Méthodes non paramétriques

- F. Farnir, A. Rives, L. Massart, N. Moula
- Faculté de Médecine Vétérinaire
- Université de Liège

Pourquoi ?

- Les tests vus jusqu'à présent font des hypothèses sur les distributions des résidus
  - Indépendance
  - Normalité des données
  - Homoscédasticité
- Que fait-on quand ces hypothèses ne sont pas vérifiées ?

Objectif

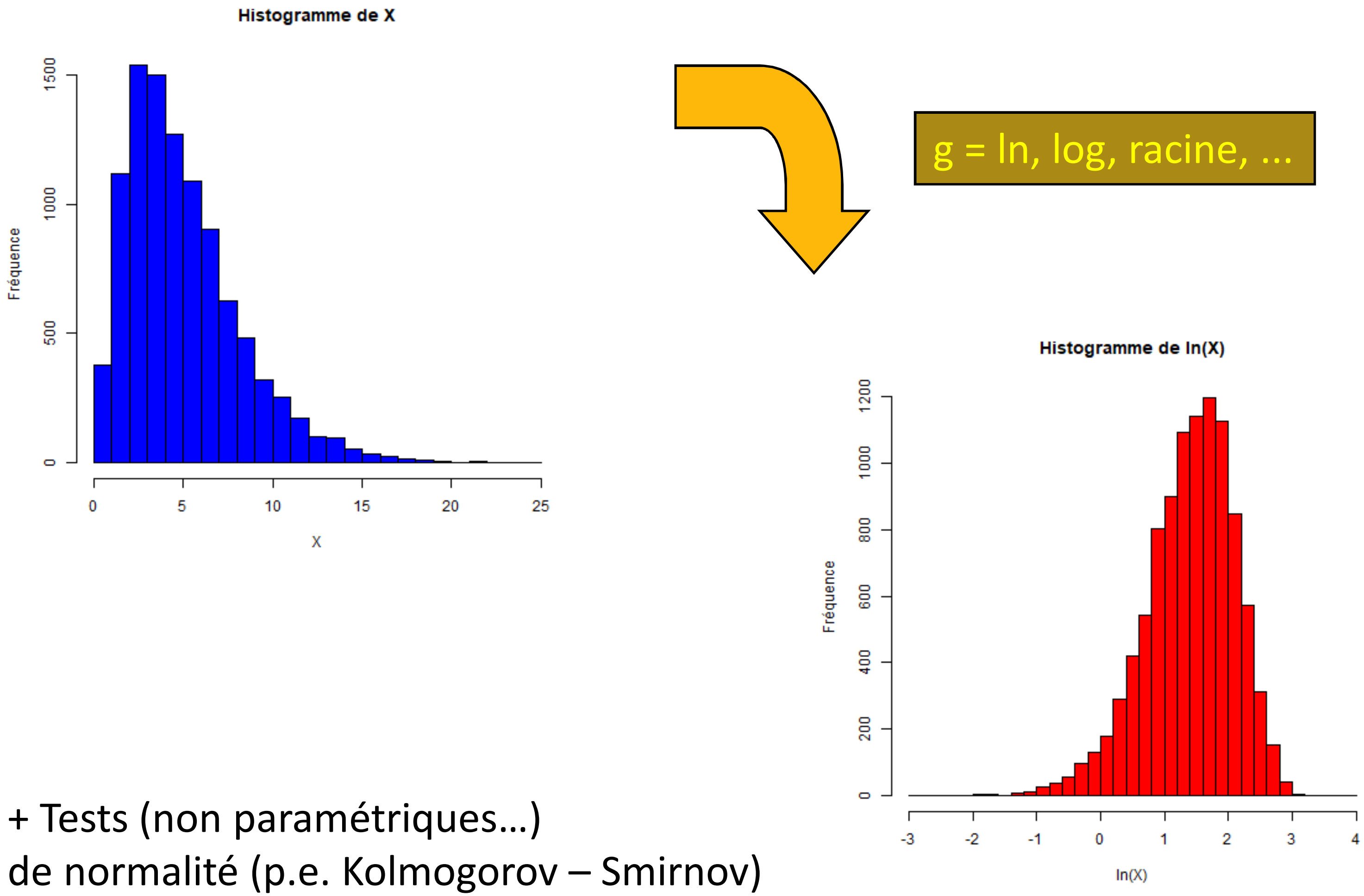
- Remplacer les tests basés sur ces hypothèses ("tests paramétriques") par d'autres tests qui sont "robustes" quand ces hypothèses ne sont pas vérifiées.
  => Tests "non-paramétriques"

Comment ?

