Comparaison de moyennes

Tests visant à mettre en évidence une corrélation une variable quantitative et une variable qualitative

M. L. Delignette-Muller
VetAgro Sup

7 octobre 2020

Objectifs pédagogiques

- Comprendre les différences entre un test paramétrique et un test non paramétrique
- Savoir réaliser à la main les deux tests de Student (séries indépendantes ou appariées) et les tests non paramétriques associés (somme des rangs et rangs signés).*
- Connaître le principe de l'analyse de variance
- Connaître le principe des méthodes de comparaisons multiples
- Savoir interpréter les résultats d'un test de normalité et d'un test de comparaison de variances et en connaître les limites.
- Savoir choisir et réaliser le test adapté pour comparer deux ou plusieurs séries d'une variable quantitative en fonction de la question posée, du plan d'expérience et des données.**

* *savoir faire évalué uniquement en S5*

** *savoir faire évalué uniquement en S6 après entraînement en TD*

Plan

Exemple de comparaison de moyennes sur 2 séries indépendantes

Un essai randomisé a été réalisé sur 18 chiens, afin d'évaluer l'efficacité d'un supplément alimentaire contre la formation de tartre sur les dents de l'animal.
Neuf chiens reçoivent une alimentation supplémentée (**groupe supplément**) et neuf chiens ne reçoivent aucune supplémentation (**groupe témoin**).
La formation de tartre est quantifiée par un index combinant la proportion de dents atteintes et l'épaisseur de la couche de tartre formée. Les index moyens observés sont respectivement de 0.747 pour le groupe supplément et de 1.089 pour le groupe témoin.
Cette différence est-elle significative ?

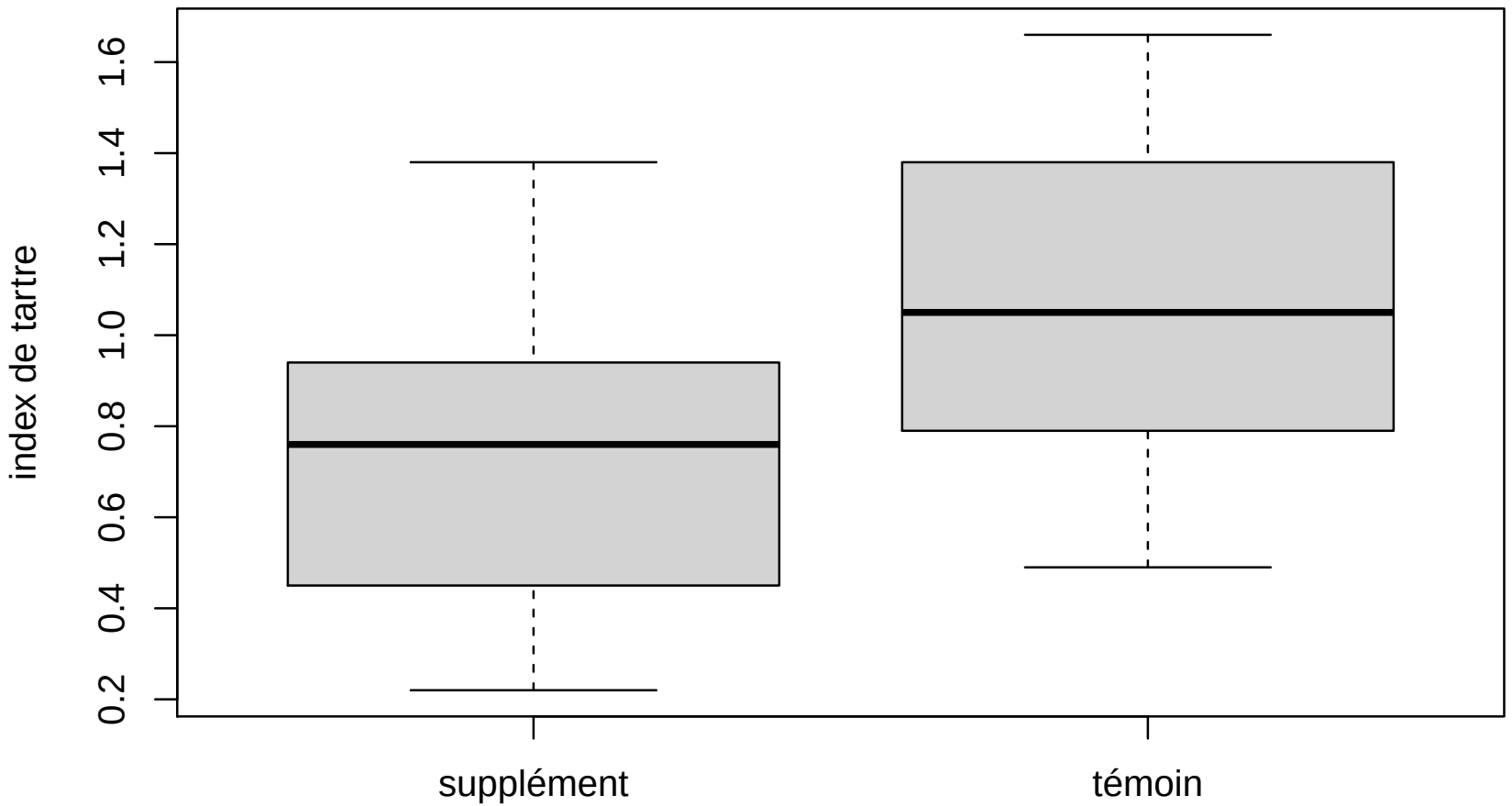
Les données observées

Données brutes (telles que saisies informatiquement) :

index	traitement
0.49	temoin
1.05	temoin
0.79	temoin
1.35	temoin
0.55	temoin
1.38	temoin
1.55	temoin
1.66	temoin
1.00	temoin
0.34	supplement
0.76	supplement
0.45	supplement
0.69	supplement
0.87	supplement
0.94	supplement
0.22	supplement
1.07	supplement
1.38	supplement

Visualisation des données brutes

Représentation classique : diagrammes en boîte



Test paramétrique de Student

La démarche paramétrique va supposer
que le **théorème de l'approximation normale s'applique**.
Le test de Student va de plus supposer **les variances égales**.
Variable de décision et sa loi sous H_0 :

