

Atelier d'intégration en méthodes quantitatives :

Comment améliorer l'impact de vos
recherches par les intervalles de confiance.

Dominic Beaulieu-Prévost
Centre de Recherche Fernand Seguin

Retour sur les tests de signification statistique (TSS)

Depuis plus de 50 ans, les tests de signification statistique (TSS) ont été considérés comme LE standard pour l'analyse et la présentation des résultats scientifiques et pour décider de la valeur scientifique des hypothèses en sciences sociales quantitatives.

Depuis ce temps, plus de 300 articles ont démontré que cette approche était extrêmement problématique et inadéquate pour juger de l'importance pratique ou de la pertinence d'un résultat de recherche.

Mais quel est le problème avec les TSS?

1- Les TSS sont TRÈS difficiles à interpréter.

Les interprétations fautives des TSS est très courante chez les chercheurs (Lecoutre, Poitevineau & Lecoutre, 2003).

Une enquête chez des chercheurs en psychologie a démontré que seulement 11% d'entre eux étaient capables d'interpréter adéquatement un TSS (Oakes, 1986).

Ex: Étude de traitement pour la dépression

Mesure d'efficacité: Beck Depression Inventory (BDI)

Traitement expérimental: Réduction de 8 pts (ÉT = 6)

Traitement placebo: Réduction de 4 pts (ÉT = 5)

Il y a 20 participants par condition et l'alpha est de 0.05.

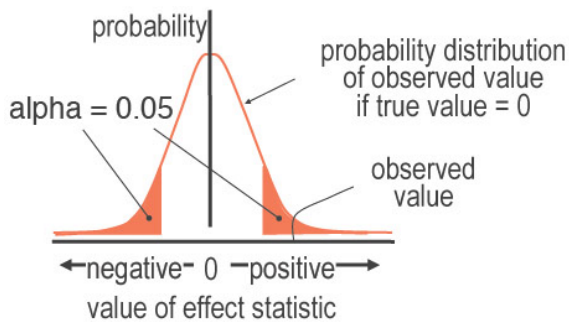
Résultats classiques: Test T pour groupes indépendants

$t(38) = 2.29, p = 0.028$ (ou) $p < 0.05$.

Le résultat est statistiquement significatif. Que peut-on inférer?

- Il y a moins de 5% de chances que l'hypothèse nulle soit vraie.
- H1 est probablement vrai (plus de 95% des chances).
- Le résultat n'est probablement pas du au hasard (moins de 5% des chances).
- La probabilité d'obtenir des résultats aussi extrêmes (c.-à-d. 4 pts de différence) est de moins de 5% EN SUPPOSANT que H0 est VRAI.

Le p est donc simplement un indicateur de surprise "conditionnelle".



Mais quel est le problème avec les TSS?

2- Le p dépend de la taille d'échantillon.

Un résultat statistiquement significatif sera TOUJOURS obtenu si n est assez grand, sauf si la taille d'effet est EXACTEMENT zéro.

3- L'hypothèse nulle n'est pas plausible.

Dans les sciences du vivant (i.e. de la biologie à la sociologie), presque toutes les variables sont corrélées entre elles jusqu'à un certain point (Meehl, 1990).

4- L'hypothèse nulle est toujours fausse.

Comme H_0 est représenté par un point sur une échelle continue, il est logiquement peu probable (1/!) qu'elle soit vraie.

5- L'hypothèse alternative est non-falsifiable

Comme H_0 est toujours fausse, H_1 est toujours vraie. De plus, si l'on n'a pas réussi à rejeter H_0 , on peut toujours déclarer que la taille d'échantillon était trop petite. H_1 est donc non-falsifiable.

Retour sur les tests de signification statistique (TSS)

En 1999, le "Statistical task force" de l'APA a failli recommander de bannir les TSS de leurs revues (Wilkinson et al., 1999).

En 2009, la nouvelle édition du Guide de publication de l'APA a finalement pris position: Les résultats statistiques devront maintenant inclure des indicateurs de taille d'effet et d'intervalle de confiance.

Objectif de l'atelier

S'attarder aux solutions et non aux problèmes.

Pour mieux comprendre les critiques des TSS, voir les références fournies en fin d'atelier.