- L'idée est de remplacer la (les) variable(s) par une (des) autre(s), respectant l'ordre de la variable initiale, et de distribution connue

- Trois techniques:
  - Normalisation des données
  - Permutations
  - Utilisation des rangs

A - Normalisation des données



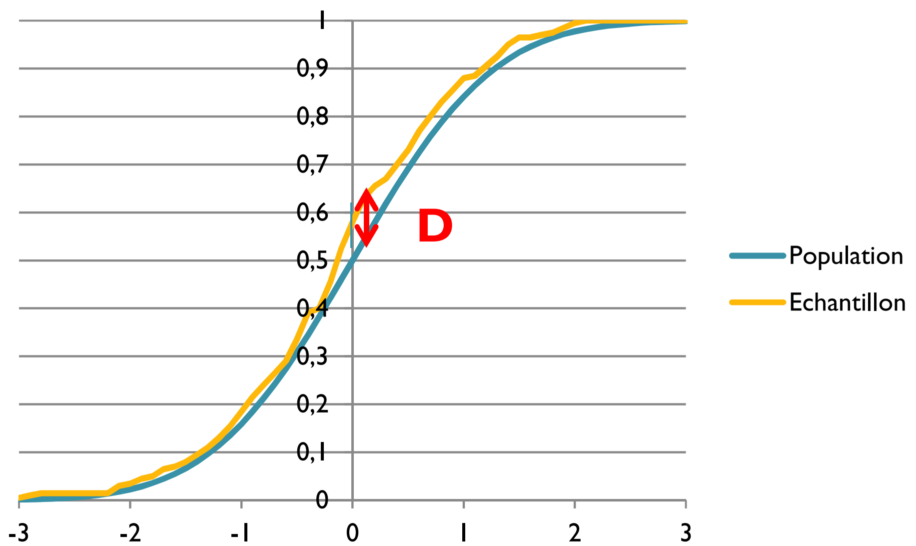
+ Tests (non paramétriques…)
de normalité (p.e. Kolmogorov – Smirnov)

Test de Kolmogorov - Smirnov

- Test non-paramétrique permettant de tester si une variable suit une distribution théorique particulière (p.e. normale)
- Principe:
  - On représente graphiquement la densité théorique de probabilité cumulée
  - On représente graphiquement la densité empirique de probabilité cumulée

Test de Kolmogorov - Smirnov

- Principe (suite):
  - On calcule l'écart maximal D entre les deux distributions



Test de Kolmogorov - Smirnov

- Principe (suite):
  - Les valeurs maximales de D sont tabulées en fonction de l'effectif

Test de Kolmogorov-Smirnov

$$\mathbb{P}\{\|F - F_n^\star\|_\infty \leq D_n\} = p$$

(F est une fonction de répartition continue et $F^\star$ la fonction de répartition observée.)

p=	0,80	0,9	0,95	0,975	0,98	0,99	0,995	=p
n								n
1	0,900000	0,950000	0,975000	0,987500	0,990000	0,995000	0,997500	1
2	0,683772	0,776393	0,841886	0,888197	0,900000	0,929289	0,950000	2
3	0,564810	0,636045	0,707598	0,767921	0,784557	0,829002	0,864279	3
4	0,492653	0,565216	0,623939	0,673881	0,68887	0,734238	0,776393	4
5	0,44698	0,509449	0,563275	0,612618	0,62718	0,668531	0,705431	5
6	0,410373	0,467993	0,519262	0,564043	0,577407	0,616607	0,652865	6
7	0,381476	0,436069	0,483424	0,525645	0,53844	0,575812	0,609753	7
8	0,358313	0,409623	0,454267	0,494464	0,506543	0,541793	0,574291	8

Test de Kolmogorov - Smirnov

- Exemple:

```
> x <- rnorm(50) # loi normale
> y <- rnorm(40) # loi normale
> z <- runif(30) # loi uniforme
> ks.test(x,y)
        Two-sample Kolmogorov-Smirnov test
data:  x and y
D = 0.145, p-value = 0.686
alternative hypothesis: two-sided
> ks.test(x,z)
        Two-sample Kolmogorov-Smirnov test
data:  x and z
D = 0.5, p-value = 9.065e-05
alternative hypothesis: two-sided
>
```

B - Permutations

- On a vu le principe plus haut dans le cours
- On va l'illustrer sur un exemple de comparaison de deux groupes non pairés, mais on peut facilement généraliser à d'autres situations

B - Permutations

- Exemple: Comparaison des anomalies de spermatozoïdes dans deux milieux. Chaque donnée représente des comptages dans un lot.

Milieu I	Milieu II
1	10
2	150
120	155
	220

Moyenne I	Moyenne II
41	133.75

Différence	92.75

B – Permutations - principe

- La question: s’il n’y a pas de différence entre milieux, quelle est la probabilité d’observer une différence aussi grande (voire plus grande) ?
- Principe: sous H_0, il n’y a pas de différence entre les deux milieux. Les données sont donc attribuées aléatoirement à chaque milieu si H_0 est vraie.