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T(n_1 + n_2 - 2)$$

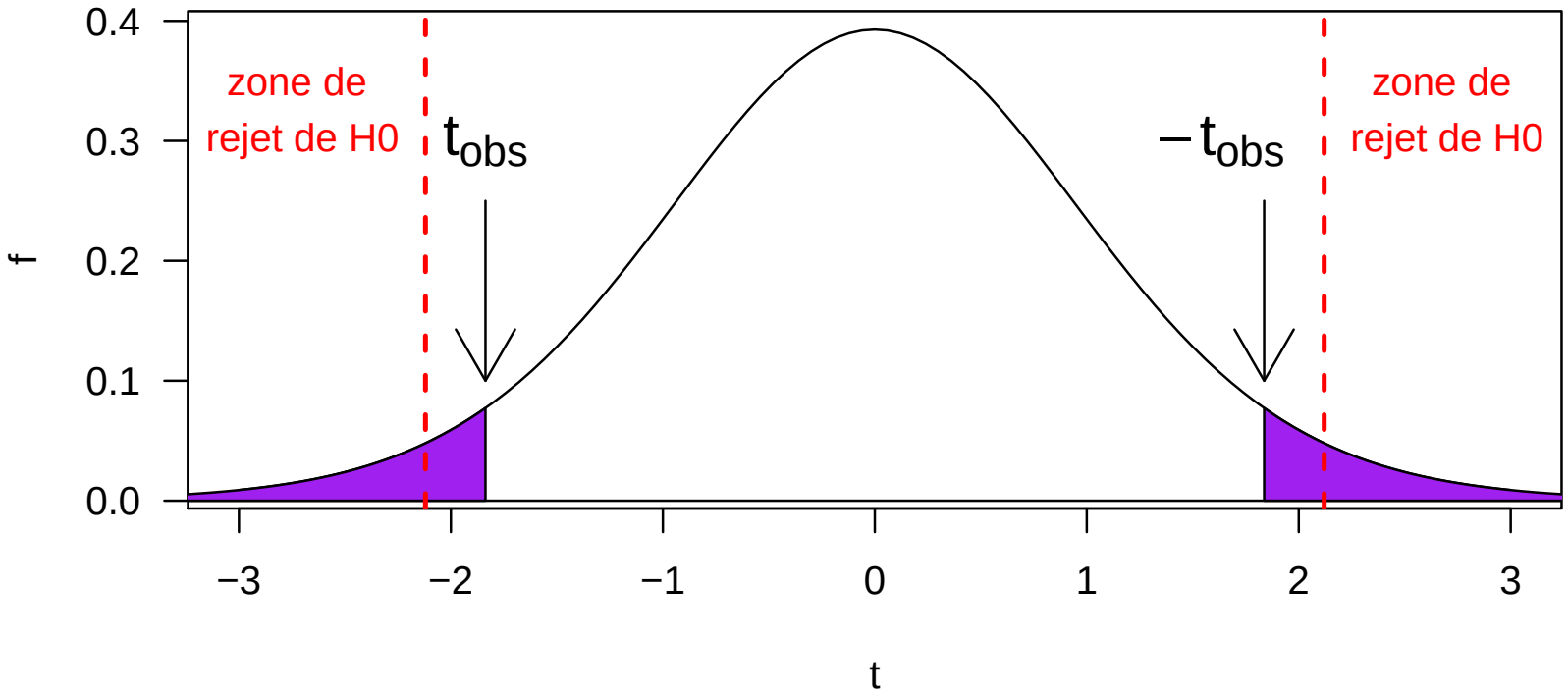
avec $\hat{\sigma} = \sqrt{\frac{(n_1-1)\hat{\sigma_1}^2 + (n_2-1)\hat{\sigma_2}^2}{n_1+n_2-2}}$
et $T(n_1 + n_2 - 2)$ la loi de Student de degré de liberté $n_1 + n_2 - 2$

Calcul de la valeur de p à partir de la valeur de t observée

$p = Pr(|t| > |t_{obs}|)$ (aire violette sur le graphe)

Dans cet exemple $t_{obs} = -1.84$ ce qui correspond à $p > 0.05$.

On ne peut donc pas conclure à une différence significative.



Intervalle de confiance associé au test paramétrique de Student

Intervalle de confiance sur la différence entre les 2 moyennes :

$$\mu_1 - \mu_2 = \overline{x}_1 - \overline{x}_2 \pm t_{n_1+n_2-2;1-\frac{\alpha}{2}} \times \hat{\sigma} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

avec $t_{n_1+n_2-2;1-\frac{\alpha}{2}}$ le quantile à $1 - \frac{\alpha}{2}$ de la loi de Student de degré de liberté $n_1 + n_2 - 2$.

On **rejette en fait l'hypothèse** H_0 d'égalité des moyennes par le test de Student **dès que l'intervalle de confiance** à 95% sur la différence entre les 2 moyennes **ne contient pas la valeur 0**.

Cet intervalle de confiance est de plus **informatif quel que soit le résultat du test**.

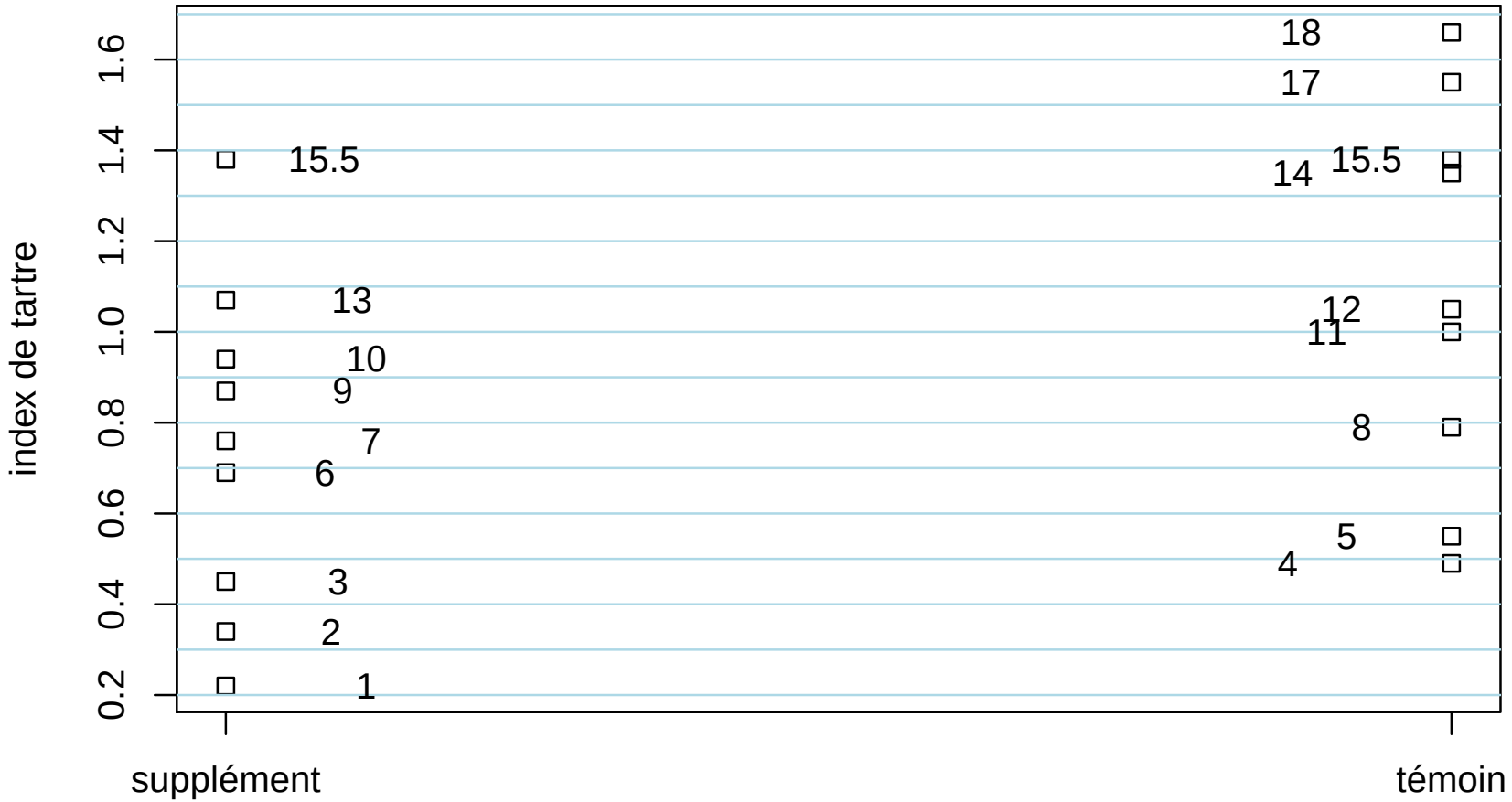
Estimation de la différence avec son intervalle de confiance à 95% : -0.34 $[-0.74; 0.05]$

(estimation très imprécise mais pas en faveur d'une différence nulle)