Les deux utilités des ICs

1. *L'estimation de paramètres*

2. *Les tests d'hypothèse*

L'estimation de paramètres

Estimation de paramètres

L'intervalle de confiance, un outil de base

Le modèle de base: $IC = M \pm Vc * ET$

M= Taille de l'effet dans l'échantillon

Vc= Valeur critique pour le niveau de confiance précisé

ET= Erreur-type

Au lieu de simplement spécifier si la taille de l'effet est différente de zéro, les ICs nous donnent accès à :

- la taille de l'effet
- la précision de l'estimé

Table 2
Mean Systematic Error for the whole sample and for the three sub-samples of Attitude Towards Dreams

Subgroup	<i>n</i>	<i>M</i> (attitude)	<i>M</i> (error)	CI (95%)	<i>p</i>
All participants	82	4.94	−0.58	−1.06 to −0.10	0.020
Low Attitude	27	3.57	−1.40	−2.17 to −0.62	0.001
Average Attitude	27	5.09	−0.64	−1.41 to 0.14	0.105
High Attitude	28	6.12	0.26	−0.67 to 1.19	0.568

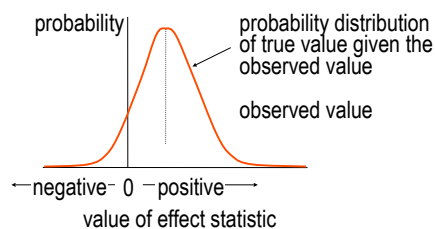
Note. CI = confidence interval for the Systematic Error.

Estimation de paramètres

L'intervalle de confiance, un outil de base

Il est possible de calculer des ICs pour des:

- moyennes
- corrélations
- proportions
- différences



Estimation de paramètres

L'intervalle de confiance, un outil de base

Il est possible de calculer des ICs pour des:

- moyennes
- corrélations
- proportions
- différences

En fait, les ICs sont mathématiquement équivalents aux tests statistiques. Il est donc possible de les calculer pour tous les test statistiques existants!

Comment interpréter un IC

Selon une interprétation fréquentiste:

- "L'intervalle des valeurs considérées équivalentes (étant donné l'erreur échantillonnale) avec un niveau de confiance de 95%.
- "Une estimation de la moyenne populationnelle avec un niveau de confiance de 95%.

Si on pouvait calculer des ICs pour une infinité d'échantillons aléatoires provenant de la même population, la moyenne de la population serait incluse dans 95% d'entre eux.

Un niveau de confiance de 95% représente donc une *fréquence relative à long-terme* et non l'idée qu'il y a 95% de chances que la moyenne populationnelle soit incluse dans l'intervalle.

Comment interpréter un IC

Selon une approche bayésienne:

Il est possible de calculer un IC dans lequel il y a 95% de chances que la moyenne populationnelle soit incluse.

- Cela nécessite de calculer un intervalle de confiance à priori représentant les connaissances accumulées en plus des données de l'expérience. C'est un principe similaire à une méta-analyse.
- Mais si on postule un manque de connaissance à priori, l'IC bayésien coïncide avec son homologue fréquentiste. C'est ce que l'on appelle un *à priori agnostique* ou non-informatif.

Comment interpréter un IC

En résumé:

Les ICs traditionnels représentent donc des intervalles qui ont 95% de chances d'inclure la moyenne populationnelle à la condition que l'on postule un *à priori agnostique*.

Par extension, on peut considérer la distribution reliée à l'IC comme la distribution des valeurs probables.

Table 2
Mean Systematic Error for the whole sample and for the three sub-samples of Attitude Towards Dreams

Subgroup	<i>n</i>	<i>M</i> (attitude)	<i>M</i> (error)	CI (95%)	<i>p</i>
All participants	82	4.94	−0.58	−1.06 to −0.10	0.020
Low Attitude	27	3.57	−1.40	−2.17 to −0.62	0.001
Average Attitude	27	5.09	−0.64	−1.41 to 0.14	0.105
High Attitude	28	6.12	0.26	−0.67 to 1.19	0.568

Note. CI = confidence interval for the Systematic Error.