B – Permutations – principe (2)

- Principe (suite): on peut donc vérifier, en répartissant les (7) données aléatoirement entre les deux groupes, dans quelle proportion de cas on dépasse la différence réellement observée (92.75)
- Exemple: dans le cas présent, il y a 35 manières différentes de répartir les 7 données en un groupe de 3 et un groupe de 4 => on peut tester toutes les possibilités

B – Permutations – principe (3)

Exemple:

Milieu I			Milieu II				M2 - M1	
1	2	10	120	150	155	220	156.92	>
…			…				…	
150	2	120	10	1	155	220	5.83	<
…			…				…	

B – Permutations – principe (3)

- Résultat:
  - Parmi les 35 permutations, seules 2 donnent des valeurs supérieures ou égales à 92.75

=> $P = 2/35 = 0.0571$

B – Permutations – remarque

- Remarque:
  - Si la taille des groupes ↑, le nombre de possibilités ↑ ↑
    - p.e. pour 2 groupes de 10, le nombre de possibilités est: $C_{20}^{10} = 184756$

      => recours à l'échantillonnage (méthode de Monte-Carlo)

C - Utilisation des rangs

- On substitue aux valeurs réelles leur rang dans la liste des valeurs
  - Exemple:

Valeurs	13	25	1	170	289
Rangs	2	3	1	4	5

C - Utilisation des rangs

- Problème: valeurs identiques
- Solution: utiliser le rang moyen
  - Exemple:

Valeurs	13	25	170	25	289
Rangs	1	2.5	4	2.5	5

(2 + 3) / 2 = 2.5

Distribution des rangs

- On peut trouver des formules générales pour les paramètres de la distribution des rangs
  - Exemple: somme des rangs

Rangs ↑	1	2	3	4	5
Rangs ↓	5	4	3	2	1
2*Som	6=n+1	6	6	6	6

Somme	n(n+1)/2

Distribution des rangs

- Exemple: moyenne des rangs = $\Sigma r / n = \mu_r$

Moyenne	$(n+1)/2$

- Application: moyenne des rangs de 1 à 5

Moyenne	$(5+1)/2=3$

Distribution des rangs

- Exemple: somme des rangs ²

Somme	n*(n+1)*(2n+1)/6

- Preuve
  - Quand n = 1: n*(n+1)*(2n+1)/6 = 1 => OK
  - Si ok pour (n-1), alors ok pour n => Tjs OK !

$$\sum_{k=1}^{n-1} k^2 = \frac{(n-1) * n * (2 * n - 1)}{6}$$

$$\Rightarrow \sum_{k=1}^{n} k^2 = \frac{(n-1) * n * (2 * n - 1)}{6} + n^2 = \frac{n * (n+1) * (2 * n + 1)}{6}$$

Distribution des rangs

- Exemple: variance des rangs = $\Sigma (r - \mu_r)^2 / n$
  - On montre facilement que:

Variance	$(n^2-1)/12$

- Application: variance des rangs de 1 à 5

Variance	$(5^2-1)/12=2$

Comparaison de 2 groupes non pairés

- Il s'agit de trouver une alternative non-paramétrique au test de t de comparaison de deux groupes non pairés, quand ce test n'est pas utilisable
- Idée générale:
  - Remplacer les données par leur rang
  - Sommer les rangs dans un des deux groupes
  - Voir si la somme de rangs obtenue est anormalement faible (ou élevée)

Test de Mann-Whitney

- Exemple: Comparaison des anomalies de spermatozoïdes dans deux milieux. Chaque donnée représente des comptages dans un lot.

Milieu I	Milieu II
1	10
2	150
120	155
	220

Rangs

Milieu I	Milieu II
1	3
2	5
4	6
	7

Test de Mann-Whitney

Milieu I	Milieu II
1	3
2	5
4	6
	7

- Dans cet exemple:
  - La somme des rangs du groupe "Milieu I" vaut $W = 1 + 2 + 4 = 7$, pour un groupe de taille $m = n_I = 3$.
  - Sous H_0: pas de différence entre les milieu, la valeur attendue pour W est:

$$\bar{W} = n_I * \underbrace{\frac{(n+1)}{2}}_{\boxed{\text{Rang moyen}}} = n_I * \frac{(n_I + n_{II} + 1)}{2} = 12$$

Test de Mann-Whitney

Milieu I	Milieu II
1	3
2	5
4	6
	7

- Dans cet exemple (suite):
  - On peut calculer la probabilité d'avoir une valeur de $W \leq 7$ quand $\bar{W} = 12$ en listant toutes les situations possibles

Test de Mann-Whitney

1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	3	3	3	3	4	4	4
3	4	5	6	7	4	5	6	7	5	6	7
6	7	8	9	10	8	9	10	11	10	11	12

1	1	1	2	2	2	2	2	2	2	2	2
5	5	6	3	3	3	3	4	4	4	5	5
6	7	7	4	5	6	7	5	6	7	6	7
12	13	14	9	10	11	12	11	12	13	13	14

