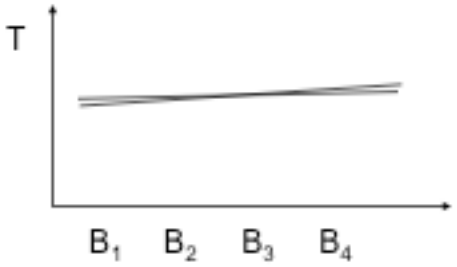
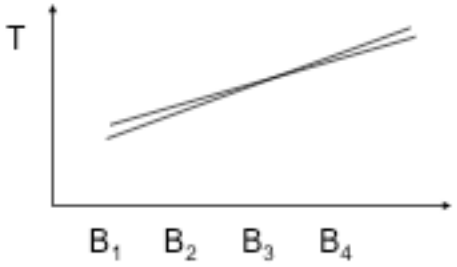
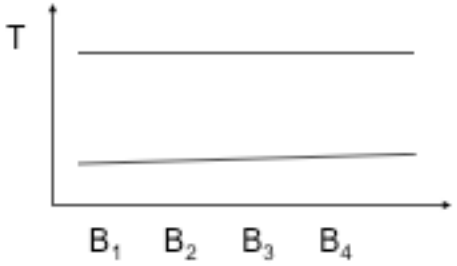


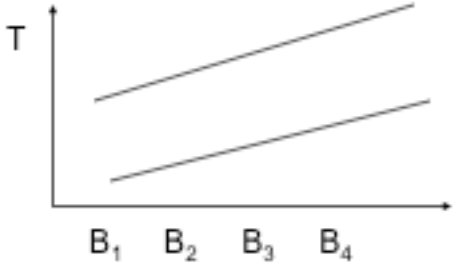
Pas d'effet significatif



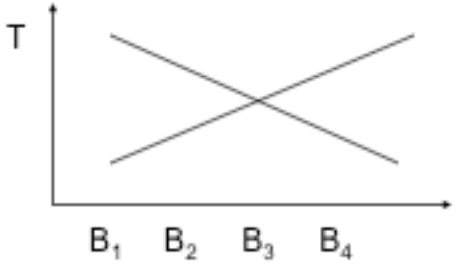
Effet significatif de Bloc. Pas d'autre effet



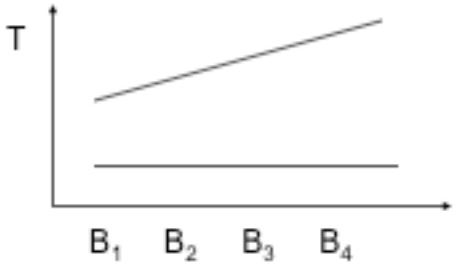
Effet significatif du périphérique. Pas d'autre effet



Effet significatif des périphériques et des blocs. Pas d'interaction



Interaction significative. Pas d'autre effet



Effet significatif des périphériques. Interaction

NIHM - ÉVALUATION

Géry Casiez http://cristal.univ-lille1.fr/~casiez
RVI Master 2 spécialité IVI – Université de Lille 1

Avez-vous déjà entendu parlé de …

- expériences contrôlées
- plan expérimental
- variables dépendantes / indépendantes
- variables continues, discrètes, nominales…
- intervalle de confiance
- test paramétrique / non paramétrique
- ANOVA
- p-value
- interaction entre des facteurs

Validation expérimentale

- Démontrer que votre proposition est meilleure que d'autres travaux de l'état de l'art suivant certains critères dans certaines conditions

Question de recherche

- Sur un téléphone portable, est-il plus rapide de rentrer du texte avec un clavier physique, un écran tactile avec les doigts ou un écran tactile avec un stylet?
- Quelles sont les hypothèses de recherche?
- Quelles sont les variables?
- Quel est le plan expérimental?
- Qu'est-ce qu'on va mesurer?
- => réalisation d'une expérience contrôlée

Expériences contrôlées

- Choix d'un petit nombre de facteurs dont on veut mesurer les effets
- Choix de mesures
- Construction d'un plan expérimental
- Choix d'un ou plusieurs groupes d'utilisateurs

Un peu de vocabulaire…

Participants

- Les gens qui participent à une expérience sont appelés “participants”
- Eviter d’utiliser le terme “sujet”
- Utiliser le terme participant pour toute référence explicite à l’expérience (ex:, “tous les participants ont obtenu un taux d’erreur important…”)
- Les commentaires généraux ou les conclusions peuvent utiliser d’autres termes: “ ces résultats suggèrent que les utilisateurs ont moins de chances de …”

Variable indépendante

- Une variable indépendante est une variable qui est manipulée à travers la conception de l'expérience
- Ex: périphérique, type de retour, apparence d'un bouton, mise en page, sexe, âge, expertise, etc.
- Les termes variable indépendante et facteur sont synonymes
- "Indépendant" signifie "indépendant des participants"

Condition de test (niveaux)

- Les valeurs prises par une variable indépendante sont les conditions de test (niveaux)
- Donner des noms à la fois aux variables indépendantes (facteurs) et aux conditions de test (niveaux)
- Ex:

Facteurs	**Niveaux**
Périphérique	Souris, trackball, joystick
Type de retour	audio, tactile, retour de force
Tâche	Pointage, dragging
Visualisation	2D, 3D, animée

Variable dépendante

- Une variable dépendante est une variable représentant les mesures ou observations d'une variable indépendante
- Ex: temps de réalisation, vitesse, précision, taux d'erreur, nombre de touches appuyées, vitesse d'apprentissage, etc.
- Donner un nom pour la variable dépendante avec les unités
- "Dépendant" signifie "dépendant des participants"
- Exemples:
  - Temps de réalisation (ms), vitesse (mots par minute, nb de sélections par minute, etc), taux d'erreur (%) ...

Variable de contrôle

- Conditions ou facteurs qui (a) peuvent influencer une variable dépendante, mais (b) qui ne sont pas étudiés et dont on peut s'accommoder d'une certaine façon
- Une façon de les contrôler – est de les traiter comme des variables de contrôle
- Une variable de contrôle est gardée constante d'un test à l'autre
- Ex: éclairage d'une pièce, bruit de fond, température
- L'inconvénient est d'avoir trop de variables de contrôle qui rendent l'expérience moins généralisable (cad., applicable à d'autres situations)

Variable de contrôle

- Nombre moyen d'images par seconde
- Latence moyenne
- Retard du réseau
- Distorsion optique
- ...

