
0,5	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,6	0,524	0,510	0,496	0,482	0,468	0,454	0,440	0,426	0,412	0,399
0,7	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,8	0,253	0,240	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,9	0,126	0,113	0,100	0,088	0,075	0,063	0,050	0,038	0,025	0,013

Quelques valeurs particulières :

u	0,574	1,282	1,645	1,960	2,325	2,576	3,090
α	0,50	0,20	0,10	0,05	0,02	0,01	0,002
u	3,291	3,891	4,417	4,992	5,327	5,731	6,109
α	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}

TEST DE KOLMOGOROV-SMIRNOV

Problème :

- 1 caractère (variable, attribut. . .),
- 2 échantillons, 2 populations,
- mesures indépendantes.

On veut comparer les lois de probabilité
dans les 2 populations.

$$\begin{array}{ll} (x_1, \ldots, x_m) & x_i \leq x_{i+1} \\ (y_1, \ldots, y_n) & y_j \leq y_{j+1} \end{array}$$

HYPOTHÈSES NULLE H_0 ET ALTERNATIVE H_1

Notons $F_X(z)$ et $F_Y(z)$ les fonctions de répartition dans chacune des populations.

$$H_0 : F_X(z) = F_Y(z) \ \forall z$$

Alternatives possibles :

$H_1\ (a) : F_X(z) \neq F_Y(z)$ $\qquad$ ($\exists z$ tel que...)

$H_1\ (b) : F_X(z) \geq F_Y(z)$ $\qquad$ ($\forall z$, et $\exists z$ tel que $>$)

$H_1\ (c) : F_X(z) \leq F_Y(z)$ $\qquad$ ($\forall z$, et $\exists z$ tel que $<$)

$H_1\ (d) : F_X(z) = F_Y(z - \theta),\ \theta \neq 0$ $\qquad$ (ou $\theta > 0$, ou $\theta < 0$)

$H_1\ (e) : F_X(z) = F_Y(\theta z),\ \theta \neq 1$ $\qquad$ (ou $\theta > 1$, ou $\theta < 1$)

STATISTIQUE DE KOLMOGOROV

On définit les fonctions de répartition empiriques :

$$\widehat{F}_{m,X}(z) = \frac{\text{nombre de } x_i \leq z}{m}$$
$$\widehat{F}_{n,Y}(z) = \frac{\text{nombre de } y_j \leq z}{n}$$

Les distributions des critères :

$$D_{m,n} = \sup_z |\widehat{F}_{m,X}(z) - \widehat{F}_{n,Y}(z)|$$
$$D^+_{m,n} = \sup_z [\widehat{F}_{m,X}(z) - \widehat{F}_{n,Y}(z)]$$

sont tabulées pour les petites valeurs de m et n.

Au delà, on connaît la distribution asymptotique.

TESTS

Bilatéral :

$H_0 : F_X = F_Y$,
$H_1 : F_X(z) \neq F_Y(z)$ pour au moins une valeur de z.

On utilise $D_{m,n}$: si $D_{m,n} > d_\alpha \implies$ rejet de H_0.

Unilatéral :

$H_0 : F_X = F_Y$,
$H_1 : F_X > F_Y$.

Si $D^+_{m,n} > d^+_\alpha \implies$ rejet de H_0.

Remarques :
– les tests sont moins puissants qu'un test spécifique
portant seulement sur un décalage de position ;
– on ne connaît pas la puissance exacte.

EXEMPLE

x :		39,1	40,6	41,2	45,2	46,2
$F_{m,X}(x)$:	$\frac{0}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$	$\frac{5}{10}$

x :	47,2	48,4	48,7	52,1	55,0
$F_{m,X}(x)$:	$\frac{6}{10}$	$\frac{7}{10}$	$\frac{8}{10}$	$\frac{9}{10}$	$\frac{10}{10}$

y :		24,3	29,1	32,4	32,6	34,4
$F_{n,Y}(y)$:	$\frac{0}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$	$\frac{5}{10}$

y :	35,2	38,1	39,2	40,9	41,8
$F_{n,Y}(y)$:	$\frac{6}{10}$	$\frac{7}{10}$	$\frac{8}{10}$	$\frac{9}{10}$	$\frac{10}{10}$

On a $\forall z, F_{m,X}(z) \leq F_{n,Y}(z)$. Si l'on choisit $H_1 : F_Y > F_X$, on cherche $D^+_{m,n} = \sup_z[\widehat{F}_{n,Y}(z) - \widehat{F}_{m,X}(z)]$, qui est atteint par exemple pour z compris entre 38,1 et 39,1, et vaut $\frac{7}{10}$.

Pour $\alpha = 0,01$, la table donne $d^+_\alpha = 0,700 \Rightarrow$ rejet de H_0.

Si l'on utilise l'approximation asymptotique, la statistique $4(D^+_{m,n})^2 \frac{mn}{m+n}$ vaut $4 \times (\frac{7}{10})^2 \times \frac{10\times 10}{20} = 9,8 > 9,21 = \chi^2_2(0,01)$. On rejette donc encore H_0.