2	3	3	3	3	3	3	4	4	4	5
6	4	4	4	5	5	6	5	5	6	6
7	5	6	7	6	7	7	6	7	7	7
15	12	13	14	14	15	16	15	16	17	18

$$\#Cas = C_7^3 = \frac{7!}{3! * 4!} = 35$$

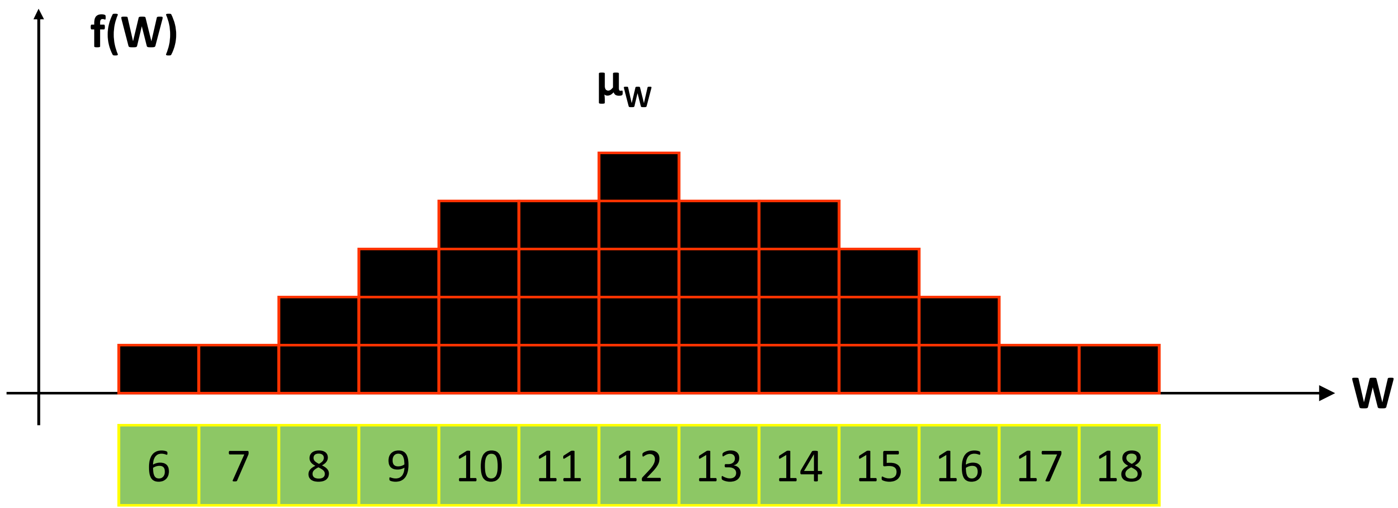
$$\#Cas(W \leq 7) = 2$$

$$\Rightarrow P(W \leq 7) = \frac{2}{35} = 0.057$$

=> Presque significatif au seuil (unilatéral) α = 5%

Test de Mann-Whitney

- On peut déduire la distribution de W pour $n_I = 3$ et $n_{II} = 4$.



- Par simple comptage, on peut calculer toutes les probabilités utiles pour tester des hypothèses
- Ces probabilités sont tabulées pour n_I=m et n_{II}=n <10

Table de Mann-Whitney pour $n_I = 3$

TABLE J
Lower- and upper-tail probabilities for W_x, the Wilcoxon-Mann-Whitney rank-sum statistic*

Entries are $P[W_x \leq c_L]$ and $P[W_x \geq c_U]$. W_x is the rank-sum for the smaller group.