Variable aléatoire

- Au lieu de contrôler tous les facteurs, certains peuvent varier de manière aléatoire
- De tels facteurs sont des variables aléatoires
- Plus de variabilité est introduite dans les mesures (---), mais les résultats sont plus généralisables (+++)

Variable de confusion

- Une variable qui varie systématiquement avec une variable indépendante est une variable de confusion
- Ex: Si trois périphériques sont toujours testés dans le même ordre, la performance des participants peut s'améliorer avec l'entraînement; ex., de la 1ère à la 2nde condition, et de la 2nde à la 3e condition; par conséquent "l'apprentissage" est une variable de confusion (parce qu'elle varie systématiquement avec le "périphérique")

Plan expérimental

Intra-sujets, Inter-sujets

- L'administration des niveaux d'un facteur est soit intra ou inter-sujets
- Si chaque participant est testé sur chacun des niveaux, le facteur est dit intra-sujets (within-subject)
- Si chaque participant est testé sur seulement un niveau, le facteur est dit inter-sujets (between subject). Dans ce cas, des groupes séparés de participants sont utilisés dans chaque conditions.
- Les termes "mesures répétées" (repeated-measures) et "intra-sujets" sont synonymes.

Intra vs. inter Sujets

- Question: Lors de la conception d'une expérience, vaut-il mieux utiliser des facteurs intra-sujets ou inter-sujets?
- Réponse: Ca dépend!
- Discussion:
  - Parfois un facteur doit être inter-sujets (e.g., sexe, age)
  - Parfois un facteur doit être intra-sujets (e.g., session, bloc)
  - Parfois on a le choix. Dans ce cas, il faut faire un compromis
  - Avantage intra-sujets: la variance due aux pré-dispositions des participants est normalement la même dans toutes les conditions (cf. inter-sujets)
  - Avantage inter-sujets: évite les phénomènes d'interférences (ex: utiliser deux claviers avec une organisation différente des touches)

Plan d'expérience

- Le plan d'expérience fait référence à l'organisation des facteurs, niveaux, procédures … dans une expérience
- Exemple:
  - "Plan 3 x 2 intra-sujets" correspond à une expérience avec deux facteurs, ayant 3 niveaux dans le premier, et 2 niveaux dans le second. Il y 6 conditions de test au total. Chacun des facteurs est intra-sujets signifiant que tous les participants testent toutes les conditions
- Note: Une conception mixte est aussi possible
  - Dans ce cas, les niveaux d'un facteur sont administrés à tous les participants (intra-sujets) alors que les niveaux d'un autre facteur sont administrés à des groupes différents (inter-sujets).

Contre balancement

- Pour une conception intra-sujets, la performance des participants peut s'améliorer avec l'entraînement d'une condition de test à une autre.
- Pour compenser, l'ordre de présentation des conditions est contre-balancé.
- Les participants sont divisés en groupes, et un ordre différent est administré à chacun des groupes
- L'ordre suit un carré latin

Carré latin

- La caractéristique définissant un carré latin (Latin square) est que chaque condition apparaît seulement une fois dans chaque ligne et colonne.
- Ex:

Carré latin 3 X 3

A	B	C
B	C	A
C	A	B

Carré latin 4 X 4

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

Carré latin 4 X 4 balancé

A	B	C	D
B	D	A	C
D	C	B	A
C	A	D	B

Note: Dans un carré latin balancé chaque condition précède et suit chaque autre condition un nombre égal de fois

Types de variables

Variables catégorielles

- Mesures qui ne représentent pas une quantité
- Variables nominales
  - plantes, champignons, poissons, vertébrés
  - homme/femme (variable binaire)
- Variables ordinales
  - catégories ordonnées
  - exemple: échelle de Likert
    -2 (très mauvais) à +2 (excellent), en passant par zéro (indifférent)

Variables continues

- Intervalles
  - une variable est un intervalle quand des intervalles identiques sur l'échelle représentent des différences identiques de la propriété mesurée.
  - exemple: noter un produit avec une note de 1 à 5
- Ratios
  - En plus: les ratios ont du sens
  - exemple: temps.

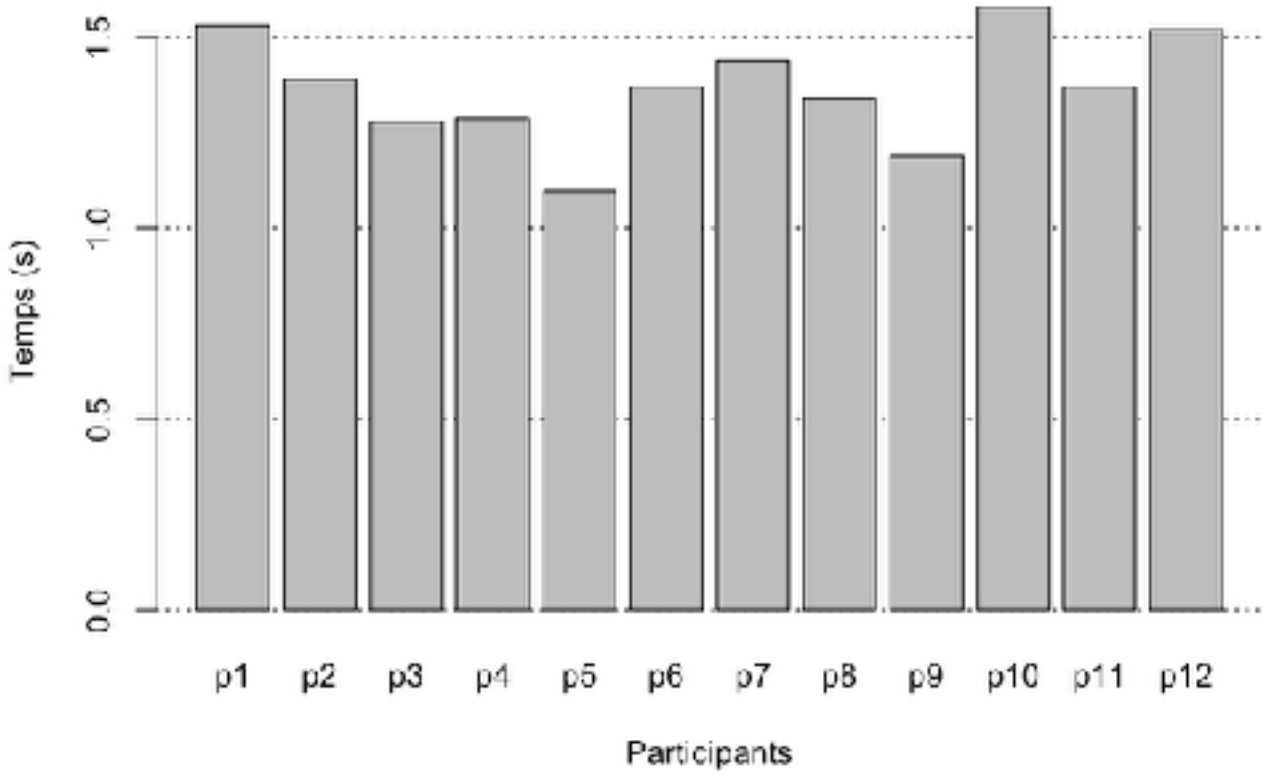
Statistiques

Analyse statistique

- Statistiques descriptives
- Exemple
  - La performance de 12 participants a été mesurée pour 2 périphériques (souris et tablette)
  - Plusieurs mesures de performance ont été réalisées
  - Chaque mesure de performance est une variable dépendante
  - Une des variables est le temps de réalisation (T)
  - La variable indépendante est le périphérique

