Institut d'Enseignement à Distance
de l'Université de Paris 8

DEUG de Psychologie deuxième année

Inférence Statistique

Résumés et exercices

Jean-Marc Meunier

Référence : R 2442 T
Classe 371

Introduction.

Cette introduction est surtout une mise en garde contre la tentation de croire que l'étude de ce document puisse permettre de se dispenser de l'étude du cours proprement dit. Le propos de ce document est de vous proposer une aide à l'étude du cours. Il est organisé dans le respect de la structure de votre cours. Vous y trouverez :

- Une définition simple des principaux concepts.
- Un résumé du cours.
- Quelques exercices.
- Les principaux pièges à éviter.
- Une foire aux questions.

. La réalisation des exercices proposés n'a aucun caractère obligatoire, mais est vivement conseillée surtout dans les parties du cours que vous avez du mal à appréhender. **Ces exercices ne doivent pas être envoyés à la correction**. Pour chacun d'eux, vous trouverez un corrigé vous permettant de vous évaluer et de progresser.

Définitions des principaux concepts.

Protocole.

Ensemble d'observations sur une ou plusieurs variables.

Échantillon.

Ensemble d'individus statistiques sur lesquels sont recueillies les données constituant le protocole. L'échantillon est un sous-ensemble de la population.

Population parente.

Également appelée population, c'est l'ensemble des individus statistiques d'où est extrait l'échantillon. La population parente est de taille finie.

Espace des échantillons.

C'est l'ensemble de tous les échantillons possibles obtenu par combinatoire.

Distribution d'échantillonnage.

C'est la distribution, pour une statistique donnée, de l'ensemble des échantillons possibles. Pour les variables numériques, la distribution d'échantillonnage est faite sur la moyenne. Pour les variables nominales ou catégorisée, on utilise généralement la fréquence pour construire la distribution d'échantillonnage.

Échantillonnage dans une population.

C'est l'extraction d'un échantillon dans ensemble de référence de taille finie. L'échantillonnage dans une population peut être vu comme un tirage sans remise.

Échantillonnage dans une distribution.

C'est l'extraction d'un échantillon dans un ensemble de référence de taille infinie. Cette forme d'échantillonnage peut être assimilée à un tirage avec remise dans une population finie.

Modèle d'échantillonnage.

C'est l'ensemble des hypothèses que l'on fait sur le mode de constitution de l'échantillon à partir de la population.

Principes et méthodologie de l'inférence statistique.

Résumé.

Objectif de l'inférence statistique.

L'objectif de l'inférence statistiques est de tester la généralisabilité des conclusions de l'analyse statistique descriptive pour trois objectifs statistiques :

a) Comparaison d'un groupe d'observation à une distribution connue.
b) Comparaison de deux groupes d'observations.
c) Évaluation de la liaison entre deux variables.

Choix du modèle d'échantillonnage.

Le modèle d'échantillonnage est l'ensemble des hypothèses que l'on fait sur le mode de constitution de l'échantillon à partir de la population.

Dans tous les cas, on peut se placer dans le cadre du modèle combinatoire qui consiste à considérer le protocole observé comme un élément de l'ensemble des protocoles possibles (espace des échantillons). La conclusion se formulera en termes de typicité (comparaison d'un groupe d'observation à une distribution de référence) ou d'homogénéité (comparaison de deux groupes d'observations).

On peut parfois se situer dans le cadre du modèle fréquentiste qui consiste à considérer que la distribution d'échantillonnage est une distribution des probabilités d'obtenir un échantillon de telle ou telle moyenne. Pour cela on fait l'hypothèse supplémentaire que l'échantillon a été tiré au hasard dans l'ensemble des protocoles possibles. Cette hypothèse n'est justifiée que si la procédure expérimentale fait intervenir le hasard (sondage ou aléatorisation de la répartition des sujets) ou si l'expérience vise véritablement à tester une hypothèse.

Choix de la distribution d'échantillonnage.

Ce choix dépend de l'échelle de mesure de la variable dépendante. Chaque fois que cela est possible, on préférera la distribution exacte à la distribution approchée.

Echelle de la VD	*Distribution exacte*	*Distribution approchée*
Nominale ou catégorisée.	Hypergéométrique ou binomiale.	Distribution de Z ou X^2
Numérique.	Combinatoire	Z ou T de Student

Mise en œuvre du test.

La mise en œuvre du test dépend de la question posée (objectif statistique) et du choix de la distribution d'échantillonnage. Elle ne dépend pas du modèle d'échantillonnage. Elle sera présentée en détail dans les chapitres suivants.

La démarche générale de l'inférence comporte quatre étapes :

d) Choix du modèle d'échantillonnage (combinatoire ou fréquentiste).
e) Déterminer la distribution d'échantillonnage soit en adoptant une distribution d'échantillonnage approchée, soit en déterminant par combinatoire l'ensemble des protocoles possibles (distribution d'échantillonnage exacte).
f) Situer le protocole observé dans la distribution d'échantillonnage en calculant (ou en lisant dans la table) la proportion d'échantillons plus extrêmes ou égaux que le protocole observé.
g) Comparer cette proportion au seuil-repère .025 (unilatéral) ou .05 (bilatéral).

Formulation de la conclusion.

La formulation de la conclusion repose toujours sur une comparaison entre la proportion observée (calculée ou lue dans une table) et un seuil de significativité fixé par convention à .025 (seuil unilatéral) ou à .05 (seuil bilatéral). Lorsque la proportion observée est inférieure au seuil, le test est déclaré significatif. Le choix entre un seuil unilatéral ou bilatéral dépend du type de comparaison que l'on fait et de la question qu'on se pose (voir le tableau ci-dessous).

Dans tous les cas, le seuil bilatéral est égal à la somme des seuils unilatéraux supérieurs et inférieurs. Dans le cas particulier des distributions d'échantillonnage symétriques, le seuil bilatéral est égal au double d'un des seuils unilatéraux.

Seuil.	*Seuil-repère*	*Échantillon vs population*	*Groupe 1 vs groupe 2*
Unilatéral supérieur	.025	L'échantillon est-il extrême du côté des valeurs élevées ?	Le groupe 1 est-il supérieur au groupe 2 ?
Unilatéral inférieur	.025	L'échantillon est-il extrême du côté des valeurs faibles ?	Le groupe 1 est-il inférieur au groupe 2 ?
Bilatéral.	.05	L'échantillon est-il différent de la population ?	Le groupe 1 est-il différent du groupe 2 ?

Interprétation des résultats du test.

D'un point de vue statistique, l'interprétation du test dépend du type de comparaison que l'on fait et du modèle d'échantillonnage choisi (voir les deux tableaux ci-dessous). Le test est significatif si $p_{obs} \leq$ seuil repère.

Modèle combinatoire	*Échantillon vs population*	*Groupe 1 vs groupe 2*
$p_{obs} \geq$ seuil-repère	Typicité de l'échantillon	Homogénéité des groupes
$p_{obs} \leq$ seuil-repère	Atypicité de l'échantillon	Hétérogénéité des groupes

Modèle fréquentiste	*Échantillon vs population*	*Groupe 1 vs groupe 2*
$p_{obs} \geq$ seuil-repère	Echantillon = ou ≠ population	Groupe 1 = ou ≠ groupe 2
$p_{obs} \leq$ seuil-repère	Echantillon ≠ population	Groupe 1 ≠ groupe 2

Il faut garder à l'esprit deux points importants :

h) Une analyse inférentielle ne dit rien sur l'importance d'une différence, elle permet seulement de se prononcer ou non sur son existence.

i) L'analyse inférentielle est le prolongement de l'analyse descriptive.

j) Toute analyse statistique vise à permettre au chercheur de mieux comprendre les phénomènes psychologiques. Une interprétation statistique des résultats doit donc être accompagnée d'une interprétation psychologique (qu'est-ce que cela m'apprend sur les phénomènes que j'étudie ?).

S'entrainer

Exercice 1 :

« Cette recherche étudie l'impact de l'organisation des procédures sur le processus de planification impliqué dans la rédaction d'instructions. À l'aide de boutons de commandes, trente-quatre sujets devaient sélectionner une lettre (écrire) et/ou sélectionner une cellule (Assembler) pour construire une séquence cible de trois items. Ensuite, ils devaient rédiger une notice d'instructions sur une tablette graphique. Trois situations ont été construites : une situation dans laquelle la macro-commande *écrire* avait une position hiérarchique fonctionnelle superordonnée à la macro-commande *assembler* ; une situation où c'est l'inverse ; une situation contrôle dans laquelle les deux macro-commandes étaient au même niveau. Les résultats indiquent que les propriétés fonctionnelles du matériel sont déterminantes pour la planification des instructions lors de la rédaction : position hiérarchique textuelle des instructions et durée des pauses. »

Potelle H., Passerault J.M. (2002) Effet de l'organisation hiérarchique des procédures sur la planification des textes procéduraux. *Psychologie Française*, Mars 2002, 47, 1.

Exercice 2

« L'objectif de cette étude est de préciser l'évolution de l'organisation gestuelle au cours du vieillissement normal. Pour cela, nous nous sommes appuyés sur le modèle fonctionnaliste de Roy et Square (1985) postulant l'existence de deux systèmes, l'un conceptuel et l'autre de production, à la base de toute action. Le système conceptuel a été étudié à l'aide d'épreuves permettant d'analyser strictement l'aspect conceptuel ou représentationnel de l'action (épreuves n'impliquant aucune production motrice). Le système de production a été étudié à l'aide d'épreuves sur imitation simultanée de gestes n'impliquant qu'un recours minimal à des processus mnésiques ou conceptuels. Quarante sujets âgés de cinquante à quatre-vingt-neuf ans ont bénéficié d'une évaluation avec ce protocole original. Les analyses quantitatives puis qualitatives des résultats aux épreuves évaluant le système conceptuel ont montré, pour tous les sujets, une bonne

connaissance de l'utilisation et de la fonction des objets et outils, une bonne connaissance des actions indépendamment de l'usage des objets, ainsi qu'une bonne connaissance de la sériation des mouvements constitutifs d'une action. Aucun effet du facteur âge n'a été révélé lors de la passation de ces épreuves évaluant la représentation interne et abstraite de l'action. Au contraire, nous avons constaté un effet significatif de l'âge sur chacune des épreuves évaluant le système de production. En référence au modèle théorique de Roy et Square (1985), il existe un effet significatif de l'âge sur les deux niveaux du système de production : le premier permettant la transformation de la représentation interne et abstraite de l'action (élaborée dans le système conceptuel) en programmes généraux et le deuxième assurant la transformation et l'actualisation de ces programmes en action concrète selon les contraintes temporelles et spatiales des différents éléments d'une séquence gestuelle ou d'une action. »

Blondel A., Desgranges B., De La Sayette V., Bouchard S., Lechevallier B., Eustache F., (2001) Les modifications de l'organisation gestuelle au cours du vieillissement normal : une altération spécifique du système de production ? *Revue de neuropsychologie*, 11, 4, pp. 485 – 519.