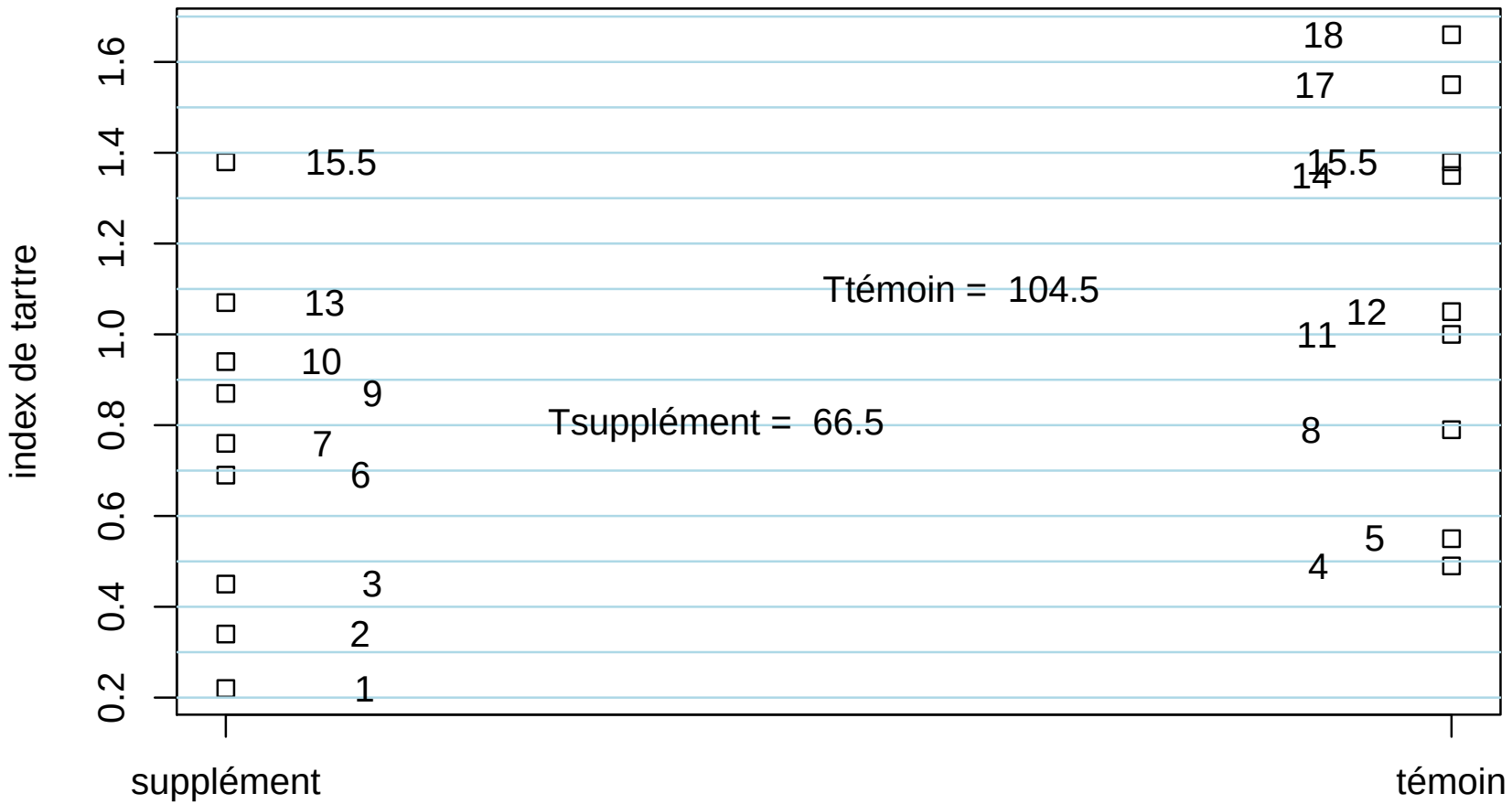
Approches paramétrique et non paramétrique

- **Test paramétrique**
  La variable de décision est calculée **à partir d'un paramètre statistique caractérisant une loi donnée** (souvent la loi normale).
- **Test non paramétrique**
  On ne fait plus d'hypothèse quant à la forme des distributions et on utilise le plus souvent des **statistiques de rang** qui n'utilisent comme information que l'ordonnancement des observations entre elles (plus robustes, c'est-à-dire moins sensibles aux valeurs extrêmes).

Sur l'exemple précédent, calcul des rangs des observations dans un classement global



Calcul des sommes des rangs par groupe



Principe du test de Mann-Whitney-Wilcoxon - test de la somme des rangs

- **Classement global des observations**
  Affectation de son rang à chaque observation en moyennant les rangs des ex-aequos
- **calcul de la somme des rangs de chacun des groupes**
  Dans l'exemple $T_{supplement} = 66.5$ et $T_{temoin} = 104.5$
- **Comparaison, à l'aide d'une variable de décision adapatée, des 2 sommes des rangs**
  Dans l'exemple on obtient une valeur de $p > 0.05$ (cf. fiche technique pour réalisation complète).
  On ne peut donc pas conclure à une différence significative.

Comment choisir entre test paramétrique et test non paramétrique ?

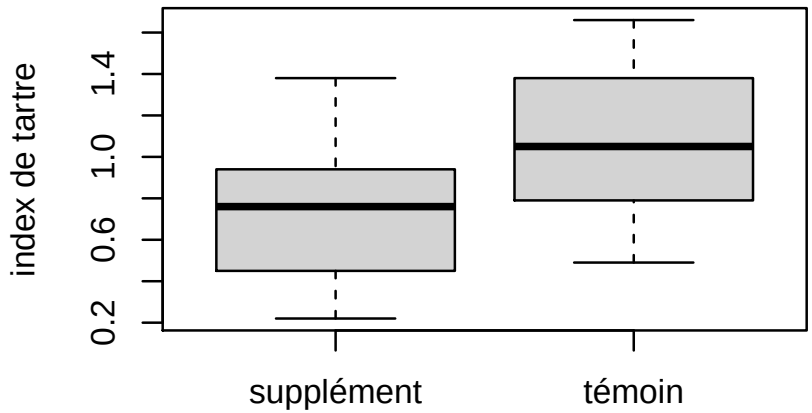
- **Test paramétrique**
  Hypothèse forte sur la forme des distributions.
  Conditions d'utilisation assez restrictives.
  Intervalle de confiance associé pouvant s'avérer très informatif surtout en cas de non rejet de H_0.
- **Test non paramétrique**
  Pas d'hypothèse forte quant à la forme des distributions, mais dégradation de l'information initiale qui peut induire à une perte de puissance.
  Pas d'intervalle de confiance associé.

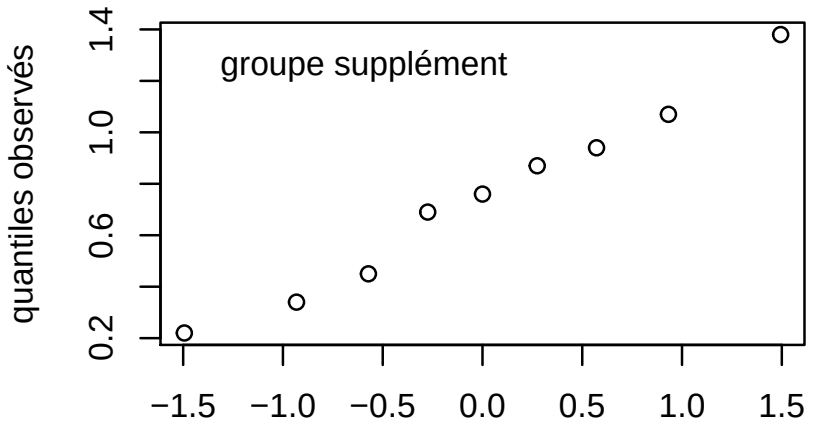
Test paramétrique à privilégier, si possible, éventuellement après transformation de variable.