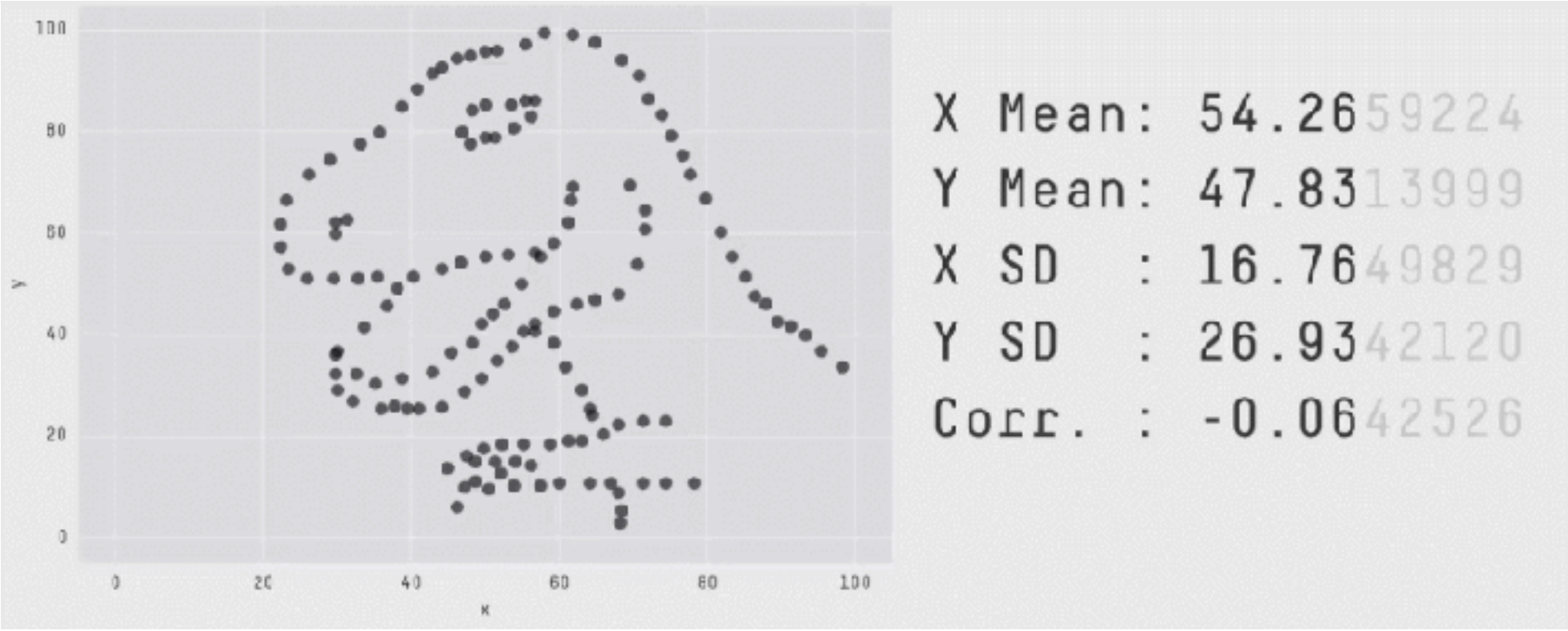
Statistiques Descriptives

participant	souris
p1	1.53
p2	1.39
p3	1.28
p4	1.29
p5	1.10
p6	1.37
p7	1.44
p8	1.34
p9	1.19
p10	1.58
p11	1.37
p12	1.52



moyenne: 1.37 s
écart-type: 0.14 s

Toujours visualiser des données!



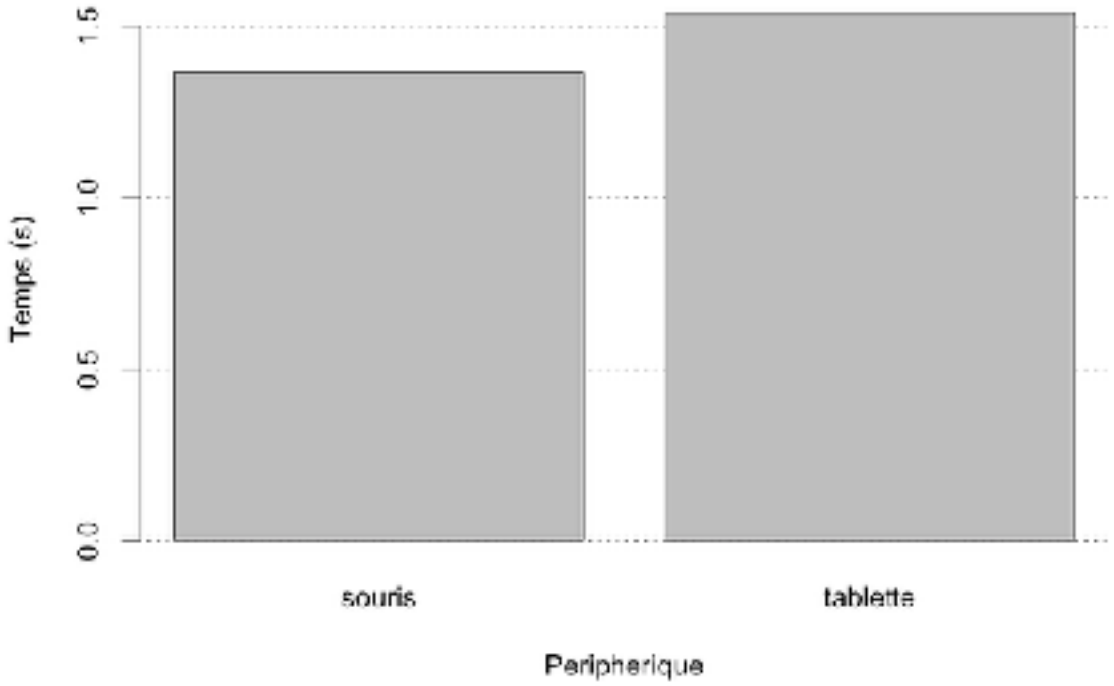
Justin Matejka and George Fitzmaurice. 2017. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17). ACM, New York, NY, USA, 1290-1294. DOI: https://doi.org/10.1145/3025453.3025912

Evaluations comparatives

- Le résultat précédent, seul, n'est pas très intéressant
- L'objectif est souvent de comparer une ou plusieurs conditions
- Les conditions sont les niveaux de la variable indépendante
- Dans l'exemple, la variable indépendante est "Périphérique" et les niveaux sont "Souris" vs. "Tablette"

Evaluations comparatives

participant	souris	tablette
p1	1.53	1.35
p2	1.39	1.74
p3	1.28	1.81
p4	1.29	1.16
p5	1.10	1.43
p6	1.37	1.30
p7	1.44	1.46
p8	1.34	1.73
p9	1.19	1.81
p10	1.58	1.68
p11	1.37	1.66
p12	1.52	1.36



Les études comparatives sont plus intéressantes mais est-ce que ce résultat est plus intéressant?