Exercice 3

Pour cet exercice, on postulera que les sujets ne sont confrontés qu'à une des modalités de chacun des facteurs.

« Le but de cette étude est d'analyser les processus mnésiques qui permettent de rendre compte de l'expertise en danse. Généralement, l'effet de l'expertise sur le rappel d'enchaînements est imputé à la base de connaissances des sujets. Mais d'autres facteurs seraient susceptibles d'intervenir comme la nature du codage de l'information en mémoire. Nous avons donc comparé les performances de rappel moteur de sujets experts en danse en fonction du degré de structuration des enchaînements présentés, des conditions d'encodage (normal, avec tâche d'interférence verbale et motrice) et du style de danse (classique et moderne). Les résultats ont mis en évidence un effet du style de danse ainsi qu'une interaction entre le degré de structuration des enchaînements et le type de tâche d'interférence. De plus, nous constatons que la tâche d'interférence verbale détériore légèrement le rappel et que les enchaînements structurés sont mieux rappelés que les enchaînements non structurés en condition contrôle. Néanmoins, des recherches complémentaires s'avèrent nécessaires étant données la nature de nos propres résultats et l'hétérogénéité des résultats expérimentaux déjà obtenus par ailleurs. »

Jean J., Cadopi M., Ille A. (2001) How are dance sequences encoded and recalled by expert dancers? *Cahiers de psychologie cognitive*. 20 , 5 , pp. 325 - 337

Exercice 4 :

« L'acquisition de l'écriture s'inscrit à la fois dans l'évolution de l'espèce humaine et dans le développement ontogénétique de l'individu. La différenciation entre le dessin et l'écriture s'est faite progressivement au cours de l'histoire. L'enfant baigne aujourd'hui dans un monde fortement sémiotisé et très tôt, il construit des connaissances relatives aux différents domaines de notation. L'objectif de notre travail est d'analyser la façon dont l'enfant, au cours du développement et des apprentissages, différencie le dessin et l'écriture. Nous demandons à des sujets de 3, 4, 5 et 6 ans de dessiner des objets de formes simples (ballon et crayon), d'écrire le mot correspondant, de choisir ce mot parmi un ensemble de cartes et d'expliciter les raisons de leur choix. Les stratégies de réponse utilisées par les enfants font l'objet d'une analyse de la variance et d'une description en termes d'effectifs. Les résultats mettent en évidence quatre stratégies : pictographique, représentation générale, sémiotique et phonographémique. La première est caractéristique des enfants de 3 ans et diminue fortement avec l'âge, la dernière est caractéristique des enfants de 6 ans. Les stratégies de représentation générale (minoritaire) et sémiotique caractérisent les enfants de 4 et 5 ans. Parallèlement, si les connaissances que les enfants de 3 ans ont sur l'écriture sont largement implicites, on observe des progrès dans la capacité à expliciter les déterminants de la réponse, notamment entre 4 et 5 ans. En conclusion on s'interroge sur la pertinence de certaines méthodes pédagogiques pour favoriser la nécessaire différenciation entre dessin et écriture. »

Noyer M., Baldy R. (2002) Du dessin à la lecture et à l'écriture, *Psychologie & éducation*, no 49, pp. 73–88.

Comparaison d'un groupe d'observations à une population ou à une distribution.

Résumé.

Cas d'une variable numérique.

Dans le cas des variables numériques, la distribution d'échantillonnage est la distribution des moyennes associée à l'espace des échantillons. Elle a les propriétés suivantes :

a) La moyenne de la distribution d'échantillonnage de la moyenne est égale à la moyenne de l a distribution parente.
b) Lorsque n/N est petit, la variance de la distribution d'échantillonnage est approximativement égale à la variance de la population parente divisée par la taille de l'échantillon.
c) Plus n est grand, plus la forme de la distribution d'échantillonnage est proche d'une distribution normale.

Échantillonnage dans une population. On fait la combinatoire de tous les échantillons de taille n possibles dans la population. L'espace des échantillons est l'ensemble des échantillons possibles. Le nombre d'échantillons possibles est donné par la formule :

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

On fait ensuite la distribution des échantillons sur la moyenne (distribution d'échantillonnage) et on repère la proportion d'échantillons plus extrêmes que l'échantillon observé (du côté des valeurs faibles ou élevées ou les deux). Pour les distributions d'échantillonnage approchées, on utilisera les distributions de z ou de t comme dans l'échantillonnage sur une distribution.

Échantillonnage dans une distribution.

- Si μ_0 et σ_0 sont connues, on utilise la distribution de Z en appliquant la formule suivante : $Z = \frac{M - \mu_0}{\sigma_0 / \sqrt{n}}$ et en comparant z_{obs} avec la valeur lue dans la table pour le seuil-repère.

Si seule μ_0 est connue, on utilise la statistique T à ν=n-1 degrés de liberté en appliquant la formule suivante : $t_{obs} = \frac{(m_{obs} - \mu_0)}{S / \sqrt{n}}$ puis on compare avec la valeur de la table pour le seuil repère à n-1 degrés de liberté.

Cas d'une variable nominale ou catégorisée.

Échantillonnage dans une population.

On peut, comme pour une variable numérique, faire la combinatoire de tous les protocoles possibles (dans ce cas, la statistique de test est généralement la fréquence). On fera ensuite la distribution des fréquences dans cet espace des échantillons (distribution d'échantillonnage) pour évaluer la proportion d'échantillons extrêmes que l'échantillon observé. On peut également calculer directement cette proportion en utilisant la distribution hypergéométrique représentée par le tableau ci-dessous :

	Catégorie visée.	*Autres catégories.*	*Total.*
Échantillon.	k	n-k	n
Reste après échantillonnage.	A-k	N-A-(n-k)	N-n
Population.	A	N-A	N

Et dans laquelle la proportion p_k recherchée est donnée par la formule :

$$p_k = \frac{\binom{A}{k}\binom{N-A}{n-k}}{\binom{N}{n}} = \frac{A!(N-A)!n!(N-n)!}{k!(n-k)!(A-k)!(N-A-(n-k))!N!}$$

Échantillonnage dans une distribution.

Dans le cas de l'échantillonnage dans une distribution, la distribution exacte est la distribution binomiale. La formule permettant de calculer p_k, avec P la proportion de référence dans la population et Q=1-P, est la suivante :

$$pk = \binom{n}{k} P^k Q^{n-k}$$

Dans le cas de l'échantillonnage dans une distribution, la distribution approchée est la distribution de X^2, les conditions d'application de cette distribution sont:

a) tous les effectifs théoriques sont supérieurs à 5
b) on appliquera une correction de continuité.

La table à consulter est la table de X^2 à un degré de liberté.

S'entraîner.

Exercice 5.

Pour ce premier exercice, nous prendrons un exemple analogue à celui de votre cours. On fait passer à une population de 9 sujets, composée de 3 filles et 6 garçons un test noté sur 50. On a, pour la population la distribution suivante :

Note au test	*Effectif*
18	1
23	2
28	3
33	2
38	1
Total	9

Les trois filles ont obtenu les notes 18, 23 et 28. Peut-on dire que les filles ont des notes faibles à ce test ? On utilisera pour répondre la distribution d'échantillonnage exacte.

Exercice 6.

Reprendre l'exercice 1 en utilisant une distribution d'échantillonnage approchée.

Exercice 7.

On fait passer à l'ensemble des 300 élèves de 3ème d'un collège, dont 25 étudient le latin, un test de compréhension verbale où la note représente le nombre de bonnes réponses sur 40 questions. On se demande si l'étude du latin favorise le développement de ce type de compétence. Sachant que les latinistes ont obtenu une moyenne de 24,60 avec un écart-type de 12,580 et l'ensemble des élèves de 3ème, une moyenne de 19,30, peut-on dire que les latinistes ont une meilleure réussite à ce test ?

Exercice 8.

Recommencez l'exercice 3 sachant que la variance des notes au test pour tous les élèves de 3ème est de 183,684.

Exercice 9.

On a fait passer un test de raisonnement à une population de 10 sujets et on observe dans cette population une fréquence de réussite de 30 %. Dans cette population, 3 sujets sont mathématiciens et un seul d'entre eux a échoué au test. Peut-on dire que les mathématiciens réussissent plus souvent ce test de raisonnement que le reste de la population ? Pour répondre à cette question, on déterminera la distribution d'échantillonnage par combinatoire.

Exercice 10.

Nous reprendrons, pour cet exercice, les données de l'exercice 5. La question posée reste la même. On prendra pour distribution d'échantillonnage, soit la distribution hypergéométrique, soit la distribution binomiale. Justifiez votre choix.

Exercice 11.

Nous allons reprendre les données de l'exercice 9. Peut-on dire que les 10 sujets ayant participé à l'expérience ont fourni des réponses qui diffèrent du hasard ? On utilisera pour répondre une distribution d'échantillonnage exacte.

Exercice 12.

Même énoncé que le 7, mais avec une distribution d'échantillonnage approchée.

Comparaison de deux groupes d'observations.

Résumé.

Cas d'une variable numérique pour deux groupes appariés (plan S*T$_2$).