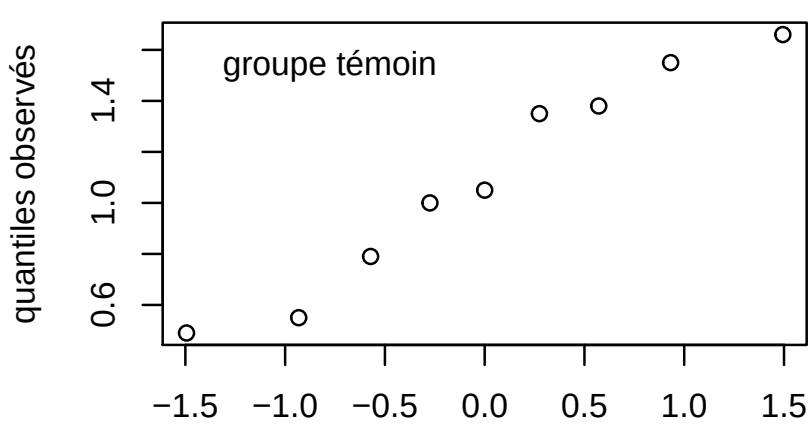
Revenons à l'exemple

Peut-on utiliser un test paramétrique ?
C'est-à-dire **peut-on appliquer le théorème de l'approximation normale ?**
La variable est un index combinant diverses informations (variable de type score). Rien ne garantit à l'avance la normalité de sa distribution.
Les effectifs ne sont pas très grands (deux groupes de 9).
Qu'en est-il de la forme des distributions observées ?

Examen des distributions







Choix de la démarche sur l'exemple

L'observation des données ne conduit pas à remettre en cause l'hypothèse de normalité des distributions.
Néanmoins les effectifs ne sont vraiment pas très grands.
On est ici dans un **cas un peu limite** ou certains choisiraient une démarche paramétrique et d'autres une démarche non paramétrique.

Dans le cas du choix d'une démarche paramétrique, il serait **raisonnable de supposer les variances égales**
(écarts types du même ordre de grandeur,
0.37 pour le groupe supplément et 0.42 pour le groupe témoin
et dispersions comparables d'après les diagrammes en boîte).

Pourquoi est-il inapproprié d'utiliser des tests pour vérifier les hypothèses de normalité et d'égalité des variances ?

De nombreux scientifiques utilisent le **test de Fisher de comparaison de variances** pour vérifier l'égalité des variances avant de comparer les moyennes,
et le **test de Shapiro-Wilk de normalité** pour vérifier la normalité d'une distribution.
CES TESTS NE PEUVENT EN AUCUN CAS DE REMPLACER UN EXAMEN VISUEL DES DISTRIBUTIONS !

Pourquoi ?

A quoi sert le test de Fisher de comparaison de 2 variances ?

H_0 : "égalité des variances"
Seule conclusion possible sans calcul de puissance au préalable (cas classique) :
rejet de H_0
donc **mise en évidence d'une différence entre 2 variances**
(qui peut être intéressant en soi par exemple si on étudie la répétabilité d'une mesure réalisée avec deux méthodes différentes).
Ce test n'a d'intérêt que pour mettre en évidence une différence significative entre 2 variances,
mais ne permet jamais de conclure à l'égalité entre 2 variances.

A quoi sert un test de normalité ?

H_0 : "normalité de la distribution"
Seule conclusion possible sans calcul de puissance au préalable (cas classique) :
rejet de H_0
donc **mise en évidence d'un écart à la normalité**.
Ce test **ne permet pas de montrer la normalité d'une distribution**,
mais dans certains cas il peut mettre en évidence un écart à la normalité (de nature ou non à remettre en cause l'applicabilité du théorème de l'approximation normale)

Le test de normalité : jamais là quand on a besoin de lui !

- **Cas des petits effectifs**
  On a vraiment besoin de savoir si la distribution est normale pour pouvoir appliquer le théorème de l'approximation normale mais le test a peu de chance de mettre en évidence un écart à la normalité (**faible puissance**).
- **Cas des grands effectifs**
  Le test mettra en évidence des écarts à la normalité même faibles (**forte puissance**),
  qui ne devront pas forcément remettre en cause l'applicabilité du théorème de l'approximation normale, d'autant moins que l'effectif est grand.

Plan

Comparaison d'une moyenne observée à une moyenne théorique (un seul échantillon)

Exemple

Un laboratoire d'analyse indique comme valeur moyenne de l'urée plasmatique chez les chats sains, une valeur de 8.5 mmol/l. Suite à un remplacement de ses appareils de mesure, le laboratoire dose l'urée sur un échantillon aléatoire de 140 chats en bonne santé.
Valeurs obtenues :
m = 9.7 mmol/l et SD = 2.6 mmol/l

La moyenne observée est-elle significativement différente de la valeur moyenne de référence indiquée par le laboratoire ?

Comparaison d'une moyenne observée à une moyenne théorique - les tests

- **Test paramétrique**
  test de conformité de **Student** si le théorème de l'approximation normale s'applique
  (dans cet exemple $t_{obs} = 5.46 \rightarrow p < 0.001$, cf. fiche technique pour détails et représentation des données)
- **Test non paramétrique**
  test de la médiane sinon
  Principe : **la valeur théorique est-elle au milieu des observations ?**
  on compte les effectifs observés de part et d'autre de la valeur théorique, et on les compare aux effectifs théoriques 50% - 50%, à l'aide d'un test du χ^2 d'ajustement
  (cf. comparaison d'une fréquence observée à une théorique).

Comparaison de moyennes sur 2 échantillons indépendants

cf. exemple introductif

- **Tests paramétriques**
  si le théorème de l'approximation normale s'applique
  - **test de Student avec "variances égales"**
    s'il est raisonnable de supposer les écarts types égaux
  - **test de Welch**
    appelé aussi test de Student avec variances inégales
    si les écarts types semblent différents
    et qu'il reste intéressant de comparer les moyennes.
- **Test non paramétrique**
  test de la somme des rangs de Mann-Whitney-Wilcoxon

Comparaison de moyennes sur 2 échantillons dépendants (appariés)

Exemple

On veut comparer une nouvelle méthode de dosage de l'urée urinaire (méthode 2) à la méthode de référence (méthode 1). Pour cela on a dosé l'urée par les 2 méthodes chez 12 animaux.
On obtient respectivement des moyennes de 27.7 et 28.8 g/24h pour les méthodes 1 et 2.

La différence observée entre ces moyennes est-elle significative ?