Hypothèse nulle

- Déclarer comme “hypothèse statistiquement nulle" quelque chose qui est logiquement l’opposé de ce que l’on croit.
- Appeler cette hypothèse *H0*
- Montrer à partir des données que *H0* est fausse, et doit être rejetée
- En rejetant *H0*, on confirme ce en quoi on croit

Hypothèse nulle

- L'hypothèse nulle est rejetée ou non

		Etat du monde	
		H0	H1
Décision	H0	Acceptation correcte	Erreur de type II β
	H1	Erreur de type I α	Rejet correct

ANOVA

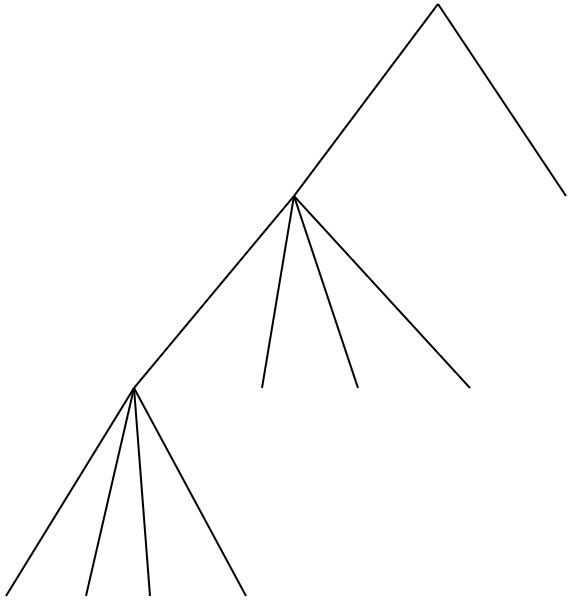
- C'est le principal outil statistique utilisé dans le domaine de l'interaction homme-machine pour évaluer des expériences
- Utilisé pour répondre à des questions du type "Est-ce que le temps pour accomplir telle tâche varie différemment suivant le type de technique d'interaction utilisé? "

Plan expérimental

Facteur 1 (ex: Périphérique): 2 niveaux (e.g. souris and tablette)

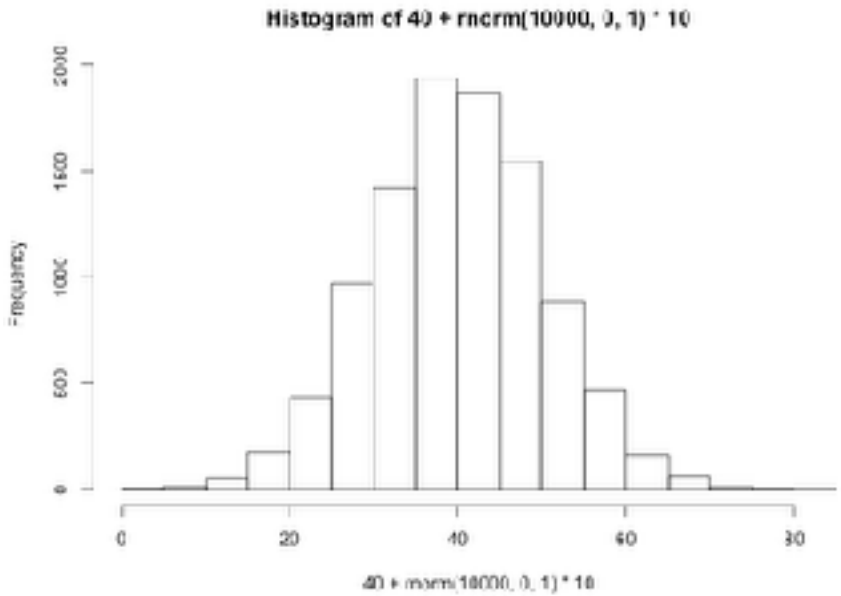
Facteur 2 (ex: Bloc): 4 niveaux

Facteur 3 (ex: ID): 4 niveaux



Conditions d'utilisation

- Indépendance des données
- Distributions normales
- Homogénéité des variances
  (condition d'homoscédasticité, Homoscedasticity)

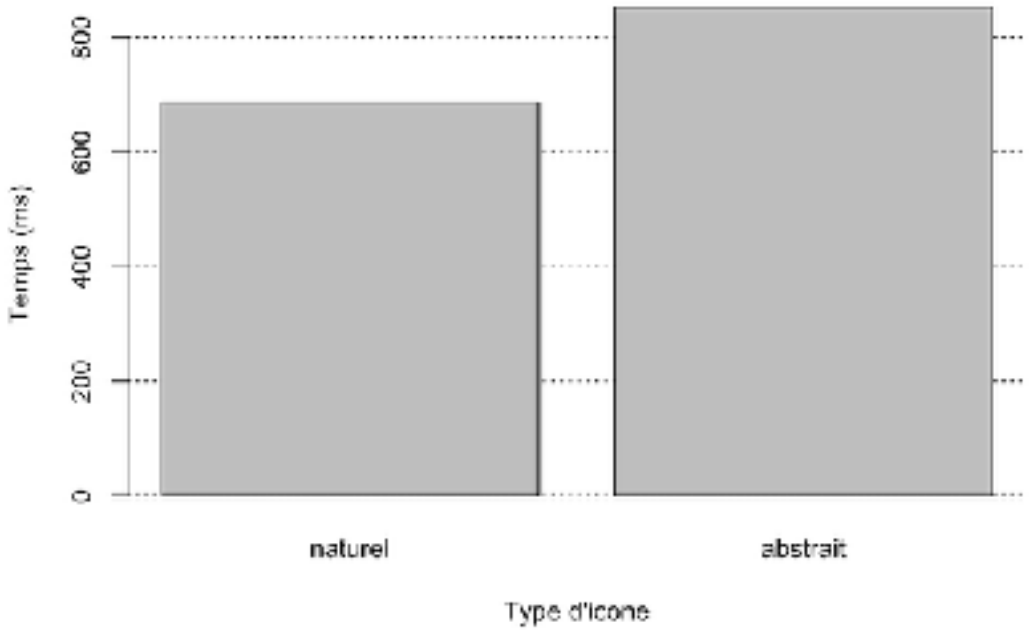


ANOVA avec un facteur intra-sujets

- Exemple
  - Apparence d'une icône avec 2 niveaux: naturel et abstrait
  - Mesure du temps de sélection (ms)
  - 10 participants

ANOVA avec un facteur intra-sujet

participant	naturel	abstrait
p1	741	936
p2	727	890
p3	671	838
p4	681	797
p5	693	839
p6	680	822
p7	721	854
p8	613	867
p9	637	845
p10	697	844



ANOVA avec un facteur intra-sujet

Test de la normalité des données

```
shapiro.test(data[,"naturel"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[, "naturel"]
## W = 0.95859, p-value = 0.7697
```

```
shapiro.test(data[,"abstrait"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[, "abstrait"]
## W = 0.92602, p-value = 0.4099
```

p>0.05 donc données considérées comme suivant une distribution normale

Test d'homogénéité des variances

```
library(car)
d3$type = factor(d3$type)
kable(leveneTest(temps ~ type, data=d3))
```

	Df	F value	Pr(>F)
group	1	0.1420698	0.7106387
	16	NA	NA

p>0.05 donc les variances des 2 groupes ne sont pas statistiquement différentes.