Distribution d'échantillonnage exacte. Dans ce cas, l'espace des échantillons est l'ensemble des échantillons qu'on peut obtenir en permutant les observations d'un ou plusieurs couples de données. Le nombre de permutations possibles est de 2^n. La distribution d'échantillonnage est la distribution, sur tous les échantillons possibles, des moyennes des différences entre les observations de chacun des couples.

Distribution d'échantillonnage approchée.

Cas où σ^2 est connue. Dans ce cas, on utilise la distribution z avec la formule suivante : $z_{obs} = \frac{(d - \delta_0)}{\sigma_0 / \sqrt{n}}$ où d est la moyenne des différences individuelles du protocole et δ_0=0. On lit ensuite la proportion associée à z_{obs} dans la table de la distribution normale réduite.

Cas où σ2 n'est pas connue. Dans ce cas, on utilise la variable T en prenant comme estimateur de σ^2 la variance corrigée du protocole des différences individuelles notée ici s^2. La formule à appliquer est la suivante :
$t_{obs} = \frac{(d - \delta_0)}{s / \sqrt{n}}$ où d est la moyenne des différences individuelles du protocole et δ_0=0. On lit ensuite la proportion associée à t_{obs} à n-1 degrés de liberté.

Cas d'une variable numérique pour deux groupes indépendants (plan S<G$_2$>).

Distribution d'échantillonnage exacte. Dans ce cas, l'espace des échantillons est l'ensemble de toutes les partitions possibles de n' et n'' éléments, c'est-à-dire le nombre d'observations dans chacun des deux groupes. Le nombre d'échantillons possibles est donné par la formule suivante :
$\binom{N}{n} = \frac{N!}{n!(N-n)!}$

La distribution d'échantillonnage est la distribution, sur tous les échantillons possibles, des différences de moyennes entre les deux groupes.

Distribution d'échantillonnage approchée.

Cas où σ^2 est connue. Dans ce cas, on utilise la variable Z. La formule est alors la suivante :

$z_{obs} = \frac{m' - m''}{\sigma_0 \sqrt{\frac{1}{n'} + \frac{1}{n''}}}$ où m' et m'' sont les moyennes des groupes et n' et n'' sont les effectifs des deux groupes. On consulte ensuite la table de la distribution z.

Cas où σ^2 n'est pas connue. Dans ce cas, on utilise la variable T. La formule est alors la suivante :

$t_{obs} = \frac{m' - m''}{s \sqrt{\frac{1}{n'} + \frac{1}{n''}}}$ où s est l'écart-type corrigé intragroupe obtenu à partir de la variance corrigée intragroupe avec la formule suivante : **Erreur! Signet non défini.**
On consulte la table à n'+n''-2 degrés de liberté.

Cas d'une variable nominale ou catégorisée pour deux groupes appariés (plan $S*T_2$).
Distribution d'échantillonnage exacte. Comme pour une variable numérique, l'espace des échantillons est l'ensemble des échantillons qu'on peut obtenir en permutant les observations d'un ou plusieurs couples de données. Le nombre de permutations possibles est de 2^n. La distribution d'échantillonnage est la distribution, sur tous les échantillons possibles, des fréquences en faveur de l'une ou l'autre des deux tâches, en négligeant les cas d'égalité. La distribution d'échantillonnage exacte est la distribution binomiale. La formule permettant de calculer p_k, avec P la proportion de référence dans la population et Q=1-P, est la suivante :

$pk = \binom{n}{k} P^k Q^{n-k}$

Distribution d'échantillonnage approchée. Elle correspond à la distribution de X^2 à un degré de liberté. Comme pour la comparaison à une distribution de référence, on vérifiera que tous les effectifs théoriques sont supérieurs à 5 et on appliquera une correction de continuité. On calculera ensuite $X^2_{corr} = \sum \frac{\left(|e_{obs} - e_{théo}| - 0,5\right)^2}{e_{théo}}$, puis on consulte la table de X^2 à un degré de liberté.

On peut également utiliser le test de Mc Némar qui fait également appel à la variable X^2. Les données sont alors présentées sous la forme d'un tableau à 4 cases comme celui-ci :

		Deuxième tâche. +	Deuxième tâche. -
Première tâche.	+	A	B
	-	C	D

Et X^2 se calcule de la façon suivante : $X^2_{corr} = \frac{(|A+D|-1)^2}{A+D}$

Cas d'une variable nominale ou catégorisée pour deux groupes indépendants (plan S<G$_2$>).

Distribution d'échantillonnage exacte. Comme pour les variables numériques, l'espace des échantillons possibles est l'ensemble des partitions possibles de n' et n'' éléments, c'est-à-dire le nombre d'observations dans chacun des deux groupes. Le nombre d'échantillons possibles est donné par la formule suivante :

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

La distribution d'échantillonnage est la distribution des fréquences des échantillons possibles.

k) Dans le cas où on a que deux groupes d'observations et deux catégories de réponse (soit donc un tableau à 4 cases) la distribution exacte est la distribution hypergéométrique avec comme paramètres :

	Catégorie visée.	*Autres catégories.*	*Total.*
Groupe 1	k	n-k	n
Groupe 2	A-k	N-A-(n-k)	N-n
Total	A	N-A	N

La formule de p_k est la même, soit : $p_k = \frac{\binom{A}{k}\binom{N-A}{n-k}}{\binom{N}{n}} = \frac{A!(N-A)!n!(N-n)!}{k!(n-k)!(A-k)!(N-A-(n-k))!N!}$

a) Si nous avons plus de deux catégories de réponses et/ou plus de deux groupes d'observations, la distribution exacte est la distribution multinomiale qui n'est pas abordée dans ce cours. Dans ce cas, on utilisera comme distribution approchée la distribution de X^2 à plusieurs degrés de liberté.

Distribution d'échantillonnage approchée.

Dans les deux cas, on utilise la distribution de X^2 à (L-1)*(C-1) degrés de liberté. L est le nombre de lignes et C le nombre de colonnes du tableau de données.

a) **- Dans le cas d'un tableau à quatre cases**, on vérifiera qu'aucun effectif théorique n'est inférieur à 5 et on appliquera un correction de continuité à X^2 en appliquant la formule $X^2_{corr} = \sum \frac{\left(|e_{obs} - e_{théo}| - 0,5\right)^2}{e_{théo}}$, puis on consulte la table de X^2 à un degré de liberté.

b) **Dans le cas d'un tableau à plus de quatre cases**, on vérifiera que moins de 20% des effectifs théoriques sont inférieurs à 5. Dans ce cas, la correction de continuité n'est pas nécessaire. On appliquera donc la formule suivante : $\chi 2 = \sum \frac{(e_{obs} - e_{théo})2}{e_{théo}}$. On consultera la table à (L-1)*(C-1) degrés de liberté.

S'entrainer

L'expérience suivante porte sur l'acquisition de connaissances sur l'utilisation de logiciels de traitement de textes. On cherche à comparer les effets d'un cours préalable à l'utilisation du logiciel et l'utilisation d'un didacticiel sur la résolution de problèmes liés à l'utilisation du logiciel. On a ainsi constitué trois groupes de 9 sujets. Le premier groupe suit un cours sur l'utilisation du logiciel (groupe formation), puis doit résoudre deux problèmes. Le second groupe apprend à se servir du logiciel à l'aide d'un programme d'apprentissage (groupe didacticiel) avant de résoudre les deux problèmes. Le troisième groupe ne suit aucun apprentissage (groupe contrôle) avant de résoudre les deux problèmes. Deux variables dépendantes ont été étudiées: le temps de résolution (T) exprimé en minutes et le nombre d'actions (E) que le sujet fait pour résoudre chacun des problèmes. Les résultats de cette expérience sont donnés dans le protocole suivant :

Apprentissage	*Temps*		*NB essai*	
	T1	*T2*	*E1*	*E2*
Contrôle	5,8	5,9	20	16
Contrôle	5,7	5	21	15
Contrôle	5,2	4,7	20	14
Contrôle	5,6	5,2	19	16
Contrôle	6,5	5,2	18	17
Contrôle	2,7	2,3	20	19
Contrôle	5,9	5,2	12	12
Contrôle	4	4,2	16	15
Contrôle	1,8	2,3	16	15
Formation	6,8	4,4	15	12
Formation	3,7	5,5	13	14
Formation	4,4	2,6	14	12

Formation	5,2	2,5	15	12
Formation	4,5	2,5	12	11
Formation	3,4	2,5	12	10
Formation	4,7	4,3	14	13
Formation	3,6	3,1	12	11
Formation	6,9	4,5	13	10
Didacticiel	3,5	3	15	11
Didacticiel	3,2	2	14	12
Didacticiel	5,6	5,6	13	11
Didacticiel	6,9	5,7	12	12
Didacticiel	2,4	2	12	13
Didacticiel	4,4	4,3	13	12
Didacticiel	4,5	3,4	15	12
Didacticiel	6,4	5,2	13	12
Didacticiel	5,6	5,2	12	10

Exercice 13.

Peut-on dire, tous groupes confondus que les sujets résolvent le second problème plus vite que le premier ? On répondra à cette question sur le plan descriptif et inférentiel . On utilisera une distribution d'échantillonnage approchée.

Exercice 14.

Même question que l'exercice 13 pour le groupe « contrôle ».

Exercice 15.

Même exercice que le 13 pour le groupe « formation ».

Exercice 16.

Même exercice que le 14 pour le groupe « didacticiel ».

Exercice 17.

Comparer les temps de résolution aux deux problèmes pour les cinq premiers sujets en utilisant la distribution d'échantillonnage exacte. Concluez.

Exercice 18.

Du point de vue du temps de résolution, peut-on dire que l'enseignement (formation ou didacticiel) permet de résoudre plus rapidement le premier problème que dans le groupe « contrôle » ? Répondez à cette question d'un point de vue descriptif et inférentiel en utilisant une distribution d'échantillonnage approchée.

Exercice 19.

Peut-on dire que les deux groupes ayant suivis une formation mettent le même temps pour résoudre le problème 1 ?

Quelques exercices de plus ?

Pour vous entrainer, vous pouvez recommencer les exercices de 13 à 19 sur la variable "nombre d'actions".

Liaison entre deux groupes d'observations appariés.

Résumé.

Cas de variables numérique.