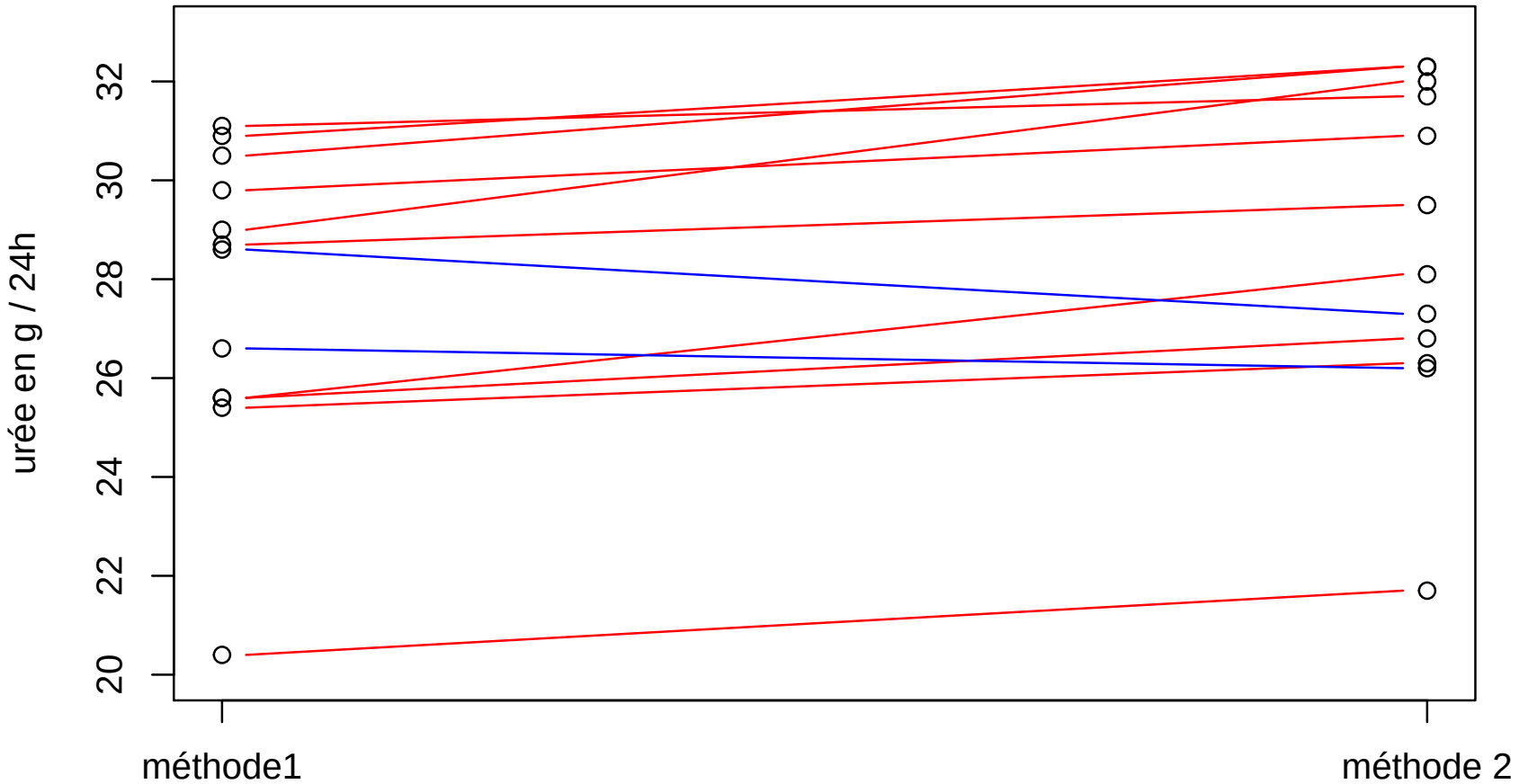
Les données observées

Données brutes (telles que saisies informatiquement) :

	indiv	meth1	meth2
indiv1		20.4	21.7
indiv2		25.4	26.3
indiv3		25.6	26.8
indiv4		25.6	28.1
indiv5		26.6	26.2
indiv6		28.6	27.3
indiv7		28.7	29.5
indiv8		29.0	32.0
indiv9		29.8	30.9
indiv10		30.5	32.3
indiv11		30.9	32.3
indiv12		31.1	31.7

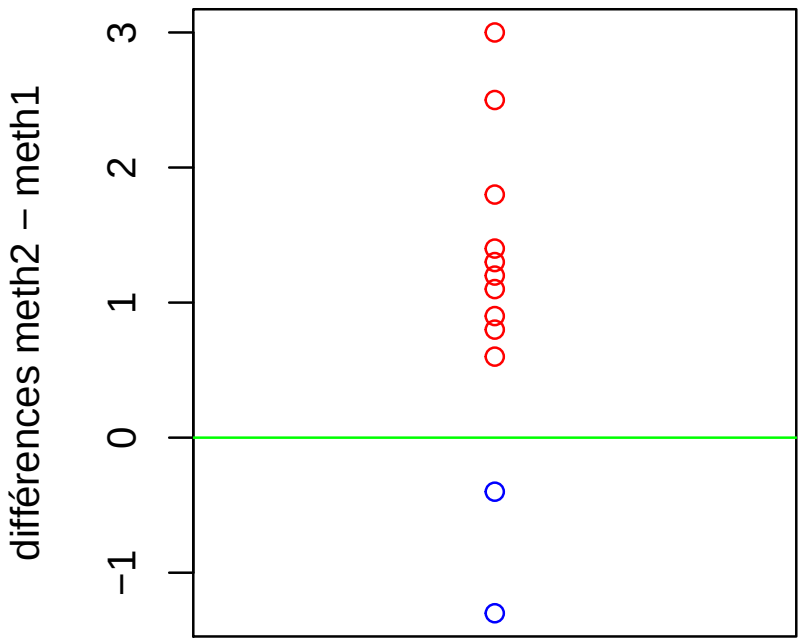
Visualisation des données brutes

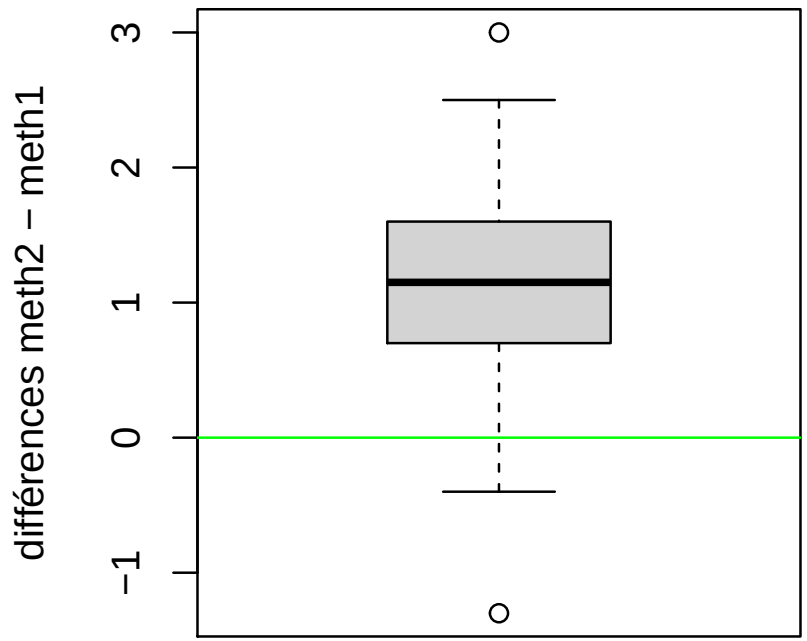
avec visualisation de l'appariement.
Différences *meth*2 – *meth*1 positives en rouge et négatives en bleu.



Examen de la distribution des différences *meth*2 – *meth*1

Comparer les moyennes des 2 groupes revient à comparer la moyenne des différences à 0.
On est ramené à un test de comparaison d'une moyenne observée sur un échantillon (des différences) à 0.





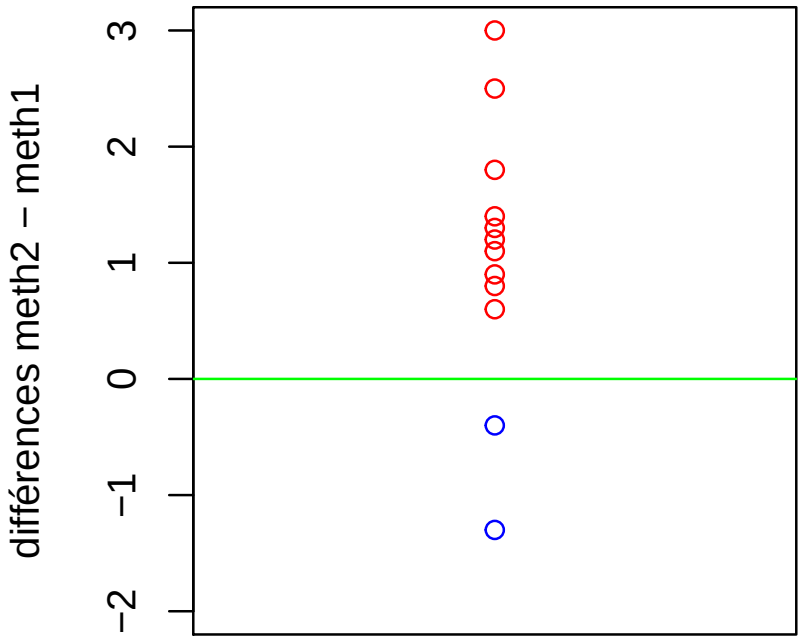
Comparaison de moyennes sur 2 échantillons dépendants (appariés)