ANOVA avec un facteur intra-sujet

	Effect	DFn	DFd	F	p	p<.05	ges
2	type	1	9	164.5409	4e-07	*	0.8367692

Effet significatif $F_{1,9}$ = 164.5 p < 0.001 de l'apparence de l'icône sur le temps d'acquisition.

Nombre de degrés de liberté (nb de niveaux – 1),
(nb de niveaux – 1) x (nb de participants – 1)

Si p < 0.05, il y a 95% de chances que la différence observée n'est pas due au hasard

ANOVA avec un facteur intra-sujet

- Retour à l'exemple des périphériques

	Effect	DFn	DFd	F	p	p<.05	ges
2	peripherique	1	11	4.660843	0.0538073		0.1926046

$F_{1,11} = 4.66$ p=0.054

p-value > 0.05 pas d'effet significatif sur le temps. Peut-on conclure qu'il n'y a pas de différence entre les deux périphériques?

Taille de l'effet (effect size)

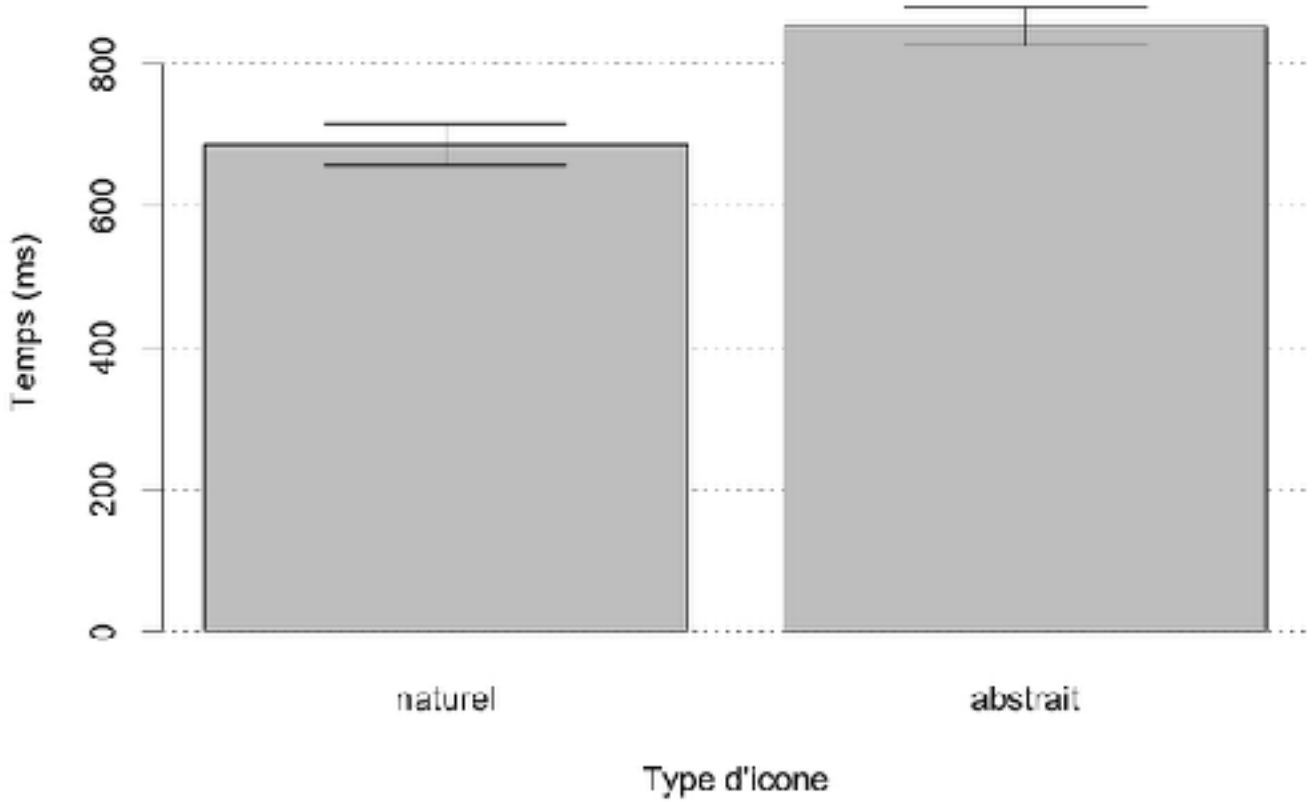
- Importance de l'effet observé et interprétation

	Small	Medium	Large
Cases: → / **Measures ↓**	Difference between the heights of 15 and 16 year old girls in the USA.[1]	M vs. F verbal fluency[2]	Difference between the heights of 13 and 18 year old girls in the USA.[3]
Cohen's d	0.2 - 0.3	≈ 0.5	0.8 - or larger
r (correlation coeff.) [7]	0.1 - 0.3	0.3 - 0.5	0.5 or larger
R^2: % of variance explained	1%	9%	25%
R^2: (≈≈ Partial eta-squared)	0.02	0.13	0.26
Partial eta-squared (η^2)	0.01	0.06	0.14
Generalized eta-squared η^2_G [8]	0.02	0.13	0.26