Comme pour les groupes appariés considérés précédemment, l'espace des échantillons s'obtient en faisant toutes les permutations possibles. La distribution d'échantillonnage est la distribution des coefficients de corrélation (r de Bravais-Pearson) des échantillons de même taille issus de la distribution de référence. On ne considère ici que le cas particulier d'une distribution de r = 0.

Dans ce cas, on approche la distribution d'échantillonnage en utilisant la statistique $T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, qui suit approximativement une distribution de Student à n-2 degrés de libertés.

Cas de variables nominales.

Dans ce cas, la distribution d'échantillonnage exacte est la distribution hypergéométrique ou multinomiale, selon qu'on a un tableau à quatre cases ou plus. La distribution approchée est la distribution de X^2 à (L-1)*(C-1) degrés de liberté.

Comme précédemment, dans le cas des tableaux à quatre cases, on vérifiera qu'aucun effectif n'est inférieur à 5 et on appliquera une correction de continuité. Nous rappelons ci-dessous la formule à appliquer :

$$X^2_{corr} = \sum \frac{\left(|e_{obs} - e_{théo}| - 0,5\right)^2}{e_{théo}}$$

On consulte ensuite la table de X^2 à 1 degré de liberté.

Dans le cas où on a un tableau à plus de quatre cases, il n'est pas nécessaire d'appliquer le coefficient de continuité. On vérifiera cependant que moins de 20% des effectifs théoriques sont inférieurs à 5. On consulte alors la table de X^2 à (L-1)*(C-1) degrés de liberté.

S'entraîner.

Exercice 20.

Dans cet exercice, nous reprendrons les données de l'exercice 12. Pour plus de facilité, nous les reproduisons ci-dessous. Nous nous demanderons si le temps de résolution au

problème 1 est corrélé au temps de résolution au problème 2. La question sera examinée au niveau descriptif et au niveau inférentiel.

Apprentissage	*Temps*		*NB essai*	
	T1	*T2*	*E1*	*E2*
Contrôle	5,8	5,9	20	16
Contrôle	5,7	5	21	15
Contrôle	5,2	4,7	20	14
Contrôle	5,6	5,2	19	16
Contrôle	6,5	5,2	18	17
Contrôle	2,7	2,3	20	19
Contrôle	5,9	5,2	12	12
Contrôle	4	4,2	16	15
Contrôle	1,8	2,3	16	15
Formation	6,8	4,4	15	12
Formation	3,7	5,5	13	14
Formation	4,4	2,6	14	12
Formation	5,2	2,5	15	12
Formation	4,5	2,5	12	11
Formation	3,4	2,5	12	10
Formation	4,7	4,3	14	13
Formation	3,6	3,1	12	11
Formation	6,9	4,5	13	10
Didacticiel	3,5	3	15	11
Didacticiel	3,2	2	14	12
Didacticiel	5,6	5,6	13	11
Didacticiel	6,9	5,7	12	12
Didacticiel	2,4	2	12	13
Didacticiel	4,4	4,3	13	12
Didacticiel	4,5	3,4	15	12
Didacticiel	6,4	5,2	13	12
Didacticiel	5,6	5,2	12	10

Exercice 21.

On reprendra les mêmes données et le même questionnement qu'à l'exercice 16, en examinant cette fois la variable "nombre d'actions".

Exercice 22.

Voici un exemple tiré du sujet d'examen de juin 2002. Dans une enquête, on désire étudier la mobilité sociale des femmes, c'est-à-dire le changement de catégories socioprofessionnelles (CSP) d'une génération à une autre. On relève donc auprès de 130 familles la CSP de la mère et de la fille. Le résultat de cette enquête est présenté dans le tableau ci-dessous. Analysez ces données au niveau descriptif et inférentiel. Commentez et concluez.

	Filles

Effectifs observés	Ouvrières	Employées	Cadres	Total
Mères Ouvrières	25	13	11	49
Mères Employées	10	23	12	45
Mères Cadres	11	12	13	36
Total	46	48	36	130

Corrigés des exercices

Corrigé de l'exercice 1.

Dans cette étude, les individus statistiques sont les sujets. Ils sont au nombre de 34. Ils ont tous la même tâche à réaliser : construire une séquence cible en sélectionnant une lettre ou une cellule. Ce n'est donc pas une variable. L'expérimentateur demande ensuite de rédiger une notice d'instruction. Ce n'est pas non plus une variable, tous les sujets ont à faire cela.

Ce qui varie, en revanche, c'est la position hiérarchique fonctionnelle des commandes. C'est une variable nominale à trois modalités (écrire superordonnée à assembler, assembler superordonné à écrire, même niveau pour les deux actions). Bien que ce ne soit pas explicité dans le résumé, on peut supposer que chacun des sujets ne voit qu'une de ces trois conditions (sinon la comparaison n'aurait guère de sens). Il existe donc une relation d'emboîtement entre le facteur sujet et le facteur « groupe ».

Dans ce résumé, deux variables dépendantes (observées) sont évoquées :

m) La position hiérarchique textuelle.

n) La durée des pauses.

Nous n'avons pas de précision sur l'échelle de mesure utilisée.

L'objectif de cette recherche est la comparaison de plusieurs groupes indépendants d'observations sur les deux variables dépendantes. Dans le cadre de l'approche combinatoire, la question revient à tester l'homogénéité des groupes. On peut supposer que dans cette recherche, la répartition des sujets dans les groupes a été faite au hasard. Par ailleurs, il s'agit bien de tester l'hypothèse de l'influence de la position hiérarchique fonctionnelle des commandes sur la rédaction des notices. On peut donc se placer dans le cadre de l'inférence fréquentiste et interpréter le résultat de l'analyse en termes de rejet ou de conservation de l'hypothèse nulle.

Corrigé de l'exercice 2 .

Dans cette expérience, les individus statistiques sont les sujets. Un premier facteur est constitué par l'âge des sujets. On ne nous dit rien de l'échelle de mesure de ce facteur. Mais les sujets ne pouvant appartenir qu'à une tranche d'âge, il existe une relation d'emboîtement entre ce facteur et le facteur sujet.

Nous avons deux variables dépendantes : performance du système conceptuelle et performance du système de production. Aucune précision n'est donnée dans ce résumé sur l'échelle de mesure utilisée. On comprend cependant aisément que la problématique de cette étude est la comparaison des groupes d'âges sur les deux variables dépendantes. Dans l'approche combinatoire, cela revient à tester l'homogénéité de

groupes indépendants sur les deux variables dépendantes. Dans cette étude, on ne peut pas se placer dans le cadre de l'approche fréquentiste. La première raison c'est que les sujets ne sont pas répartis de manière aléatoire dans les groupes. Ensuite on ne cherche pas vraiment à tester une hypothèse, mais plutôt l'équivalence entre les groupes. Nous resterons donc, pour cette étude, dans une approche combinatoire.

Corrigé de l'exercice 3 .

Dans cette expérience, la variable dépendante (VD) ou variable observée est la performance du rappel moteur. Aucune précision n'est donnée dans ce résumé quant à la façon dont cette mesure est prise. Tous les sujets sont experts en danse, c'est donc une constante et non une variable.

La première variable indépendante (VI) est le degré de structuration des enchaînements (structuré versus non-structuré). C'est une variable nominale dichotomique . Comme cela est demandé en introduction, nous considèrerons que ce facteur entretient une relation d'emboîtement avec le facteur sujet.

La seconde VI est constituée par les conditions d'encodage. C'est une variable nominale à trois modalités (normal, interférence verbale, interférence motrice). Pour la même raison que précédemment, nous considèrerons qu'il existe une relation d'emboîtement entre ce facteur et le facteur sujet.

La troisième VI est le style de danse (moderne ou classique). C'est une variable nominale dichotomique. Comme pour les deux autres VI, nous considèrerons une relation d'emboîtement entre ce facteur et le facteur sujet.

Cette expérience relève donc de la comparaison de groupes indépendants d'observations. Dans le cadre d'une approche combinatoire, il s'agit d'évaluer l'homogénéité des groupes. On peut considérer que les sujets sont répartis de manière aléatoire dans les groupes. On pourra donc se placer dans le cadre de l'inférence fréquentiste et interpréter le résultat en termes de rejet ou de conservation de l'hypothèse nulle.

Corrigé de l'exercice 4 .

Dans cette étude, les individus statistiques sont les enfants. Le premier facteur est l'âge. Les enfants sont répartis dans 4 groupes d'ages. Dans la mesure où ces groupes sont exclusifs, il existe une relation d'emboîtement entre ce facteur et le facteur sujet.

On demande aux enfants d'effectuer plusieurs tâches (dessin, écriture, sélection de cartes, explicitation des choix). Il s'agit ici d'une répétition de mesure. Il y a donc une relation de croisement entre ce facteur et le facteur sujet.

Les stratégies de réponses ne constituent pas un facteur puisqu'elles sont identifiées à partir des résultats. La variable dépendante est constituée par les réponses données dans chacune des tâches. Aucune précision n'est donnée dans ce résumé sur la nature de

cette variable (par exemple, compte-t-on les bonnes ou les mauvaises réponses ?), ni sur l'échelle de mesure utilisée.
L'objectif statistique, dans cette expérience, est la comparaison de groupes d'observations. Dans le cadre de l'inférence combinatoire, on cherche à se prononcer sur l'homogénéité des groupes d'observations. La répartition des enfants dans les groupes d'âges ne relève pas du hasard. Il ne s'agit pas non plus de tester une hypothèse (il s'agit ici d'identifier des différences de stratégie de réponse selon le groupe d'âge). On ne peut donc se placer dans le cadre de l'inférence fréquentiste.

Corrigé de l'exercice 5.

La question posée peut se reformuler en termes statistiques de la façon suivante : Les filles sont-elles atypiques ou extrêmes du coté des notes basses à ce test ? D'un point de vue descriptif, on peut voir que la moyenne des filles (m=23) à ce test est plus faible que la moyenne des garçons (m=28). Il semblerait donc que les filles aient de mauvais résultats à ce test[1]. Peut-on prolonger ce constat au niveau inductif ? Autrement dit, une telle moyenne est-elle suffisamment exceptionnelle dans la population pour penser que la moyenne des filles constituent un groupe à part dans la population ?