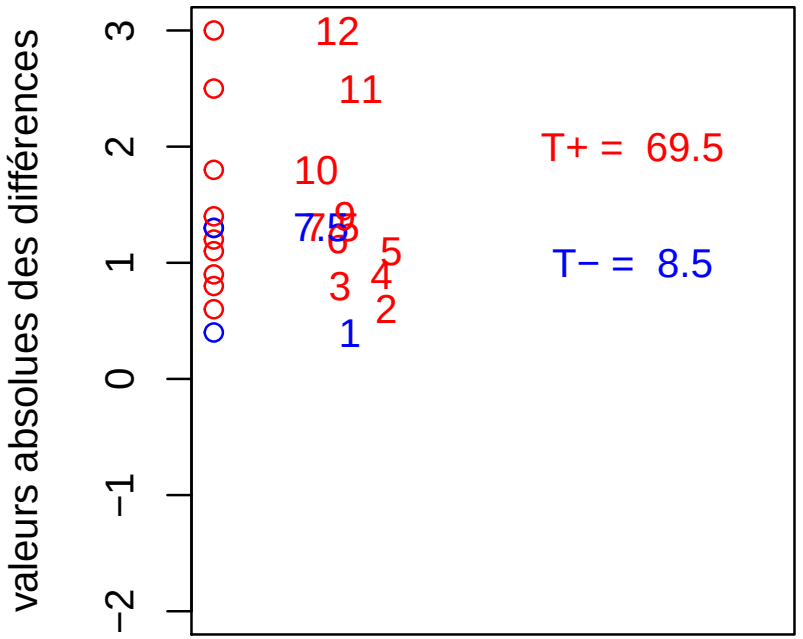
Revient à tester l'égalité à 0 de la moyenne des différences.

- **Test paramétrique**
  Test de Student des séries appariées si le théorème de l'approximation normale s'applique si la distribution des différences.
  Sur cet exemple on obtiendrait
  $t_{obs} = 3.23 \rightarrow 0.001 < p < 0.01$ et une différence estimée à 1.075 avec un intervalle de confiance à 95% de $[0.34; 1.81]$.
- **Test non paramétrique**
  Test des rangs signés de Wilcoxon sinon.

Principe du test des rangs signés de Wilcoxon

Classement des différences en valeur absolue puis comparaison de la somme des rangs T+ des différences positives à la somme des rangs T- des différences négatives (on obtiendrait ici $0.01 < p < 0.05$, cf. fiche technique pour réalisation complète).





Comparaison de moyennes sur plusieurs échantillons indépendants

Exemple

A partir d'un échantillon de 928 chiennes d'élevage, on voudrait savoir si **la durée de gestation dépend de la taille des races**

4 groupes :

- races géantes (XL) : $\overline{x}_1 = 62.5$ sur $n_1 = 77$ chiennes,
- grandes races (L) : $\overline{x}_2 = 61.4$ sur $n_2 = 281$ chiennes,
- races moyennes (M) : $\overline{x}_3 = 61.6$ sur $n_3 = 242$ chiennes
- et petites races (S) : $\overline{x}_4 = 61.6$ sur $n_4 = 328$ chiennes,

Autrement dit,

les **durées moyennes de gestation sont-elles différentes** entre les 4 groupes de taille de race ?

(exemple inspiré de la thèse vétérinaire de Mathilde Poinssot, Maisons Alfort, 2011)

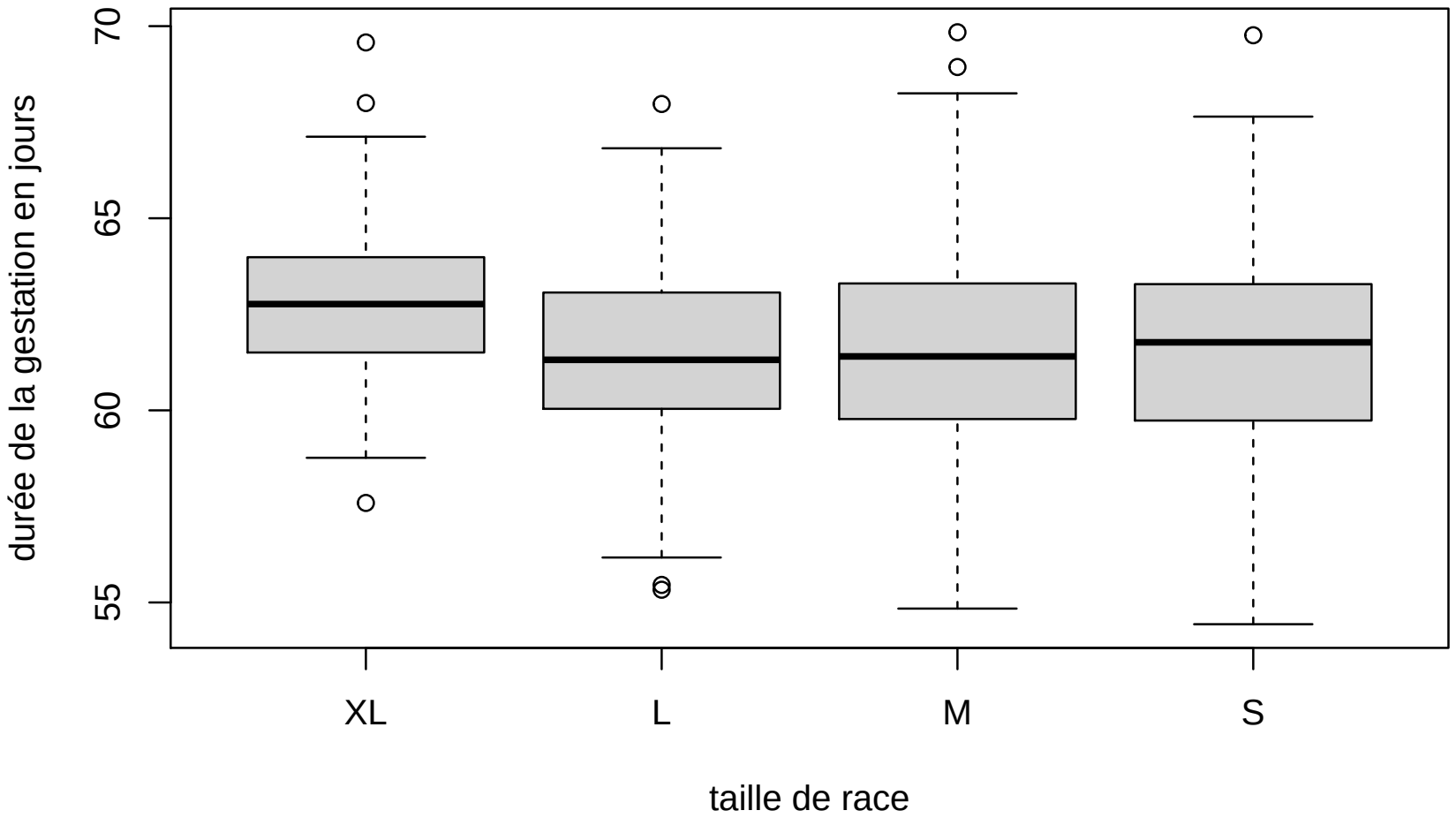
Les données observées

Données brutes (telles que saisies informatiquement) :

taille	duree
XL	64.5
XL	63.9
XL	63.5
XL	63.4
XL	63.9
XL	62.0
XL	66.9
XL	66.8
XL	59.8
XL	65.3
XL	60.6
XL	62.2

...