http://www.psy.gla.ac.uk/~steve/best/effect.html

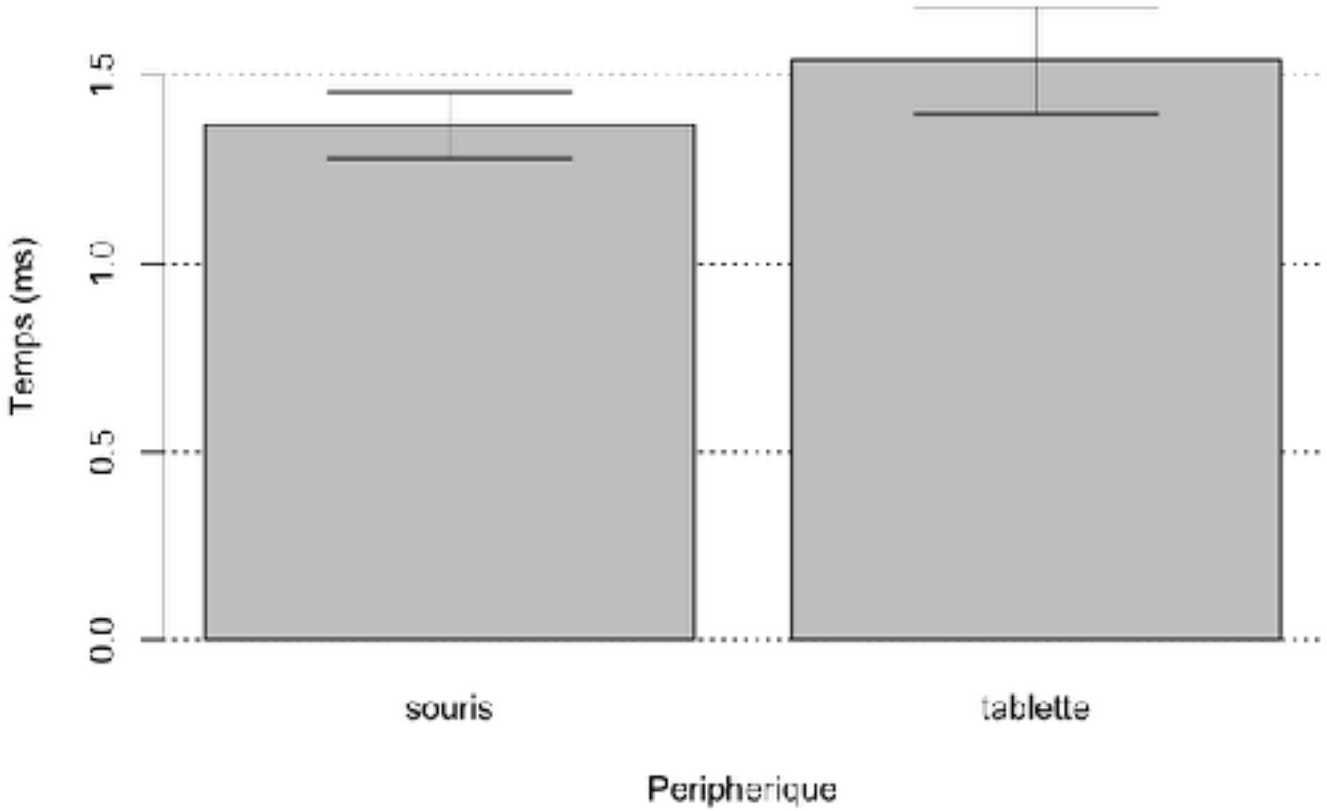
Intervalles de confiance à 95%

- Cas des icônes



Intervalles de confiance à 95%

- Cas des périphériques



95% CI

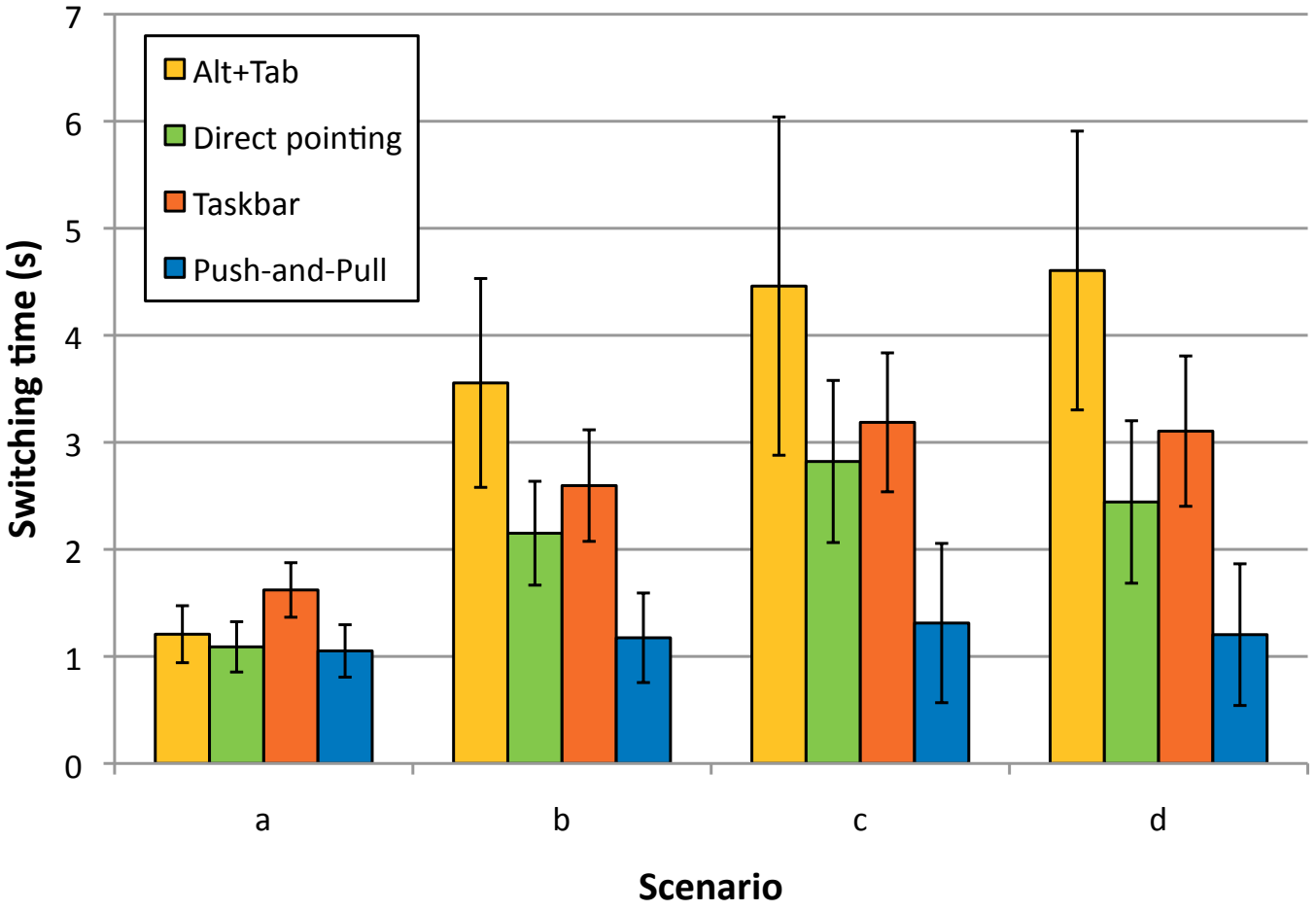


Figure 4. Mean switching time for SWITCHING TECHNIQUE and SCENARIO. Error bars represent 95% confidence interval.

ANOVA avec un facteur inter-sujets

- Exemple:
  - Tester si une interface ou une technique d'interaction fonctionne mieux avec les gaucher ou les droitiers (ou hommes vs femmes)
  - 2 groupes de participants sont nécessaires: 5 gauchers (G) et 5 droitiers (D)
  - La variable indépendante est la latéralité avec 2 niveaux, Gauche et Droite
  - La variable dépendante est le temps (secondes) pour accomplir la tâche.