Les tests d'hypothèse

La logique des tests d'hypothèse

Hypothèses ponctuelles

et

Hypothèses par intervalle

Les tests d'hypothèse

Plusieurs tests d'hypothèse peuvent être effectués *sans calcul additionnel* à partir d'un IC.

Le principe est le même pour tous les types d'ICs. Il suffit de définir les hypothèses qui nous intéressent.

Les principales hypothèses d'intérêt

- 1) L'hypothèse nulle
- 2) L'hypothèse d'un effet substantiel
- 3) L'hypothèse d'un effet néfaste
- 4) L'hypothèse d'un effet non-pertinent

Notion d'effet substantiel

C'est un effet dont la taille est assez grande pour être intéressante.

C'est l'équivalent d'un effet *cliniquement significatif* mais ça ne se limite pas aux contextes cliniques.

La valeur minimale d'un effet substantiel est définie en fonction de#

- L'importance théorique de l'effet
- L'utilité pratique du phénomène
- Le coût potentiel de l'intervention
- La sensibilité de l'échelle de mesure

Types d'hypothèse

L'Hypothèse d'un effet substantiel évalue si l'effet est au moins égal à l'*effet substantiel minimal*.

L'Hypothèse d'un effet néfaste est défini comme étant l'opposé de l'hypothèse d'un effet substantiel. Elle permet d'évaluer la probabilité d'un effet néfaste ou contre-intuitif de taille substantielle.

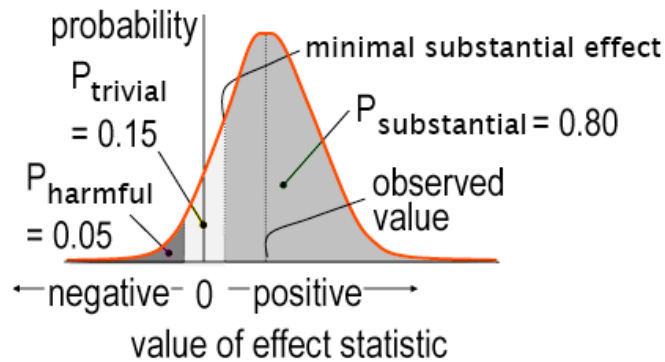
L'Hypothèse d'un effet non-pertinent évalue si l'effet est entre l'effet substantiel minimal et l'effet néfaste minimal. Lors d'une comparaison de moyennes, cette hypothèse correspond à un *test d'équivalence*.

Tester des hypothèses avec des intervalles de confiance

L'hypothèse est **corroborée** si l'intervalle de confiance est totalement incluse dans l'intervalle de l'hypothèse ($p > 0.95$ si $\alpha = 0.05$).

L'hypothèse est **falsifiée** si l'intervalle de confiance est totalement exclue de l'intervalle de l'hypothèse ($p < 0.05$ si $\alpha = 0.05$).

L'hypothèse est **indéterminée** si l'intervalle de confiance est partiellement exclue de l'intervalle de l'hypothèse ($0.05 < p < 0.95$ si $\alpha = 0.05$)



Un exemple complet

Relation entre l'estime de soi et la performance scolaire dans une école secondaire ($n = 200$).

Hypothèse d'effet substantiel: $r > 0.30$

Résultat standard# $r = 0.37$ ($p < 0.05$)

Résultat par intervalle de confiance# $[0.25 < r < 48]$

- H. d'un effet substantiel# indéterminée ($p_{\text{substantiel}} = 0.86$)
- H. d'un effet non-pertinent# indéterminé ($p_{\text{non-pertinent}} = 0.14$)
- H. d'un effet néfaste# falsifié ($p_{\text{néfaste}} = 0.00$)

Interprétation

Dans ce cas, même si l'hypothèse d'un effet substantiel n'a pas été corroborée, des décideurs pourraient quand même aller de l'avant avec un projet d'intervention ciblant l'estime de soi s'ils évaluent que 86% de chance d'effet substantiel est un risque acceptable (particulièrement en considérant que les chances d'effet néfaste sont minces).