Choix du modèle d'échantillonnage.
Le modèle d'échantillonnage retenu est l'approche combinatoire. On ne peut en effet pas se situer dans le cadre de l'approche fréquentiste puisque les sujets ne sont pas sélectionnés au hasard.

Détermination de la distribution d'échantillonnage.
L'énoncé de cet exercice est explicite, on doit utiliser une distribution d'échantillonnage exacte.
Nous sommes dans le cas d'un échantillonnage dans une population. Notre variable étant numérique, nous n'avons pas d'autre choix que de faire la combinatoire .

1) Choix de la statistique d'échantillonnage.
Avant de déterminer la distribution d'échantillonnage, il nous faut décider de la statistique sur laquelle sera faite cette distribution. Puisque notre variable est numérique, la moyenne constitue une statistique tout indiquée.

[1] On notera ici que j'évite de dire que les filles sont moins bonnes que les garçons, non pas parce que je suis spécialement féministe, mais parce qu'une telle formulation relève de la comparaison de deux groupes d'observations, ce qui relève d'une autre démarche d'analyse.

2) Constitution de l'espace des échantillons.

La population comptant neuf éléments (N=9) et l'échantillon, trois (n=3), nous aurons donc 84 échantillons possibles.

L'espace des échantillons est le suivant :

I	18	23	23	28	28	28	33	33	38	Σ	M	I	18	23	23	28	28	28	33	33	38	Σ	M
1	18	23	23							64	21,33	43		23			28				38	89	29,67
2	18	23		28						69	23,00	44		23				28	33			84	28,00
3	18	23			28					69	23,00	45		23				28		33		84	28,00
4	18	23				28				69	23,00	46		23				28			38	89	29,67
5	18	23					33			74	24,67	47		23					33	33		89	29,67
6	18	23						33		74	24,67	48		23					33		38	94	31,33
7	18	23							38	79	26,33	49		23						33	38	94	31,33
8	18		23	28						69	23,00	50			23	28	28					79	26,33
9	18		23		28					69	23,00	51			23	28		28				79	26,33
10	18		23			28				69	23,00	52			23	28			33			84	28,00
11	18		23				33			74	24,67	53			23	28				33		84	28,00
12	18		23					33		74	24,67	54			23	28					38	89	29,67
13	18		23						38	79	26,33	55			23		28	28				79	26,33
14	18			28	28					74	24,67	56			23		28		33			84	28,00
15	18			28		28				74	24,67	57			23		28			33		84	28,00
16	18			28			33			79	26,33	58			23		28				38	89	29,67
17	18			28				33		79	26,33	59			23			28	33			84	28,00
18	18			28					38	84	28,00	60			23			28		33		84	28,00
19	18				28	28				74	24,67	61			23			28			38	89	29,67
20	18				28		33			79	26,33	62			23				33	33		89	29,67
21	18				28			33		79	26,33	63			23				33		38	94	31,33
22	18				28				38	84	28,00	64			23					33	38	94	31,33
23	18					28	33			79	26,33	65				28	28	28				84	28,00
24	18					28		33		79	26,33	66				28	28		33			89	29,67
25	18					28			38	84	28,00	67				28	28			33		89	29,67
26	18						33	33		84	28,00	68				28	28				38	94	31,33
27	18						33		38	89	29,67	69				28		28	33			89	29,67
28	18							33	38	89	29,67	70				28		28		33		89	29,67
29		23	23	28						74	24,67	71				28		28			38	94	31,33
30		23	23		28					74	24,67	72				28			33	33		94	31,33
31		23	23			28				74	24,67	73				28			33		38	99	33,00
32		23	23				33			79	26,33	74				28				33	38	99	33,00
33		23	23					33		79	26,33	75					28	28	33			89	29,67
34		23	23						38	84	28,00	76					28	28		33		89	29,67
35		23		28	28					79	26,33	77					28	28			38	94	31,33
36		23		28		28				79	26,33	78					28		33	33		94	31,33
37		23		28			33			84	28,00	79					28		33		38	99	33,00
38		23		28				33		84	28,00	80					28			33	38	99	33,00
39		23		28					38	89	29,67	81						28	33	33		94	31,33
40		23			28	28				79	26,33	82						28	33		38	99	33,00
41		23			28		33			84	28,00	83						28		33	38	99	33,00
42		23			28			33		84	28,00	84							33	33	38	104	34,67

3) Détermination de la distribution d'échantillonnage.

C'est la distribution de la statistique d'échantillonnage (dans notre cas la moyenne) sur l'espace des échantillons (ensemble des échantillons de taille n possibles). Cette distribution se fait de la même façon que les distributions que vous avez eu à faire en première année : on compte, pour chaque valeur de la moyenne , le nombre de fois où elle a été observée dans l'espace des échantillons. Cette distribution est la suivante (on s'est limité aux valeurs observées).

M	*Effectifs*	*Fréquences*	*Fréquences cumulées à gauche*	*Fréquences cumulées à droite*
21,33	1	0,012	0,012	1,000
23,00	6	0,071	0,083	0,988
24,67	10	0,119	0,202	0,917
26,33	16	0,190	0,393	0,798
28,00	18	0,214	0,607	0,607
29,67	16	0,190	0,798	0,393
31,33	10	0,119	0,917	0,202
33,00	6	0,071	0,988	0,083
34,67	1	0,012	1,000	0,012
Total	84	1,000		

Situer l'échantillon dans la distribution d'échantillonnage.

La moyenne de notre échantillon est de 23. Notre question est de savoir si cette moyenne est extrême du côté des valeurs faibles. On cherche la fréquence des moyennes inférieures ou égales à la moyenne de l'échantillon. On regardera donc les fréquences cumulées à gauche. Notre question ne portant que sur une des extrémités de la distribution, c'est donc une question unilatérale, notre seuil repère est de 0,025.Sur la ligne correspondant à 23, on peut lire que cette proportion est de 0,083.

Comparer la proportion observée au seuil repère.

Cette proportion étant supérieure au seuil repère, on ne peut pas dire que cet échantillon soit atypique. Il est donc typique. D'un point de vue psychologique, cela veut dire que les filles n'ont pas moins bien réussi le test que l'ensemble de la population.

Corrigé de l'exercice 6.

Choix du modèle d'échantillonnage.

Le modèle d'échantillonnage sera le même, pour les raisons évoquées plus haut.

Détermination de la distribution d'échantillonnage.

La statistique d'échantillonnage reste la même, c'est la moyenne. Nous connaissons la moyenne (μ_0=28) et la variance (σ^2=33,33) de la population de référence . Nous prendrons donc la distribution Z comme distribution d'échantillonnage. Il n'est pas

besoin de calculer l'ensemble de la distribution d'échantillonnage puisqu'elle est donnée par les tables de la loi normale.

Situer l'échantillon dans la distribution d'échantillonnage.
Pour cela, nous allons utiliser la formule de z adaptée à la comparaison d'un groupe d'observations à une distribution de référence.

$$z_{obs} = \frac{M - \mu_0}{\sigma_0 / \sqrt{n}} = \frac{23 - 28}{5,77 / \sqrt{3}} = -1,5$$

La valeur de z_{obs} étant négative et notre question étant unilatérale, on lira dans la table de la distribution Z cumulée à gauche. Pour z_{obs}=-1,5, on peut lire la proportion p(u<z)= 0,067.

Comparaison au seuil repère et formulation de la conclusion.
La proportion lue dans la table étant supérieure au seuil repère, on ne peut pas dire que l'échantillon soit atypique de la population. Nous retrouvons, ici, fort heureusement la même conclusion que dans l'exercice précédent. La proportion trouvée dans cet exercice n'est pas la même que dans l'exercice précédent. Rappelons que la distribution Z n'est qu'une distribution approchée et que la distribution de la moyenne tend vers une distribution normale que lorsque N est suffisamment grand (ce qui n'est pas le cas). Nous sommes, avec cette population, dans un cas limite de l'utilisation de la distribution Z.

Corrigé de l'exercice 7.

Choix du modèle d'échantillonnage.
Dans une telle étude, l'adoption du modèle combinatoire revient à se demander si les élèves latinistes de 3ème sont atypiques de la population des élèves de 3ème. On ne peut pas se placer dans le cadre de l'inférence fréquentiste, puisque les élèves de l'échantillon n'ont pas été choisis aléatoirement. Par ailleurs la question sous-jacente à cette étude est de savoir si les élèves de l'échantillon diffèrent de l'ensemble de la population, on ne teste donc pas une hypothèse sur l'effet d'un facteur. Nous resterons donc dans le modèle combinatoire.

Détermination de la distribution d'échantillonnage.
La variable étant numérique, la statistique d'échantillonnage appropriée est la moyenne. On connaît par l'énoncé la moyenne et l'écart-type de l'échantillon. On connaît également la moyenne de la population de référence, mais on ne connaît pas sa variance. La distribution exacte n'est pas utilisable, puisque nous n'avons pas les données concernant la population. La distribution d'échantillonnage sera donc la distribution du t de Student.

Situer l'échantillon dans la distribution d'échantillonnage.
Pour cela, nous allons utiliser la formule du t de Student adaptée à la comparaison d'un groupe d'observations à une distribution de référence.

$$t_{obs} = \frac{(m_{obs} - \mu_0)}{S/\sqrt{n}} = \frac{(24{,}60 - 19{,}30)}{12{,}84/\sqrt{25}} = 2{,}06$$

où s à été calculé en élevant au carré l'écart-type des latinistes ($15{,}58^2$), ce résultat est ensuite multiplier par n=25 pour obtenir la somme des carrés des écarts à la moyenne, puis divisée par n-1=24 pour obtenir la variance corrigée. On extrait ensuite la racine carrée de la variance corrigée pour obtenir s : l'écart-type corrigé.

$$s = \sqrt{\frac{12{,}58^2 * 25}{24}} = 12{,}84$$

Comparaison au seuil repère et formulation de la conclusion.
On cherche ensuite dans la table, à la ligne ν=n-1 degrés de liberté, soit donc pour notre exemple, la ligne 24, la valeur inférieure ou égale la plus proche de t_{obs}. Ici, par chance, la valeur 2,06 se trouve dans la table, mais ce n'est pas toujours le cas. On remonte ensuite en tête de colonne lire la valeur de p. Si p est inférieur ou égal au seuil repère, ce qui est le cas dans notre exemple, on déclare le résultat significatif. Cela veut dire qu'on peut déclarer l'échantillon atypique de la population.
Une autre façon de lire cette table est de chercher dans la colonne correspondant au seuil repère (.025 ou .05) la valeur inscrite sur la ligne correspondant au nombre de degrés de liberté (ici, 24). On compare ensuite la valeur observée à la valeur de la table. Le résultat est déclaré significatif, si la valeur observée est supérieure ou égale à la valeur de la table (comme c'est le cas dans notre exemple).
D'un point de vue psychologique, notre résultat significatif veut dire que les élèves latinistes ont mieux réussi le test que l'ensemble de la population.