Conception intra vs. inter sujets

- Intra-sujets: 2 fois plus puissant avec 2 fois moins de participants (si 2 niveaux)…
- … mais demande 2 fois plus de temps
- Quand c'est possible, la conception intra-sujets est préférée pour les groupes de petite taille

ANOVA avec un facteur intra-sujets et un facteur inter-sujets

- Exemple: Contrôler si le contre-balancement annule les effets d'apprentissage
- Retour sur l'exemple des icônes abstraites et concrètes
- L'ordre a été contrebalancé entre les sujets

ANOVA avec un facteur intra-sujets et un facteur inter-sujets

	Effect	DFn	DFd	F	p	p<.05	ges
2	groupe	1	8	0.5368515	0.4846522		0.0453139
3	type	1	8	146.9600526	0.0000020	*	0.8431817
4	groupe:type	1	8	0.0383684	0.8495892		0.0014018

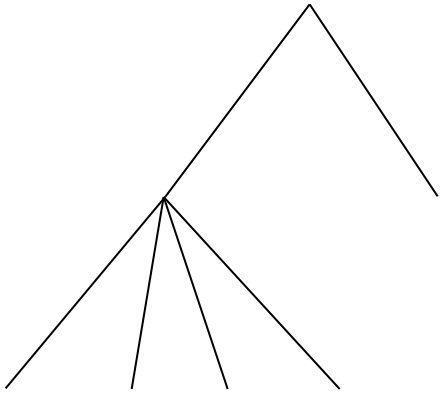
- Pas d'effet de groupe significatif ($F_{1,8} = 0.48$, ns)
- Pas d'interaction significative Type d'icône x Groupe ($F1,8 = 0.04$, ns)
  - → pas de transfert d'apprentissage asymétrique

ANOVA avec deux facteurs intra-sujets

- Exemple: 2 facteurs
  - Périphérique P1, P2
  - Bloc B1, B2, B3, B4

Facteur 1 (Périphérique): 2 niveaux

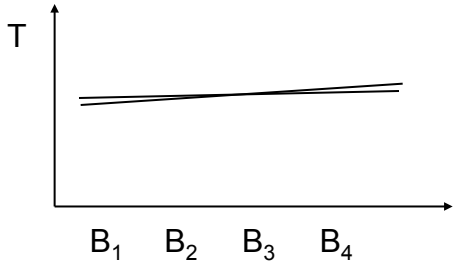
Facteur 2 (Bloc): 4 niveaux



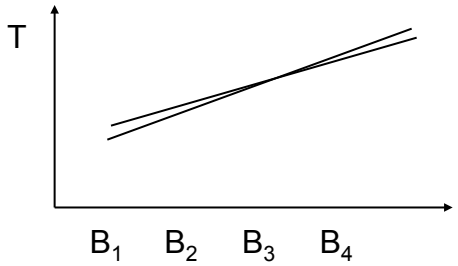
ANOVA avec deux facteurs intra-sujets

- Effet significatif principal: Périphérique ou/et Bloc
  - e.g. Effet significatif entre P1 et P2
  - On peut trouver une différence significative entre les blocs mais sans pouvoir conclure sur l'effet d'apprentissage
  - Pour connaître l'histoire complète: Etudier l'interaction Périphérique x Bloc

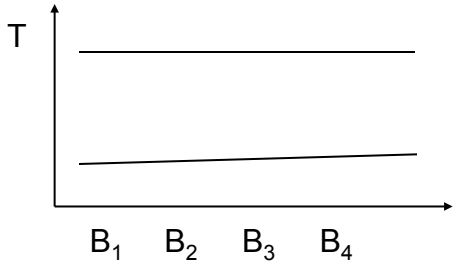
Interaction



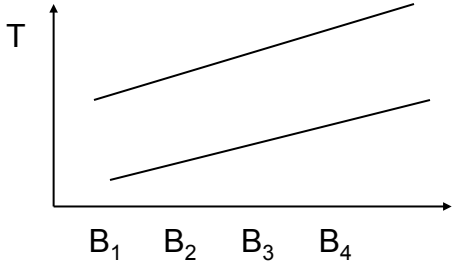
Pas d'effet significatif



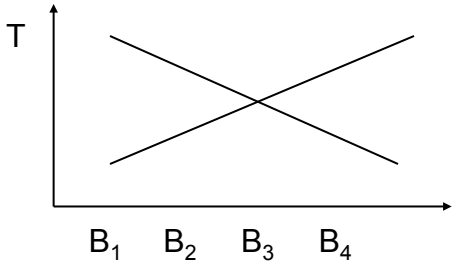
Effet significatif de Bloc. Pas d'autre effet



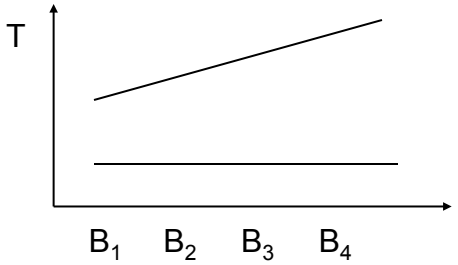
Effet significatif du périphérique. Pas d'autre effet



Effet significatif des périphériques et des blocs. Pas d'interaction



Interaction significative. Pas d'autre effet



Effet significatif des périphériques. Interaction

Test de sphéricité


```
kable(anova$`Mauchly's Test for Sphericity`)
```

	Effect	W	p	p<.05
4	block	0.9095387	0.6462531	
5	distance	0.2812080	0.0001391	*
7	task:block	0.8117481	0.2322464	
8	technique:block	0.6412757	0.0445978	*
9	task:distance	0.7087711	0.0898549	
10	technique:distance	0.7522581	0.1363227	
11	block:distance	0.4054775	0.2134726	
12	task:technique:block	0.9161025	0.5415126	
13	task:technique:distance	0.2067270	0.0000161	*
14	task:block:distance	0.4338710	0.2668584	
15	technique:block:distance	0.4575308	0.3396913	
16	task:technique:block:distance	0.3352582	0.1039435	

Test de sphéricité

A list containing one or more of the following components:

`ANOVA` A data frame containing the ANOVA results.

`Mauchly's Test for Sphericity`
If any within-Ss variables with >2 levels are present, a data frame containing the results of Mauchly's test for Sphericity. Only reported for effects >2 levels because sphericity necessarily holds for effects with only 2 levels.

`Sphericity Corrections`
If any within-Ss variables are present, a data frame containing the Greenhouse-Geisser and Huynh-Feldt epsilon values, and corresponding corrected p-values.

https://cran.r-project.org/web/packages/ez/ez.pdf

ANOVA avec deux facteurs intra-sujets et un facteur inter-sujets

- Même conception que précédemment
- Facteur inter-sujets: contre-balancement des périphériques