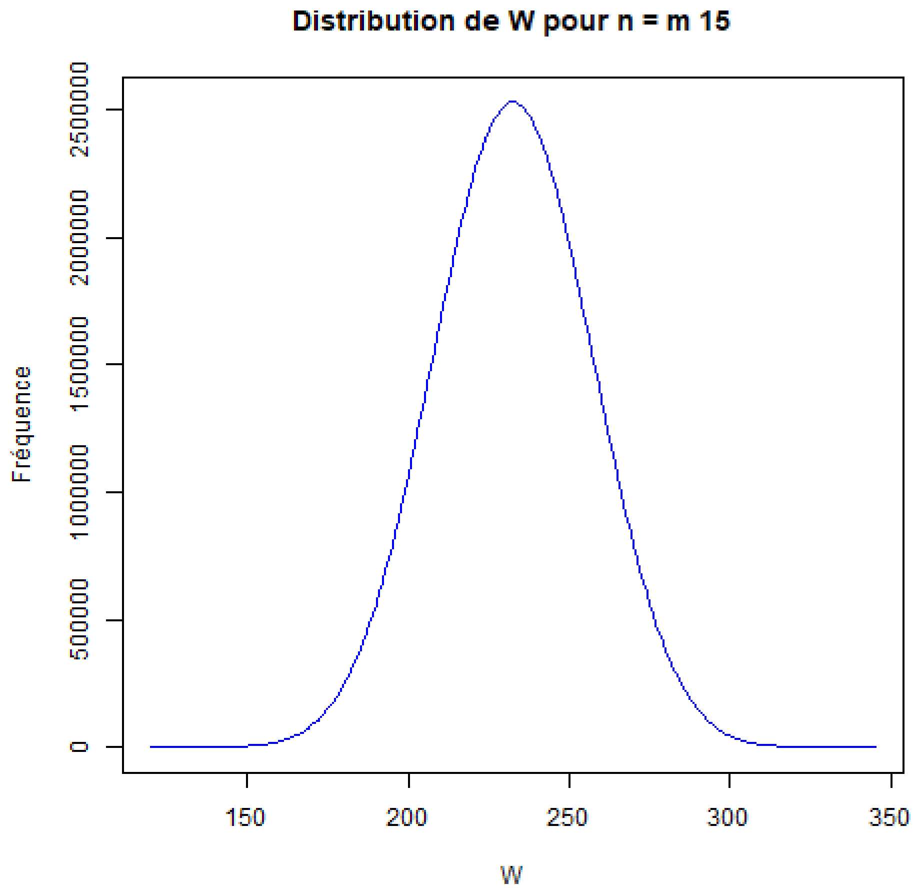
$m = 3$

c_L	$n = 3$	c_U	$n = 4$	c_U	$n = 5$	c_U	$n = 6$	c_U	$n = 7$	c_U	$n = 8$	c_U	$n = 9$	c_U	$n = 10$	c_U	$n = 11$	c_U	$n = 12$	c_U
6	.0500	15	.0286	18	.0179	21	.0119	24	.0083	27	.0061	30	.0045	33	.0035	36	.0027	39	.0022	42
7	.1000	14	.0571	17	.0357	20	.0238	23	.0167	26	.0121	29	.0091	32	.0070	35	.0055	38	.0044	41
8	.2000	13	.1143	16	.0714	19	.0476	22	.0333	25	.0242	28	.0182	31	.0140	34	.0110	37	.0088	40
9	.3500	12	.2000	15	.1250	18	.0833	21	.0583	24	.0424	27	.0318	30	.0245	33	.0192	36	.0154	39
10	.5000	11	.3143	14	.1964	17	.1310	20	.0917	23	.0667	26	.0500	29	.0385	32	.0302	35	.0242	38
11	.6500	10	.4286	13	.2857	16	.1905	19	.1333	22	.0970	25	.0727	28	.0559	31	.0440	34	.0352	37
12	.8000	9	.5714	12	.3929	15	.2738	18	.1917	21	.1394	24	.1045	27	.0804	30	.0632	33	.0505	36
13	.9000	8	.6857	11	.5000	14	.3571	17	.2583	20	.1879	23	.1409	26	.1084	29	.0852	32	.0681	35
14	.9500	7	.8000	10	.6071	13	.4524	16	.3333	19	.2485	22	.1864	25	.1434	28	.1126	31	.0901	34
15	1.0000	6	.8857	9	.7143	12	.5476	15	.4167	18	.3152	21	.2409	24	.1853	27	.1456	30	.1165	33
16			.9429	8	.8036	11	.6429	14	.5000	17	.3879	20	.3000	23	.2343	26	.1841	29	.1473	32
17			.9714	7	.8750	10	.7262	13	.5833	16	.4606	19	.3636	22	.2867	25	.2280	28	.1824	31
18			1.0000	6	.9286	9	.8095	12	.6667	15	.5394	18	.4318	21	.3462	24	.2775	27	.2242	30
19					.9643	8	.8690	11	.7417	14	.6121	17	.5000	20	.4056	23	.3297	26	.2681	29
20					.9821	7	.9167	10	.8083	13	.6848	16	.5682	19	.4685	22	.3846	25	.3165	28
21					1.0000	6	.9524	9	.8667	12	.7515	15	.6364	18	.5315	21	.4423	24	.3670	27
22							.9762	8	.9083	11	.8121	14	.7000	17	.5944	20	.5000	23	.4198	26
23							.9881	7	.9417	10	.8606	13	.7591	16	.6538	19	.5577	22	.4725	25
24							1.0000	6	.9667	9	.9030	12	.8136	15	.7133	18	.6154	21	.5275	24

Test de Mann-Whitney

- Question: Que faire quand m et n > 10 ?
- Réponse: W est une moyenne, et donc tend à se comporter de manière gaussienne !
  - Exemple: n = m = 15 (voir la dia suivante)

Test de Mann-Whitney



Test de Mann-Whitney

- Conséquence: On peut calculer une statistique similaire à z, et vérifier les limites de cette statistique, pour α donné, dans la table de Z