Un exemple pour des proportions

Exemple:

Dans un test à l'aveugle fait avec un échantillon représentatif de 300 individus, 167 personnes (55,7%) ont préféré le produit A et 133 (44,3%) ont préféré le produit B. Que peut-on dire à propos des proportions dans la population?

Un chi-carré nous indique que#

$$X^2= 3,63 \text{ et } p=0,05$$

Il y a statistiquement (mais de justesse) plus d'individus qui préfèrent A.

L'intervalle de confiance nous indique que#

$$IC(95\%)=[0,50 - 0,62]\#$$

Entre 50% et 62% des individus préfèrent A.

Présentation des ICs dans le texte

- $IC_{95\%}=[4.5 \text{ à } 8.3]; \quad 95\%CI=[4.5 \text{ to } 8.3]$
- 56% (IC 95% : 50% à 62%)
- Éviter les tirets => $IC_{95\%}=[-2.35 - 3.54]$

Présentation des ICs dans le texte

Table 4.

Confidence intervals of the size of the pretest-posttest difference for excellent, average and low pretest motivation scores.

	C.I. (95%)
For low scores (i.e. 1.34)	1.544 +/- 0.289
For average scores (i.e. 2.50)	0.576 +/- 0.128
For high scores (i.e. 3.66)	-0.404 +/- 0.289

Présentation des ICs dans le texte

- Suggestion de Louis & Zeger (2009)

In text: Compare the clarity and message of “the estimate is 1.48 ($se = 0.09$)” to “the estimate is 1.48_(0.09)” and the clarity and message of “the estimate of excess deaths is 654 (95% CI : 393 to 943)” to that of “the estimate of excess deaths is ₃₉₃654₉₄₃.” Furthermore, note both the clarity and the information content of the 5-number summary, ₃₉₃654₇₄₈₉₄₃. The recommended formats are easier to read and reinforce the message that uncertainty measures are an integral part of an estimate.

Présentation des ICs dans le texte

- Suggestion de Louis & Zeger (2009)

Table 1. Table 6 from Barnard and others (2003) (converted to recommended format): “ITT Effect of Private School Attendance on Test Scores.”

Grade at application	Applicant's school: Low		Applicant's school: High	
	Reading	Math	Reading	Math
1	-2.0 3.4 _{8.7}	3.0 7.7 _{12.4}	-7.3 1.9 _{10.3}	0.2 7.4 _{14.6}
2	-3.7 0.7 _{5.0}	-2.4 1.9 _{6.2}	-9.4 -0.9 _{7.3}	-6.2 1.5 _{9.3}
3	-4.1 1.0 _{6.1}	-0.8 5.0 _{10.7}	-9.5 -0.8 _{7.7}	-4.9 4.0 _{12.5}
4	-1.5 4.2 _{10.1}	-1.6 4.3 _{10.1}	-6.3 2.7 _{11.3}	-4.7 3.5 _{11.9}
Overall	-0.9 2.2 _{5.3}	1.4 4.7 _{7.9}	-7.1 0.6 _{7.7}	-2.6 4.2 _{10.9}

Présentation graphique des ICs

- Le standard: les barres d'erreur
Avertissement 1: Attention à l'unité!
- *SE bars vs SD bars vs CI bars*

Avertissement 2: Une différence entre deux ICs n'est pas l'IC de la différence!
- Solution: Les IC Inférentiels de Tryon (2001)

Présentation graphique des ICs

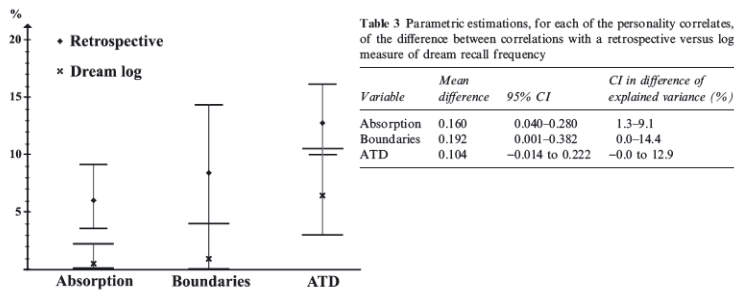


Figure 1. Parametric estimations in percentage of explained variance for each correlate of dream recall frequency represented with Tryon's (2001) inferential confidence intervals (95%).