Aussi…

- 4 facteurs…
- … n facteurs

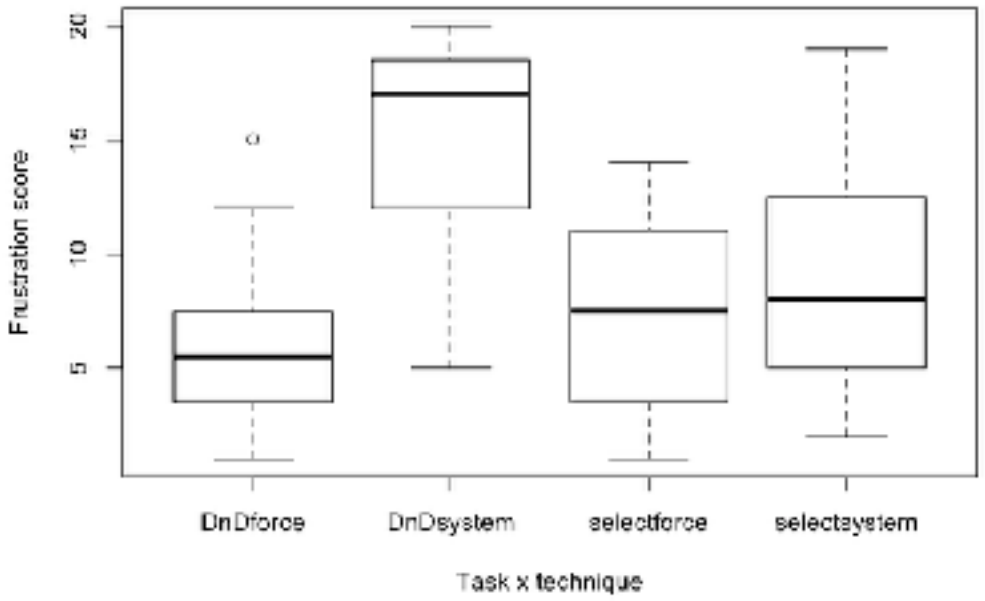
Tests non paramétriques

Quand les utiliser?

- Données de type ordinal : échelle de Likert
- Les hypothèses de l'ANOVA ne sont pas respectées
  - d'autres solutions existent (transformation log, ART…)
- Limite: pas d'analyse des interactions possible
- Souvent utilisé en IHM: Friedman (within subject)

Exemple: analyse de NASA-TLX

participant	DnDforce	DnDsystem	selectforce	selectsystem
0	4	20	5	13
1	6	9	4	11
2	1	17	2	2
3	3	13	3	2
4	8	20	11	6
5	12	20	9	12
6	3	17	8	19
7	15	17	12	15
8	4	19	14	17
9	7	18	13	7
10	6	15	2	2
11	5	11	8	6
12	7	7	11	5
13	5	5	4	5
14	1	17	1	10
15	9	14	7	8



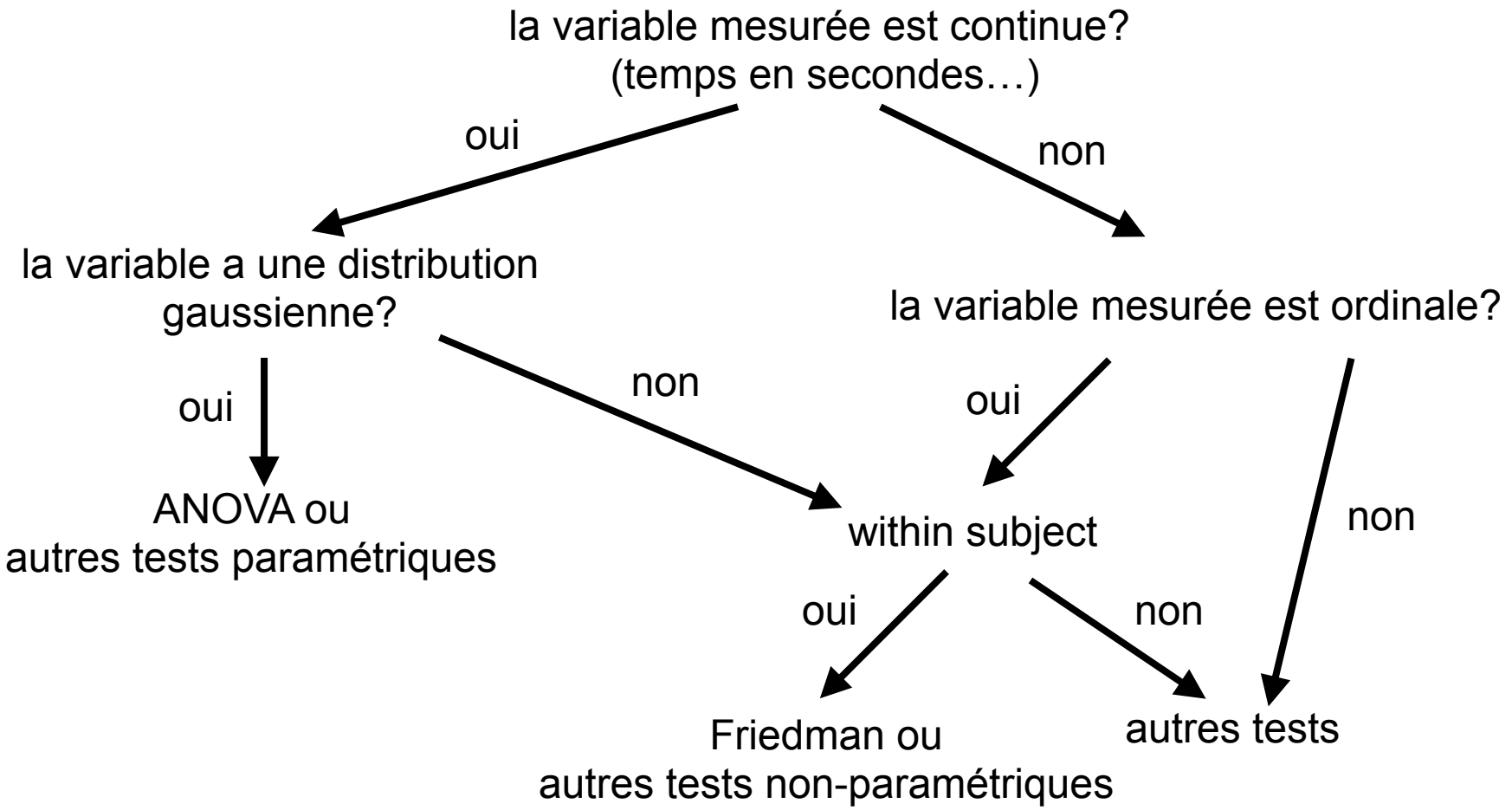
Exemple: analyse de NASA-TLX


```
 Friedman rank sum test

data:  data.matrix(data_tr)
Friedman chi-squared = 21.939, df = 3, p-value = 6.717e-05

 Pairwise comparisons using Wilcoxon rank sum test

data:  mdata$score and mdata$tasktech

             DnDforce DnDsystem selectforce
DnDsystem    0.00034  -         -
selectforce  1.00000  0.00128   -
selectsystem 0.70051  0.02153   1.00000

P value adjustment method: bonferroni
```

Coder l'expérience

- Toujours enregistrer les données brutes!
  - Evite les erreurs
  - Evite les oublis

Conclusion


la variable mesurée est continue?
(temps en secondes…)
oui
non
la variable a une distribution gaussienne?
la variable mesurée est ordinale?
oui
non
oui
non
ANOVA ou
autres tests paramétriques
within subject
oui
non
non
Friedman ou
autres tests non-paramétriques
autres tests