$$Z = \frac{W - \mu_W}{\sigma_W}$$

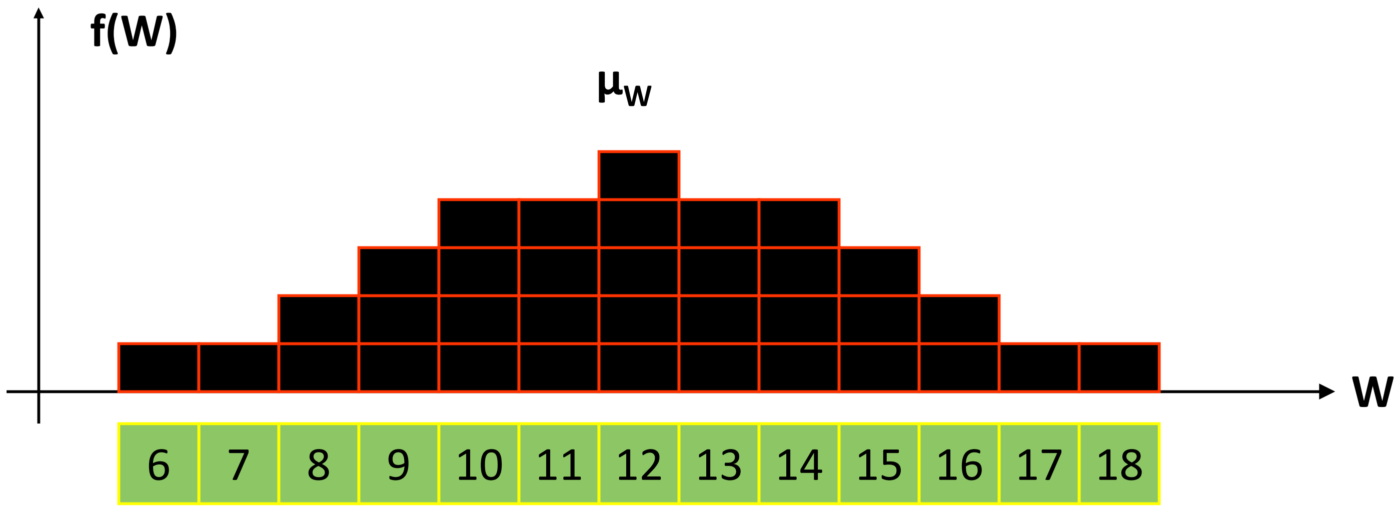
où

$$\mu_W = \frac{n * (m + n + 1)}{2}$$

$$\sigma_W = ?$$

Test de Mann-Whitney

- Question: Comment calculer σ_W ?
- Réponse: Reprenons la distribution de W pour m = 3 et n = 4.



$\Rightarrow \sigma^2_W = 1/35 * [(6 - 12)^2 + (7 - 12)^2 + 2*(8 - 12)^2 + \dots] = 8$

Test de Mann-Whitney

- Question: Comment calculer σ_W ?
- Réponse (suite): pour éviter le calcul exhaustif de σ_W, qui devient fastidieux quand m et n sont élevés, on utilise la formule:

$$\boxed{\sigma_W = \sqrt{\frac{m * n * (m + n + 1)}{12}}}$$

  - Exemple: pour le petit problème avec m = 3 et n = 4; on obtient:

$$\sigma_W = \sqrt{\frac{3 * 4 * (3 + 4 + 1)}{12}} = \sqrt{8}$$

Test de Mann-Whitney: exemple

- Exemple: on a comptabilisé les lymphocytes dans des prélèvements faits sur deux groupes de 15 individus. Les comptages n'ont pas une distribution normale, et on les a remplacés par leur rang (voir dia suivante). Existe-t-il une différence significative (au seuil α = 5%) entre les comptages des 2 groupes ?

Test de Mann-Whitney: exemple

```
> # Données
> groupe_1<-c(2050,1320,910,150,280,2210,1500,
+ 880,2740,2820,550,3050,2100,2390,1860)
> groupe_2<-c(1760,2960,1240,410,1600,1480,2490,
+ 670,1950,1010,320,2680,1100,2510,740)
> # Rangs des données
> rangs_1<-rank(c(groupe_1,groupe_2))[1:15]
> rangs_2<-rank(c(groupe_1,groupe_2))[16:30]
> # Calcul de Mann-Whitney
> W<-sum(rangs_1)
> mu_W<-15*(30+1)/2
> sigma_W<-sqrt(15*15*(15+15+1)/12)
> Z<-(W-mu_W)/sigma_W
> pZ<-1-pnorm(abs(Z))
[1] 0.3467751
```

Test de Kruskal-Wallis

- Question: Que faire quand on a plus de deux groupes ? Exprimé autrement, quelle est l'analogue non-paramétrique de l'ANOVA 1 ?
- Réponse: Dans chaque groupe g, on s'attend à ce que: $W_g = n_g * (N+1) / 2$
  où n_g est la taille du groupe g et N est le nombre total d'observations.

- Idée: Effectuer un χ^2

Test de Kruskal-Wallis

- On peut démontrer le résultat suivant:

$$\boxed{\chi^2_{k-1} = \frac{12}{N * (N+1)} * \sum_{g=1}^{k} \left(n_g * \bar{R}_g^2\right) - 3 * (N+1)}$$

où $\bar{R}_g = \frac{W_g}{n_g}$ est le rang moyen du groupe g

Test de Kruskal-Wallis: exemple

```
> # Données
> groupe_1<-c(512,752,1720,1044,692,1522)
> groupe_2<-c(1374,222,1463,1814,908,1697)
> groupe_3<-c(133,1246,1153,819,448,309)
> # Rangs des données
> rangs_1<-rank(c(groupe_1,groupe_2,groupe_3))[1:6]
> rangs_2<-rank(c(groupe_1,groupe_2,groupe_3))[7:12]
> rangs_3<-rank(c(groupe_1,groupe_2,groupe_3))[13:18]
> # Calcul de Kruskal-Wallis
> n<-length(c(groupe_1,groupe_2,groupe_3))
> n1<-length(groupe_1)
> n2<-length(groupe_2)
> n3<-length(groupe_3)
> kw<-12/(n*(n+1))*((sum(rangs_1)**2)/n1+
+                    (sum(rangs_2)**2)/n2+
+                    (sum(rangs_3)**2)/n3)-3*(n+1)
> p_value<-1-pchisq(kw,df=2)
> p_value
[1] 0.1956205
```

Tests du signe et de Wilcoxon

- Question: Que faire quand on a deux groupes de mesures pairées ? Quelle est l'analogue non-paramétrique du test de t pairé ?
- Une réponse simple: Compter les différences positives et négatives entre mesures pairées. Sous H_0, on en attend le même nombre.