Les autres types d'intervalles

- Les intervalles empiriques (IE)
Pour capturer une certaine proportion (p.e. 95%) des observations de l'échantillon.
- Calcul basé sur l'écart-type

Les limites de l'intervalle représentent des valeurs extrêmes.

ATTENTION: Ne donne pas d'info sur la population!

Les autres types d'intervalles

- Les intervalles de tolérance (IT)

Pour capturer une proportion de la population avec un certain niveau de confiance.

L'équivalent populationnel des IE

- Besoin des formules pour le calcul

Les limites de l'intervalle représentent les valeurs maximales probables pour cette proportion de la population.

Très utiles pour des calculs d'impact!

Les autres types d'intervalles

- Les intervalles de prédiction (IP)

Pour capturer l'étendu probable des futures observations avec un certain niveau de confiance.

- Besoin des formules pour le calcul

Les limites de l'intervalle représentent les valeurs "maximales" probables pour une observation future.

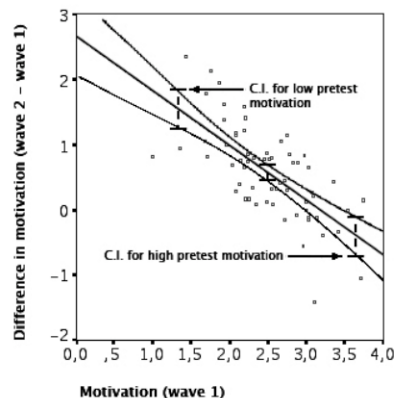
Utile pour régressions et analyses centrées sur l'individu (p.e. pour l'effet d'une thérapie)

Les autres types d'intervalles

- Les bandes de confiance (BC)

C'est l'intervalle de confiance d'une fonction

- Besoin des formules pour le calcul



Les autres types d'intervalles

- Les bandes empiriques
- Les bandes de prédiction
- Les bandes de tolérance
- Les aires de confiance
- ...

Références

La série sur les intervalles de confiance

- 1) CRITIQUE DES TSS ET THÉORIE SUR LES IC
Beaulieu-Prévost, D. (2007). Statistical decision and falsification in science: Going beyond the null hypothesis. In B. Hardy-Vallée (Dir.). *Cognitive decision-making: Empirical and foundational issues*. Cambridge# Cambridge Scholar Publishing. [Une version précédente est disponible à http://eradec.teluq.quebec.ca/IMG/pdf/CIC_2005_05.pdf]
- 2) GUIDE PRATIQUE D'INITIATION AUX IC
Beaulieu-Prévost, D. (2006). From tests of statistical significance to confidence intervals, range hypotheses and substantial effects. *Tutorials in Quantitative Methods for Psychology*, 2, 11-19.
- 3) GUIDE AVANCÉ D'UTILISATION DES IC
Beaulieu-Prévost, D. (sous presse). Gaining confidence with intervals: Practical guidelines, advices and tricks of the trade to face real-life situations. *International Journal of Psychological Research*.

Références

Site internets utiles

- 4) MON SITE PROFESSIONNEL A DES CALCULATEURS D'IC
Beaulieu-Prévost, D. *Professional web site: Statistical resources*, [Online]. <http://www.memoryproject.info/stat.html> .
- 5) UN BON SITE POUR COMPRENDRE LES IC
Hopkins, Will G. *New view of statistics: Confidence limits*, [Online]. <http://www.sportsci.org/resource/stats/generalize.html> .
- 6) UNE RÉFÉRENCE EN STATISTIQUES
NIST/SEMATECH. *e-Handbook of Statistical Methods*, [Online]. <http://www.itl.nist.gov/div898/handbook/> (Page visited october 15th, 2009).

Références

Site internet utiles

•7) LA MÉTHODE DE TRYON (INTERVALLES INFÉRENTIELLES)

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, 6, 371-386.

Tryon, W. W., & Lewis, C. (2008). An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*, 13, 272-277.

•8) À PROPOS DES ERREURS D'INTERPRÉTATION DES IC

Lecoutre, M.-P., Poitevineau, P., & Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Test. *International Journal of Psychology* 38, 37-45.