Corrigé de l'exercice 8.

Choix du modèle d'échantillonnage.
Nous conservons bien sûr, le même modèle d'échantillonnage.

Détermination de la distribution d'échantillonnage.
Dans cet exercice, on nous donne à connaître la variance de la population parente. La statistique de test restant la même, nous pouvons maintenant utiliser la distribution Z comme distribution d'échantillonnage.

Situer l'échantillon dans la distribution d'échantillonnage.

Nous utiliserons pour cela la formule de z adaptée à la comparaison d'un groupe d'observations à une distribution de différence. Avant cela, on aura extrait la racine carrée de la variance parente afin d'obtenir l'écart-type.

$$\sigma_0 = \sqrt{183,684} = 13,553$$

On applique ensuite la formule de z.

$$z_{obs} = \frac{m_{obs} - \mu_0}{\sigma_0 / \sqrt{n}} = \frac{24,60 - 19,30}{13,553 / \sqrt{25}} = 1,96$$

Comparaison au seuil repère et formulation de la conclusion.
On cherche ensuite dans la table de la distribution normale réduite (cumulée à droite puisque z_{obs} est positif) la proportion correspondante. Celle-ci est de 0,025. Dans cet exemple, p_{obs} est égal au seuil repère. On peut donc déclarer le test significatif au seuil .025. Les latinistes sont bien atypiques de la population des élèves de 3ème du point de vue du test. Leur moyenne étant plus élevée à ce test, on peut dire qu'ils ont mieux réussi ce test que l'ensemble des élèves.

Note : Vous avez pu voir la similitude entre le z et le t. Si la valeur pour p_{obs} est la même, il ne faut pas confondre ces deux statistiques. Elles ne donnent pas le même résultat et ne font pas appel à la même table statistique. Mais toutes deux aboutissent à la même conclusion.

Corrigé de l'exercice 9.

Choix du modèle d'échantillonnage.
Dans le cadre du modèle combinatoire, la question posée revient à se demander si l'échantillon des mathématiciens est atypique de la population des sujets ayant passé le test. On se trouve dans une problématique similaire à celle des élèves latinistes de 3ème que nous avons rencontrée dans l'exercice 3. Les sujets n'ont pas été choisis au hasard et la question posée revient à se demander si les mathématiciens diffèrent de la population de sujet ayant passé le test de raisonnement. Nous resterons donc dans le modèle combinatoire.

Détermination de la distribution d'échantillonnage.
Dans cet exercice, la variable est nominale. La statistique de test sera donc la fréquence. Pour plus de commodités, nous nous intéresserons à la fréquence des erreurs. Sachant que la population compte 10 sujets et que 30 % ont réussi, on peut en déduire que :
a) 10*30 %= 3 sujets ont réussi le test dans la population.
b) Et donc 10-3=7 sujets ont échoué.

Sachant que l'échantillon est de 3 sujets et qu'un seul a échoué, on en déduit que 3-1=2 sujets ont réussi. On peut synthétiser cela dans le tableau suivant (le complément est obtenu par différence entre la population et l'échantillon).

	Échec.	Réussite.	Total.
Échantillon.	1	2	3
Reste après échantillonnage.	6	1	7
Population.	7	3	10

L'espace des échantillons est donc composé de tous les échantillons de 3 éléments dans une population de 10 éléments, soit 120 échantillons :

$$\left(\frac{N}{n}\right) = \frac{N!}{n!(N-n)!} = \frac{10!}{3!(10-3)!} = \frac{10*9*8*7*6*5*4*3*2*1}{3*2*1*7*6*5*4*3*2*1} = 120$$

La combinatoire s'obtient de la même façon que pour les variables numériques. L'espace des échantillons est alors le suivant :

I	E	E	E	E	E	E	E	R	R	R
1	E	E	E							
2	E	E		E						
3	E	E			E					
4	E	E				E				
5	E	E					E			
6	E	E						R		
7	E	E							R	
8	E	E								R
9	E		E	E						
10	E		E		E					
11	E		E			E				
12	E		E				E			
13	E		E					R		
14	E		E						R	
15	E		E							R
16	E			E	E					
17	E			E		E				
18	E			E			E			
19	E			E				R		
20	E			E					R	
21	E			E						R
22	E				E	E				
23	E				E		E			
24	E				E			R		
25	E				E				R	
26	E				E					R
27	E					E	E			
28	E					E		R		
29	E					E			R	
30	E					E				R
31	E						E	R		
32	E						E		R	
33	E						E			R
34	E							R	R	
35	E							R		R
36	E								R	R
37		E	E	E						
38		E	E		E					
39		E	E			E				
40		E	E				E			

I	E	E	E	E	E	E	E	R	R	R
41		E	E					R		
42		E	E						R	
43		E	E							R
44		E		E	E					
45		E		E		E				
46		E		E			E			
47		E		E				R		
48		E		E					R	
49		E		E						R
50		E			E	E				
51		E			E		E			
52		E			E			R		
53		E			E				R	
54		E			E					R
55		E				E	E			
56		E				E		R		
57		E				E			R	
58		E				E				R
59		E					E	R		
60		E					E		R	
61		E					E			R
62		E						R	R	
63		E						R		R
64		E							R	R
65			E	E	E					
66			E	E		E				
67			E	E			E			
68			E	E				R		
69			E	E					R	
70			E	E						R
71			E		E	E				
72			E		E		E			
73			E		E			R		
74			E		E				R	
75			E		E					R
76			E			E	E			
77			E			E		R		
78			E			E			R	
79			E			E				R
80			E				E	R		

I	E	E	E	E	E	E	E	R	R	R
81			E				E		R	
82			E				E			R
83			E					R	R	
84			E					R		R
85			E						R	R
86				E	E	E				
87				E	E		E			
88				E	E			R		
89				E	E				R	
90				E	E					R
91				E		E	E			
92				E		E		R		
93				E		E			R	
94				E		E				R
95				E			E	R		
96				E			E		R	
97				E			E			R
98				E				R	R	
99				E				R		R
100				E					R	R
101					E	E	E			
102					E	E		R		
103					E	E			R	
104					E	E				R
105					E		E	R		
106					E		E		R	
107					E		E			R
108					E			R	R	
109					E			R		R
110					E				R	R
111						E	E	R		
112						E	E		R	
113						E	E			R
114						E		R	R	
115						E		R		R
116						E			R	R
117							E	R	R	
118							E	R		R
119							E		R	R
120								R	R	R

Une fois que l'espace des échantillons est déterminé, on fait la distribution des échantillons sur la statistique de test dans notre exemple, la fréquence des échecs. Notre distribution d'échantillonnage est la suivante :

Fréquences d'échec.	*Effectifs.*	*Fréquences.*	*fréq. Cum.*
0,00	1,00	0,01	0,01
0,33	21,00	0,18	0,18
0,67	63,00	0,53	0,71
1,00	35,00	0,29	1,00
Total	120	1	

Situer l'échantillon dans la distribution d'échantillonnage.

Puisque notre statistique d'échantillonnage est la fréquence des échecs, on cherche à savoir si l'échantillon est atypique du côté des valeurs faibles. Nous devons donc lire dans ce tableau la proportion des échantillons qui ont une fréquence d'échec inférieure ou égale à 1/3. Cette proportion nous est donnée dans la colonne des fréquences cumulées, sur la ligne correspondant à 1/3 (0,33) : p_{obs}= 0,18.

Comparaison au seuil repère et formulation de la conclusion.

Nous devons prendre, ici un seuil unilatéral, puisqu'on cherche à savoir si l'échantillon est atypique du côté des valeurs faibles. On peut voir que p_{obs}>.025. On ne peut donc pas dire que l'échantillon est atypique du côté des valeurs faibles. D'un point de vue psychologique, cela veut dire que les mathématiciens n'échouent pas moins souvent que l'ensemble de la population. Symétriquement, on ne peut pas dire qu'ils réussissent plus souvent

Corrigé de l'exercice 10.

Choix du modèle d'échantillonnage.

Le modèle d'échantillonnage reste le même qu'à l'exercice 5.

Détermination de la distribution d'échantillonnage.

Nous sommes dans le cadre d'un échantillonnage dans une population. La distribution d'échantillonnage exacte est donc la distribution hypergéométrique. La distribution binomiale convenant pour l'échantillonnage dans une distribution (population parente de taille infinie).

Nous allons donc utiliser la formule de la distribution hypergéométrique pour calculer la distribution d'échantillonnage.

$$p_k = \frac{\binom{A}{k}\binom{N-A}{n-k}}{\binom{N}{n}} = \frac{A!(N-A)!n!(N-n)!}{k!(n-k)!(A-k)!(N-A-(n-k))!N!}$$

Concrètement, on pose un tableau pour calculer les différentes quantités nécessaires à l'application de la formule. Nous cherchons les valeurs de p pour k=0 ; k=1 ; k=2 ;k=3

(de manière générale pour chaque valeur de k ≤ n). Notre tableau va donc comporter 4 lignes.

k	n-k	A-k	N-A-(n-k)	N!	A!	(N-A)!	n!	(N-n)!	k!	(n-k)!	(A-k)!	N-A-(n-k)!	pk
0	3	7	0	3628800	5040	6	6	5040	1	6	5040	1	0,0083
1	2	6	1	3628800	5040	6	6	5040	1	2	720	1	0,1750
2	1	5	2	3628800	5040	6	6	5040	2	1	120	2	0,5250
3	0	4	3	3628800	5040	6	6	5040	6	1	24	6	0,2917

Situer l'échantillon dans la distribution d'échantillonnage.