  => Test du Signe

- Idée: Effectuer un χ^2 (ou un test exact) pour tester H0

Test du signe

- Exemple: comptages d'un type cellulaire avant et après traitement. Le traitement a-t-il un effet sur le nombre de cellules de ce type ?

Avant	Après	Diff
100	20	80
120	20	100
18	18	0
1000	30	970
2	5	-3
70	10	60

- 5 différences non nulles
  - 4 positives
  - 1 négative

Test du signe

- Solution:
  - Si H_0 = pas de difference entre avant et après traitement
    - π = probabilité d'une difference positive = 0.5
    - n = nombre de répétitions indépendantes donnant un difference non nulle = 5
    - ⇒ Utilisation de la loi binomiale

Test du signe

- Solution:
  - Si on attend n*π = 2.5 différences positives, quelle est la probabilité d'en voir 4, voire plus ?
  - $P = B(r = 4|n = 5, \pi = 0.5) + B(r = 5|n = 5, \pi = 0.5)$

$$\Rightarrow P = \frac{5!}{4! * 1!} * 0.5^5 + \frac{5!}{5! * 0!} * 0.5^5 = \frac{6}{32} = 0.1875$$

  - On accepte donc H_0 au seuil α = 5%: la différence entre avant et après n'est pas significative (mais l'expérience est très petite…)

Test de Wilcoxon

- Remarque: dans l'exemple précédent, les différences positives étaient souvent de grande amplitude, alors que la seule différence négative était petite (-3)
- Il serait dès lors intéressant d'utiliser la notion de rang dans le test, en comparant la somme des rangs des différences positives à la somme des rangs des différences négatives

=> Test de Wilcoxon

Test de Wilcoxon

- Reprenons l'exemple:

Avant	Après	Diff	|Diff|	Rangs
100	20	80	80	3
120	20	100	100	4
18	18	0	0	0
1000	30	970	970	5
2	5	-3	3	1
70	10	60	60	2

Test de Wilcoxon

Rangs
3
4
0
5
1
2

- Exemple (suite):
  - T+ = Somme des rangs > 0 = 14
  - T- = Somme des rangs < 0 = 1
  - Si H_0 est vraie, on attendait:
    T+ = T- = N/2*(N + 1)/2 = f(N)
  - Les valeurs seuils de T+ (ou T-) sont tabulées pour N < 15.

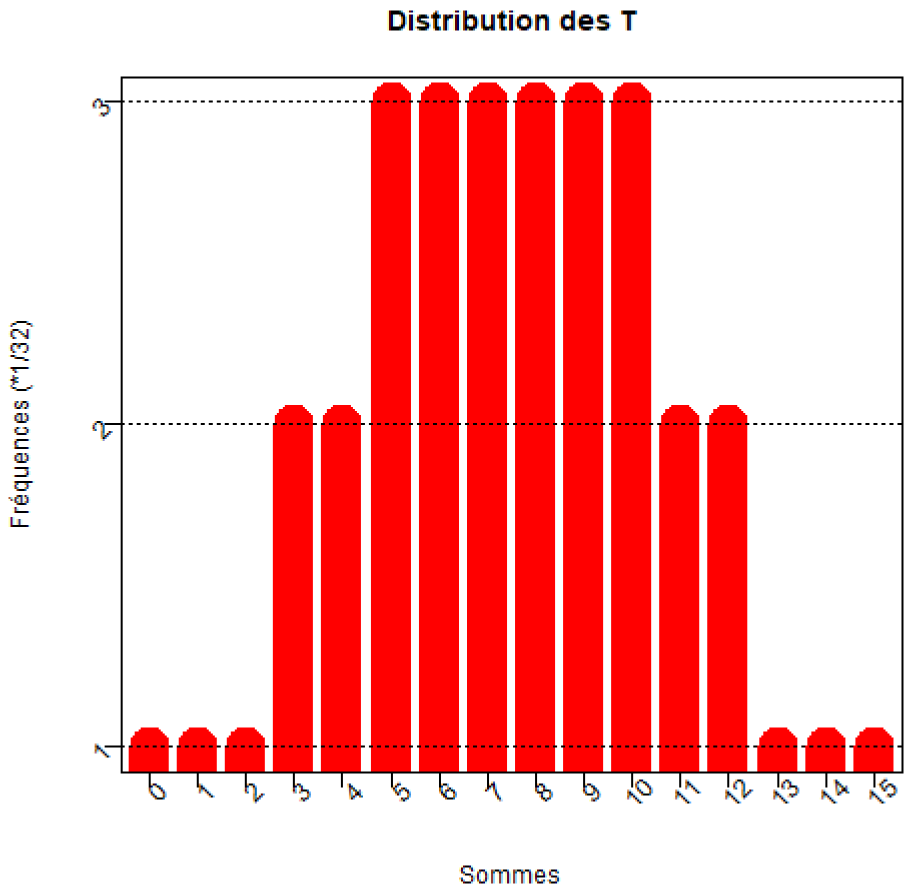
Test de Wilcoxon

- Exemple: construction de la table pour N=5

		1	2	3	4	5	1 2	1 3	1 4	1 5	2 3	2 4	2 5	3 4	3 5	4 5
$\sum$	0	1	2	3	4	5	3	4	5	6	5	6	7	7	8	9