Visualisation classique des données sous forme de diagrammes en boîte



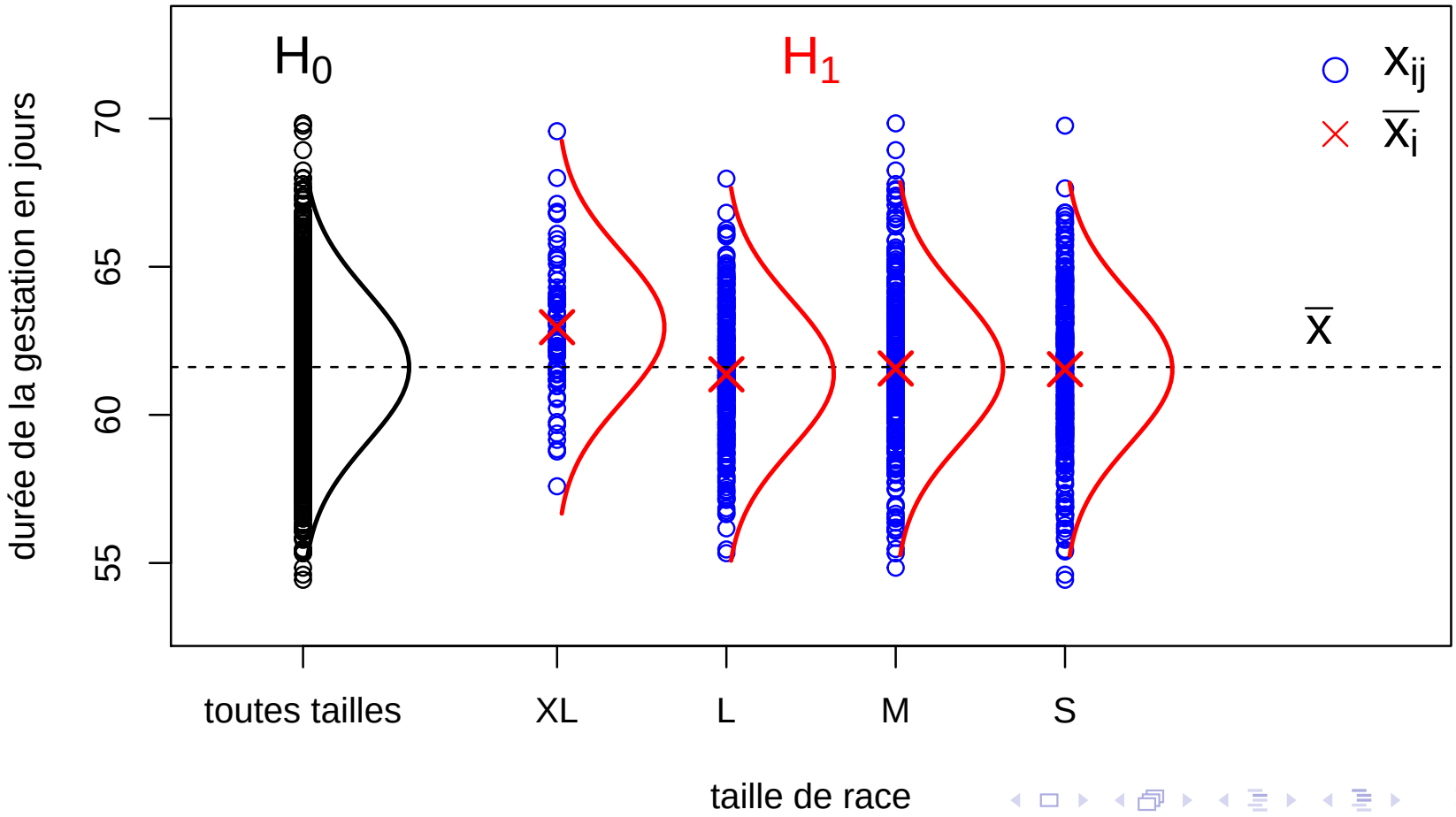
Comparaison de moyennes sur plusieurs échantillons indépendants

- **Test paramétrique**
  si le théorème de l'approximation normale s'applique et que les variances peuvent être supposées égales, réalisation d'une **analyse de variance à un facteur (ANOVA)**, généralisation du test de Student avec variances égales.
- **Test non paramétrique**
  test de la somme des rangs de Kruskal-Wallis, généralisation du test de Mann-Whitney-Wilcoxon basé exactement sur le même principe.

Modèle de l'analyse de variance à un facteur (ANOVA 1)

$x_{ij} = \mu + \alpha_i + \epsilon_{ij}$ avec $\epsilon_{ij} \sim N(0, \sigma)$

$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$

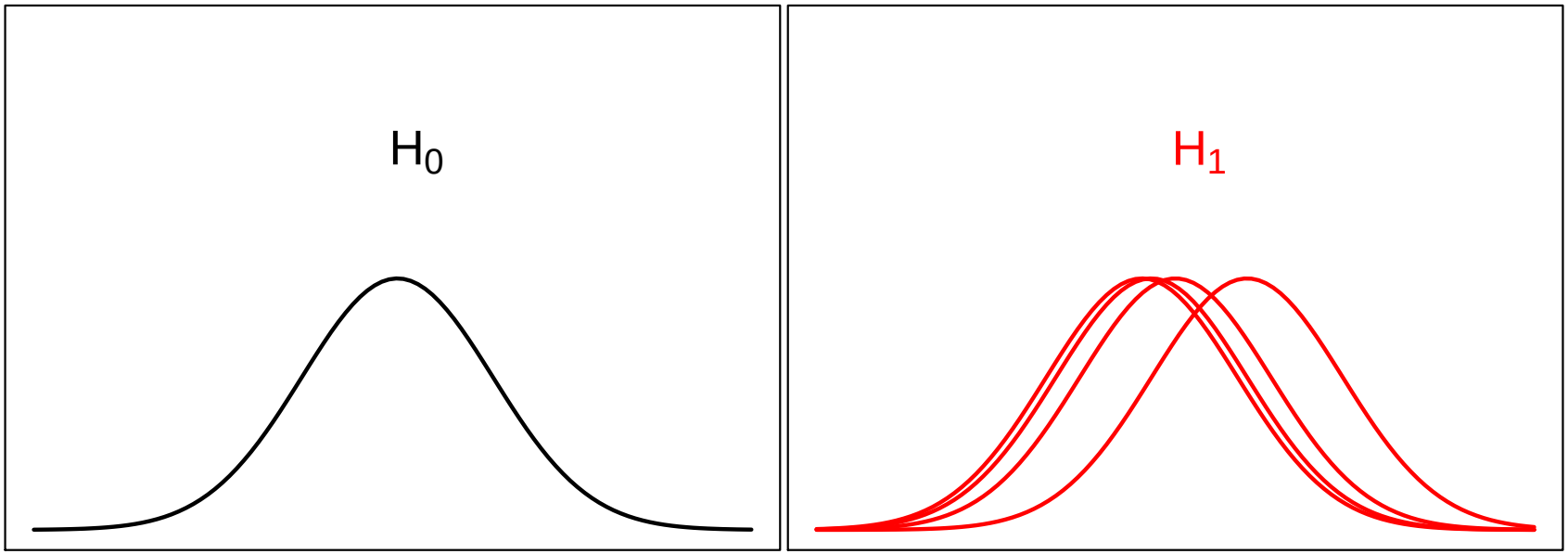


Modèle de l'analyse de variance à un facteur (ANOVA 1) autre représentation schématique

$x_{ij} = \mu + \alpha_i + \epsilon_{ij}$ avec $\epsilon_{ij} \sim N(0, \sigma)$

H_0 : α_i tous nuls $\Rightarrow$ une même dist. $N(\mu, \sigma)$

H_1 : au moins un des α_i non nul $\Rightarrow$ plusieurs dist. $N(\mu + \alpha_i, \sigma)$



Principe de l'analyse de variance à un facteur (ANOVA 1)