La proportion que nous cherchons est la proportion des échantillons pour lesquelles on observe 0 ou une erreur. Il faut donc additionner p0 et p1, puisque nous n'avons pas là une distribution cumulée, soit : 0,0083+0,175= 0,1833. Cette proportion est, bien entendu, la même que dans l'exercice 9, puisque la distribution hypergéométrique est la distribution d'échantillonnage exacte de l'échantillonnage dans une population sur une variable nominale.

Comparaison au seuil repère et formulation de la conclusion.

On se reportera à l'exercice 5 pour les commentaires de cette étape.

Corrigé de l'exercice 11.

Choix du modèle d'échantillonnage.

La question posée est toujours une question de typicalité d'un échantillon dans une population. Le modèle d'échantillonnage est donc toujours l'approche combinatoire.

Détermination de la distribution d'échantillonnage.

Ici, nous changeons de point de vue, ce qui dans les exercices 5 et 6 était la population, est regardé maintenant comme un échantillon dans une population plus large de taille infinie. Cela relève donc d'un échantillonnage dans une distribution. La distribution d'échantillonnage exacte est donc la distribution binomiale. La statistique de test reste la même : la fréquence .

Si les sujets ont répondu comme le feraient des sujets répondant au hasard, la proportion de sujets ayant réussi devrait être égale à la proportion de sujets ayant échoué, soit 50 %. Notre question peut donc être reformulée de la façon suivante : L'échantillon de 10 sujets pour lequel on observe une fréquence d'échecs de 30% est-il atypique d'une distribution dont la fréquence des échecs est de 50 %.

Concrètement, Il va falloir calculer pour chaque valeur de k (de 0 à 10) :

1) La combinatoire de n et de k (c'est le nombre d'échantillons possibles).

2) Puis élever à la puissance k la valeur de P, c'est-à-dire la proportion de l'élément considéré dans la distribution de référence.
3) On élève ensuite à la puissance n-k la valeur de Q, c'est-à-dire la proportion complémentaire de P soit Q=1-P.
4) On fait ensuite le produit des quantités calculées aux points 1, 2 et 3. Ce qui nous donne la formule suivante pour p_k.

$$pk = \binom{n}{k} P^k Q^{n-k}$$

La distribution d'échantillonnage est donc la suivante :

n	k	$\binom{n}{k}$	P^k	Q^{n-k}	p_k	p_k cum. G	p_k cum. D
10	0	1	1,0000	0,0010	0,001	0,001	1,000
10	1	10	0,5000	0,0020	0,010	0,011	0,999
10	2	45	0,2500	0,0039	0,044	0,055	0,989
10	3	120	0,1250	0,0078	0,117	0,172	0,945
10	4	210	0,0625	0,0156	0,205	0,377	0,828
10	5	252	0,0313	0,0313	0,246	0,623	0,623
10	6	210	0,0156	0,0625	0,205	0,828	0,377
10	7	120	0,0078	0,1250	0,117	0,945	0,172
10	8	45	0,0039	0,2500	0,044	0,989	0,055
10	9	10	0,0020	0,5000	0,010	0,999	0,011
10	10	1	0,0010	1,0000	0,001	1,000	0,001

Situer l'échantillon dans la distribution d'échantillonnage.

Dans cet exemple particulier, P et Q sont égaux, les distributions cumulées à droite et à gauche sont donc symétriques.

Si je considère le nombre de réussites, je cherche une typicalité du côté des valeurs élevées. Je vais donc regarder la distribution cumulée à droite des fréquences. La proportion d'échantillons une fréquence de réussite de 70% ou plus est de 0,172.

Si je considère les échecs, je cherche dans ce cas à évaluer la typicalité du côté des valeurs faibles. Je vais donc regarder la distribution cumulée à gauche. La proportion d'échantillons ayant une fréquence d'échecs de 30% ou moins est de 0,172.

Comparer la proportion observée au seuil repère.

Quelque soit le point de vue qu'on prenne, la proportion observé est supérieur au seuil repère de .025. On ne peut donc pas dire que l'échantillon des 10 sujets soit atypique d'une distribution dont la fréquence des réussites ou des échecs (puisque dans ce cas la proportion est la même) est de 50 %. On ne peut donc pas dire que les réponses des sujets diffèrent de réponses au hasard.

Corrigé de l'exercice 12.

Choix du modèle d'échantillonnage.

Le modèle d'échantillonnage reste le même que dans l'exercice 11.

Détermination de la distribution d'échantillonnage.

On nous demande dans cet exercice d'utiliser une distribution d'échantillonnage approchée. Celle qui convient à notre exemple est la distribution de X^2 à un degré de

liberté. Elle est une bonne approximation de la distribution hypergéométrique ou binomiale à condition que tous les effectifs théoriques soient supérieurs à 5, ce qui est le cas, et sous réserve de l'application d'une correction de continuité. Dans ce cas, X^2 se calcule de la façon suivante :

$$X^2_{corr} = \sum \frac{\left(|e_{obs} - e_{théo}| - 0,5\right)^2}{e_{théo}}$$

Situer l'échantillon dans la distribution d'échantillonnage.

Nous présentons le calcul de X^2 en quatre temps :

c) Dans notre exemple, la fréquence théorique est de 0,5. Les effectifs théoriques sont donc de N*0,5=10*0,5=5

d) Calcul des écarts bruts, c'est-à-dire la différence entre les effectifs observés et les effectifs théoriques. Pour les sujets ayant échoué, on aura par exemple 7-5=2.

e) Retrancher 0,5 à la valeur absolue de l'écart brut, puis élevé le résultat au carré. Ainsi pour la colonne « échecs » , on aura :

$$\left(|e_{obs} - e_{théo}| - 0,5\right)^2 = \left(|7 - 5| - 0,5\right)^2 = \left(|2| - 0,5\right)^2 = 0,45$$

d) On additionne ensuite les deux X^2 partiels : 0,45+0,45=0,9

	Echecs	*Réussites*	*Total*
Effectifs observés.	7	3	10
Effectifs théoriques.	5	5	10
Ecarts bruts	2	-2	0
X^2	0,45	0,45	0,9

On consulte ensuite la table de X^2 sur la première ligne (un degré de liberté). La valeur la plus proche inférieure au X^2 calculé est .46. On remonte ensuite lire, en tête de colonne, la valeur de p : $0,5<p_{obs}>.30$.

Comparer la proportion observée au seuil repère et formulation de la conclusion.

La valeur de p_{obs} étant supérieur à .05 (seuil bilatéral), on déclare le résultat non significatif. L'échantillon n'est pas atypique d'une distribution dans laquelle on observe 50 % d'erreurs. Autrement dit, les sujets de l'échantillon ne font pas mieux que des sujets qui répondraient au hasard (la conclusion est, bien sûr, la même qu'avec le test exact).

Corrigé de l'exercice 13.

Analyse descriptive.

Cette analyse repose sur la comparaison des deux moyennes. On peut observé que le temps moyen au premier problème (4,774 minutes) est supérieur au temps moyen au

second problème (4,011 minutes). D'un point de vue descriptif, les sujets résolvent plus rapidement le second problème. La différence est cependant faible (moins d'une minute). Aussi peut-on se demander si elle est significative d'un point de vue inférentiel.

Choix du modèle d'échantillonnage.

Dans le cadre de l'inférence combinatoire, la question posée revient à se demander si les temps de résolution dans les deux problèmes sont homogènes. On peut cependant se placer, dans cette expérience, dans le cadre de l'inférence fréquentiste puisqu'on peut considérer que les sujets sont affectés aléatoirement dans chacun des groupes et que par ailleurs, la question posée relève bien du test d'une hypothèse concernant l'effet d'apprentissage dû à la répétition des résolutions. Dans ce cas, nous testons l'absence de différence entre les temps de résolution du premier et du second problème.

Détermination de la distribution d'échantillonnage.

Nous avons ici à comparer deux groupes d'observations dans un protocole structuré par un croisement (groupes appariés). La variable dépendante est une variable numérique. Nous ne connaissons pas la variance parente. La distribution approchée est donc la distribution du t de Student.

Situer l'échantillon dans la distribution d'échantillonnage.

Pour cela, il faut calculer t_{obs} selon la formule suivante :

$$t_{obs} = \frac{(d - \delta_0)}{s / \sqrt{n}}$$

Les paramètres de cette formule sont les suivants :

Comparaison	*d*	*s*	t_{obs}	*n*
T1-T2	0,763	0,970	4,089	27

Comparer la proportion observée au seuil repère et formulation de la conclusion.

On consulte la table à n-1 degrés de liberté, soit 26. La valeur inférieure et la plus proche de tobs est 3,44. Ce qui correspond à un seuil de .001 bien inférieur au seuil-repère de .025. Le test est donc significatif au seuil .001. On peut donc rejeter l'hypothèse nulle et affirmer (en nous fondant sur l'analyse descriptive) que les sujets résolvent le second problème plus rapidement que le premier.

D'un point de vue psychologique, cet accroissement de la vitesse de résolution s'interprète comme un effet d'apprentissage. On peut donc dire que tous groupes confondus, les sujets progressent au fil de leur différente tentative de résolution de

problème. Cela ne veut pourtant pas dire que cet apprentissage soit indépendant de la formation préalablement dispensée, il faudrait pour cela montrer que cet effet se retrouve dans les trois groupes (c'est l'objet des questions de 14 à 16) et qu'il a la même importance (cette question sort du cadre de votre cours). Cela sous-entend également, qu'ili n'y a pas de différence entre les groupes (nous l'examinerons dans les questions 18 et 19).

Corrigé de l'exercice 14.

Analyse descriptive.

Dans le groupe « contrôle », le temps moyen au premier problème est de 4, 8 minutes contre 4,444 minutes au second problème. On peut donc dire, d'un point de vue descriptif, que le second problème est résolu plus vite que le premier.

Choix du modèle d'échantillonnage.