	1 2 3	1 2 4	1 2 5	1 3 4	1 3 5	1 4 5	2 3 4	2 3 5	2 4 5	3 4 5	1 2 3 4	1 2 3 5	1 2 4 5	1 3 4 5	2 3 4 5	1 2 3 4 5
$\sum$	6	7	8	8	9	10	9	10	11	12	10	11	12	13	14	15

Test de Wilcoxon



$$P(T \geq 8) = \frac{16}{32} = 0.50000$$

$$P(T \geq 9) = \frac{13}{32} = 0.40625$$

$$P(T \geq 10) = \frac{10}{32} = 0.31250$$

$$P(T \geq 11) = \frac{7}{32} = 0.21875$$

$$P(T \geq 12) = \frac{5}{32} = 0.15625$$

$$P(T \geq 13) = \frac{3}{32} = 0.09375$$

$$P(T \geq 14) = \frac{2}{32} = 0.06250$$

$$P(T \geq 15) = \frac{1}{32} = 0.03125$$

Test de Wilcoxon

Extrait de la table:

N

c	3	4	5
3	.6250		
4	.3750		
5	.2500	.5625	
6	.1250	.4375	
7		.3125	
8		.1875	.5000
9		.1250	.4063
10		.0625	.3125
11			.2188
12			.1563
13			.0938
14			.0625
15			.0313

T+

Test de Wilcoxon

Rangs
3
4
0
5
1
2

- Exemple (suite):
  - Ici P (T+ >= 14) = 0.0625

    => H_0 sera acceptée (mais on est proche de la signification, et l'échantillon est petit...)

Test de Wilcoxon

- Question: que faire quand N > 15 ?
- Réponse: on utilise le théorème de la limite centrale, qui conduit à utiliser l'approximation normale:

$$Z = \frac{W - \mu_W}{\sigma_W}$$

où

$$\mu_W = \frac{N * (N+1)}{4}$$

$$\sigma_W = \sqrt{\frac{N * (N+1) * (2 * N + 1)}{24}}$$

Corrélation de Spearman

- Question: Comment calculer une corrélation quand les données ne sont pas issues de distributions normales ?

- Idée: Utiliser les rangs
  => Corrélation de rangs

Corrélation de Spearman

- Méthode: On calcule la corrélation comme d'habitude, mais en utilisant les rangs.

$$r_S = \frac{\sum\left(R_X - \frac{N+1}{2}\right) * \left(R_Y - \frac{N+1}{2}\right)}{\sqrt{\sum\left(R_X - \frac{N+1}{2}\right)^2 * \sum\left(R_Y - \frac{N+1}{2}\right)^2}}$$

- Formule: Cette équation peut se simplifier
  - En remplaçant les deux variances (identiques) au numérateur par leur valeur (N^2 - 1) / 12
  - En développant algébriquement

Corrélation de Spearman

- Formule: on obtient alors la formule suivante:

$$\boxed{r_S = 1 - \frac{6 * \sum (R_X - R_Y)^2}{N * (N^2 - 1)}}$$

Corrélation de Spearman

Test de la corrélation:

- Pour tester H_0: $\rho_s = 0$, on peut recourir à la même approche (simulations) que pour la corrélation simple.

- De manière alternative, on peut utiliser:

$$t_{n-2} = r_S * \sqrt{\frac{N-2}{1-r_S^2}}$$

```
> # Données
> groupe_1<-c(2050,1320,910,150)
> groupe_2<-c(1760,2960,1240,410)
> # Rangs des données
> rangs_1<-rank(groupe_1)
> rangs_2<-rank(groupe_2)
> # Calculs de la corrélation de Spearman
> # Methode 1
> r_spearman_1<-cor(rangs_1,rangs_2)
> r_spearman_1
[1] 0.8
> # Methode 2
> r_spearman_2<-cor(groupe_1,groupe_2,method="spearman")
> r_spearman_2
[1] 0.8
> # Methode 3
> N<-length(groupe_1)
> r_spearman_3<-1-6*sum((rangs_1-rangs_2)**2)/(N*(N**2-1))
> r_spearman_3
[1] 0.8
> # Test de H0: rho_spearmann=0
> t<-r_spearmann_1*sqrt((N-2)/(1-r_spearmann_1**2))
> p_value<-1-pt(t,N-2)
> p_value
[1] 0.1
```

Merci pour votre attention et votre présence, parfois faible quantitativement, parfois à distance, mais, j'en suis convaincu, forte qualitativement !
Bonne chance pour la suite de votre année !