On appelle **facteur la variable qualitative définissant les groupes** (ici la taille de race)
ANOVA 1 = méthode de comparaison globale de plusieurs moyennes basée sur une décomposition de la variance totale en une **variance intra-groupe (résiduelle)** et une **variance inter-groupe (factorielle)** et sur la comparaison des ces 2 variances.
Décomposition de la variation totale (ou somme des carrés des écarts totale) :

$$SCE_T = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2 =$$
$$\sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2 + \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\overline{x}_i - \overline{x})^2 =$$
$$SCE_R + SCE_A$$

avec p le nombre de modalités du facteur A (nombre de groupes)

et n_i l'effectif du groupe i

Test de l'analyse de variance à un facteur

Estimation des variances intra-groupe et inter-groupe appelés aussi carrés moyens

$$CM_R = \frac{SCE_R}{\sum_{i=1}^{p}(n_i-1)} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2}{\sum_{i=1}^{p}(n_i-1)} = \frac{\sum_{i=1}^{p}(n_i-1)\hat{\sigma}_i^2}{N-p}$$

$$CM_A = \frac{SCE_A}{p-1} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(\overline{x}_i-\overline{x})^2}{p-1} = \frac{\sum_{i=1}^{p} n_i(\overline{x}_i-\overline{x})^2}{p-1}$$

Comparaison des variances intra-groupe et inter-groupe
Sous H_0 **(égalité de toutes les moyennes)**,
$F = \frac{CM_A}{CM_R}$ suit la **loi** $F(p-1, N-p)$ **de Fisher et Snédécor** de degrés de liberté $p-1$ et $N-p$.
Rejet de H_0 si $CM_A \gg CM_R$.

Résultats de l'analyse de variance sur l'exemple

Sortie du logiciel R

```
Analysis of Variance Table

Response: d$duree
            Df Sum Sq Mean Sq F value  Pr(>F)
d$taille     3    159    52.9    8.36 1.7e-05 ***
Residuals  924   5849     6.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

On conclut à une différence globale significative, donc à un impact de la taille de race sur la durée de gestation.

Les comparaisons multiples de moyennes sur plusieurs échantillons indépendants

Suite à la mise en évidence d'une différence globale significative entre plusieurs moyennes (par ANOVA ou test de Kruskal-Wallis), on souhaite parfois comparer les moyennes 2 à 2.

Méthode basique dite PLSD de Fisher (protected least significant difference) : statistique de Student utilisée pour chaque test mais avec σ commun estimé à partir de l'ensemble des groupes.

Problème majeur associé à ce type de comparaisons multipes : **tests répétés $\Rightarrow$ inflation du risque α global**

Nécessité de corriger le risque α (ou les valeurs de p) si on veut maîtriser le risque α global

(risque α global $=$ probabilité de détecter au moins une différence significative parmi toutes celles testées si on est sous H_0).

La méthode de Bonferroni : une méthode classique pour éviter l'inflation du risque α

Utilisable après la mise en évidence d'une différence globale significative entre plusieurs moyennes (par ANOVA ou test de Kruskal-Wallis).

- **Principe**
  Pour chaque test on corrige α ($\alpha_{cor} = \frac{0.05}{k}$) ou de façon équivalente on corrige p ($p_{cor} = p \times k$), avec k le nombre de tests réalisés, afin d'être sûr que $\alpha_{global} < 5\%$.
- **Cadre d'utilisation**
  Trop conservatif lorsque le nombre de groupes augmente.
  Il arrive alors souvent qu'une différence globale soit significative sans qu'aucune différence 2 à 2 n'apparaisse significative.

Une amélioration de la méthode de Bonferroni souvent préconisée actuellement : la méthode de Bonferroni-Holm

Même cadre d'utilisation que la méthode de Bonferroni.

- **Principe**
  - on classe les valeurs de p par ordre croissant $(p_1, p_2, \ldots, p_k)$
  - on corrige chaque p_i en le multipliant par $k + 1 - i$ $(p_{i.cor} = p_i \times (k + 1 - i))$
- **Avantage**
  méthode moins conservative que Bonferroni tout en maintenant le risque α global $< 5\%$.

Comparaisons multiples suite à l'analyse de variance sur l'exemple - méthode PLSD de Fisher

Sortie du logiciel R sans correction - méthode PLSD de Fisher

Donne les valeurs de p non corrigées.

```
        Pairwise comparisons using t tests with pooled SD

data:  d$duree and d$taille

  XL    L   M
L 1e-06 -   -
M 2e-05 0.4 -
S 9e-06 0.4 0.9

P value adjustment method: none
```

Comparaisons multiples suite à l'analyse de variance sur l'exemple - méthode de Bonferroni

Sortie du logiciel R avec correction de Bonferroni
Donne les valeurs de p corrigées, c'est-à-dire multipliées par k, en fixant à 1 les valeurs > 1.

```
        Pairwise comparisons using t tests with pooled SD

data:  d$duree and d$taille

  XL    L M
L 6e-06 - -
M 1e-04 1 -
S 6e-05 1 1

P value adjustment method: bonferroni
```

Dans ce cas le résultat est parlant : seul le groupe des chiens de races géantes se distingue significativement des autres.

Comparaisons multiples suite à l'analyse de variance sur l'exemple - méthode de Bonferroni - Holm

Sortie du logiciel R avec correction de Bonferroni-Holm
Donne des valeurs de p corrigées, c'est-à-dire multipliées par $k + 1 - i$, en fixant à 1 les valeurs > 1.

```
        Pairwise comparisons using t tests with pooled SD

data:  d$duree and d$taille

  XL    L M
L 6e-06 - -
M 1e-04 1 -
S 5e-05 1 1

P value adjustment method: holm
```

Dans ce cas les conclusions sont équivalentes à celles de Bonferroni.

Autres méthodes de comparaisons multiples

- **Comparaisons 2 à 2** : $k = \frac{p(p-1)}{2}$ comparaisons
  de très nombreuses autres méthodes disponibles (Tukey, Duncan, Rodger, Scheffé, Dunn-Sidak, ...), avec prédominance actuelle de la méthode **Bonferroni-Holm**.
- **Comparaisons à un groupe témoin** : $k = p - 1$ comparaisons
  méthode paramétrique de **Dunnett** couramment employée : statistique de Student avec estimation d'un σ global et correction des valeurs de p adaptée à ce cas particulier.
- **Comparaison multiples générales avec maîtrise du taux de fausses découvertes** : méthode de **Benjamini-Hochberg** (méthode couramment utilisée en transcriptomique - analyse de l'expression d'un très grand nombre de gènes).

Utilisation des méthodes de comparaisons multiples

- Il est indispensable de **vérifier que la différence globale est significative avant** de faire des comparaisons multiples (origine du terme ""protected" dans PLSD de Fisher).
- Il faut **corriger le risque** α lors de la réalisation de comparaisons multiples **si l'on souhaite limiter le nombre de faux positifs** (rejets à tort de H_0).
- Les comparaisons multiples suite à une comparaison globale ne sont **pas à préconiser systématiquement** : elles n'apportent parfois pas grand chose à l'analyse globale et sont souvent difficiles à interpréter.
- **NE JAMAIS OUBLIER** qu'une différence non significative ne permet pas de conclure à une non différence.
- **TOUJOURS PENSER à interpréter les effets (différences entre groupes)** : ne pas rester au niveau des valeurs de p.

Récapitulatif sur les tests de comparaison de moyennes

- **Un seul échantillon**
  test de conformité de Student ou test de la médiane
- **Deux échantillons indépendants**
  test de Student avec variances égales ou non (test de Welch) ou test de la somme des rangs de Mann-Whitney-Wilcoxon
- **Deux échantillons dépendants (appariés)**
  test de Student des séries appariés ou test des rangs signés de Wilcoxon
- **Plusieurs échantillons indépendants**
  ANOVA ou test de la somme des rangs de Kruskal-Wallis

Il est capital de bien **examiner visuellement la ou les distributions observées** afin de choisir entre un test **paramétrique** et un **test non paramétrique**.