Voir l'exercice 13.

Détermination de la distribution d'échantillonnage.

Voir l'exercice 13.

Situer l'échantillon dans la distribution d'échantillonnage.

La valeur de t_{obs} se calcul de la même façon que dans l'exercice 13.Les paramètres de la formule sont les suivants :

Comparaison	*d*	*s*	t_{obs}	*n*
T1-T2/contrôle	0,356	0,548	1,947	9

Comparer la proportion observée au seuil repère et formulation de la conclusion.

La table se consulte à n-1=8 degrés de liberté . Sur cette ligne, la valeur inférieure la plus proche de t_{obs} est 1,86. Elle correspond au seuil unilatéral de .05. Ce seuil étant supérieur au seuil-repère de .025, le test est non-significatif. On ne peut donc pas rejeter l'hypothèse nulle. On ne peut donc pas dire que les sujets résolvent le second problème plus vite que le premier.

D'un point de vue psychologique, cela signifie que les sujets du groupe « contrôle » ne progressent pas significativement entre le s deux problèmes. Ils ne bénéficient pas de l'effet d'apprentissage observé pour l'ensemble des sujets.

Corrigé de l'exercice 15.

Analyse descriptive.

Dans le groupe « formation », le temps moyen au premier problème est de 4, 8 minutes contre 3,544 minutes au second problème. On peut donc dire, d'un point de vue descriptif, que le second problème est résolu plus vite que le premier.

Choix du modèle d'échantillonnage.

Voir l'exercice 13.

Détermination de la distribution d'échantillonnage.

Voir l'exercice 13.

Situer l'échantillon dans la distribution d'échantillonnage.

La valeur de t_{obs} se calcul de la même façon que dans l'exercice 13. Les paramètres de la formule sont les suivants :

Comparaison	*d*	*s*	t_{obs}	*n*
T1-T2/formation	1,256	1,428	2,637	9

Comparer la proportion observée au seuil repère et formulation de la conclusion.

Comme dans l'exercice 14, la lecture de la table se fera à n-1=8 degrés de liberté. La valeur inférieure la plus proche est 2,31 et correspond au seuil-repère .025. Le résultat est donc significatif. On peut ainsi rejeter l'hypothèse nulle et affirmer que les sujets de ce groupes résolvent le second problème plus vite que le premier. Ces sujets progressent entre les deux problèmes.

Corrigé de l'exercice 16.

Analyse descriptive.

Dans le groupe « didacticiel», le temps moyen au premier problème est de 4, 722 minutes contre 4,044 minutes au second problème. On peut donc dire, d'un point de vue descriptif, que le second problème est résolu plus vite que le premier.

Choix du modèle d'échantillonnage.

Voir l'exercice 13.

Détermination de la distribution d'échantillonnage.

Voir l'exercice 13.

Situer l'échantillon dans la distribution d'échantillonnage.

La valeur de t_{obs} se calcul de la même façon que dans l'exercice 13. Les paramètres de la formule sont les suivants :

Comparaison	d	s	t_{obs}	n
T1-T2/didacticiel	0,678	0,497	4,092	9

Comparer la proportion observée au seuil repère et formulation de la conclusion.
Sur la ligne 8 de la table (comme dans les précédents exercices), la valeur inférieure la plus proche à t_{obs} est 3,36 . On peut lire, en tête de colonne, qu'elle correspond au seuil .005. Ce seuil étant inférieur au seuil-repère .025, On rejete l'hypothèse nulle d'absence de différence. L'analyse descriptive et inférentielle nous conduit donc à affirmer que ces sujets résolvent le second problème plus vite que le premier. Les sujets progressent donc entre les deux problèmes.

Corrigé de l'exercice 17.

Analyse descriptive.
Comme dans les exercices précédents, on calcule dans ces deux groupes d'observations les moyennes et la différence entre les moyennes. Les résultats sont résumés ci-dessous .

	S1	S2	S3	S4	S5	M
T1	5,8	5,7	5,2	5,6	6,5	5,76
T2	5,9	5,0	4,7	5,2	5,2	5,20
d	-0,1	0,7	0,5	0,4	1,3	0,56

Choix du modèle d'échantillonnage.
Voir l'exercice 13.

Détermination de la distribution d'échantillonnage.
La distribution d'échantillonnage exacte est la distribution des moyennes des différences sur l'ensemble des échantillons qu'on peut obtenir en permutant les observations d'un ou plusieurs couples (espace des échantillons).Pour 5 sujets, on a 2^5=32 échantillons possibles.

1) Les 32 échantillons sont déterminés par combinatoire.
2) On calcule ensuite, pour chaque échantillon, le protocole des différences individuelles.
3) Ces différences sont résumées par la moyenne des différences de chaque échantillon. On ordonne ensuite ces moyennes la distribution de ces moyennes constitue la distribution d'échantillonnage.

Nous donnons ci-dessous le tableau de l'espace des échantillons (O = ordre original, I = ordre inverse) .

I	Permutations possibles	Protocole des différences	moyenne	moy ord

1	0	0	0	0	0	-0,1	1	1	0	1,3	0,560	-0,600
2	I	0	0	0	0	0,1	1	1	0	1,3	0,600	-0,560
3	0	I	0	0	0	-0,1	-1	1	0	1,3	0,280	-0,440
4	0	0	I	0	0	-0,1	1	-1	0	1,3	0,360	-0,400
5	0	0	0	I	0	-0,1	1	1	-0	1,3	0,400	-0,400
6	0	0	0	0	I	-0,1	1	1	0	-1	0,040	-0,360
7	I	I	0	0	0	0,1	-1	1	0	1,3	0,320	-0,320
8	I	0	I	0	0	0,1	1	-1	0	1,3	0,400	-0,280
9	I	0	0	I	0	0,1	1	1	-0	1,3	0,440	-0,240
10	I	0	0	0	I	0,1	1	1	0	-1	0,080	-0,200
11	0	I	I	0	0	-0,1	-1	-1	0	1,3	0,080	-0,160
12	0	I	0	I	0	-0,1	-1	1	-0	1,3	0,120	-0,120
13	0	I	0	0	I	-0,1	-1	1	0	-1	-0,240	-0,120
14	0	0	I	I	0	-0,1	1	-1	-0	1,3	0,200	-0,080
15	0	0	I		I	-0,1	1	-1	0	-1	-0,160	-0,080
16	0	0	0	I	I	-0,1	1	1	-0	-1	-0,120	-0,040
17	I	I	I			0,1	-1	-1	0	1,3	0,120	0,040
18	I	I		I		0,1	-1	1	-0	1,3	0,160	0,080
19	I	I			I	0,1	-1	1	0	-1	-0,200	0,080
20	I		I	I		0,1	1	-1	-0	1,3	0,240	0,120
21	I		I		I	0,1	1	-1	0	-1	-0,120	0,120
22	I			I	I	0,1	1	1	-0	-1	-0,080	0,160
23		I	I	I		-0,1	-1	-1	-0	1,3	-0,080	0,200
24		I	I		I	-0,1	-1	-1	0	-1	-0,440	0,240
25		I		I	I	-0,1	-1	1	-0	-1	-0,400	0,280
26			I	I	I	-0,1	1	-1	-0	-1	-0,320	0,320
27	I	I	I	I		0,1	-1	-1	-0	1,3	-0,040	0,360
28	I	I	I		I	0,1	-1	-1	0	-1	-0,400	0,400
29	I	I		I	I	0,1	-1	1	-0	-1	-0,360	0,400
30	I		I	I	I	0,1	1	-1	-0	-1	-0,280	0,440
31		I	I	I	I	-0,1	-1	-1	-0	-1	-0,600	0,560
32	I	I	I	I	I	0,1	-1	-1	-0	-1	-0,560	0,600

On remarquera que la moyenne des moyennes des différences est égale à 0, ce qui correspond bien à l'idée qu'on teste l'hypothèse d'absence de différence entre les deux groupes d'observations.

Situer l'échantillon dans la distribution d'échantillonnage.

La moyenne des différences de notre échantillon est positive. On cherche à savoir si elle est extrême du côté des valeurs élevées. Il nous faut donc la proportion des échantillons supérieurs ou égaux à l'échantillon observé. Cette proportion est de 2/32= 0,0625, puisqu'il n'y a que deux échantillons qui soient supérieurs ou égaux à l'échantillon observé.

Comparer la proportion observée au seuil repère et formulation de la conclusion.

Nous raisonnons sur un seuil unilatéral puisque la question est de savoir si T2≥T1 et non si T2≠T1, ce qui correspondrait à un seuil bilatéral. Notre seuil-repère, comme dans les exercices précédents, est de .025. La proportion observée étant supérieure à ce seuil, le résultat du test est non significatif. On ne peut donc pas dire que les sujets de cet échantillon résolvent plus vite le second problème.

Corrigé de l'exercice 18.

Analyse descriptive.

Nous avons dans cet exercice à comparer deux groupes. Le premier réunit les deux groupes ayant reçu un enseignement (formation ou didacticiel) et le second est le groupe contrôle. D'un point de vue descriptif, on peut remarquer que les deux groupes résolvent le premier problème aussi rapidement l'un que l'autre. Les sujets du groupe « enseignement » résolvent le problème 1 en 4,761 minutes en moyenne. Les sujets du groupe « contrôle » mettent en moyenne 4, 8 minutes à résoudre ce même problème. D'un point de vue descriptif, la différence entre les deux groupes semble donc très faible.

Choix du modèle d'échantillonnage.

Dans le cadre de l'inférence combinatoire, la question posée revient à se demander si les deux groupes sont homogènes du point de vue du temps de résolution du problème 1. on peut également adopter le point de vue fréquentiste dans cet exercice puisque les sujets sont répartis aléatoirement dans les groupes et qu'on cherche à tester l'hypothèse de l'influence de l'enseignement sur la rapidité de résolution de problème.

Détermination de la distribution d'échantillonnage.

Nous avons ici un protocole structuré par une relation d'emboîtement. La variable dépendante est une variable numérique et l'objectif de l'analyse est la comparaison de deux groupes. Nous ne connaissons pas la variance parente. La distribution d'échantillonnage approchée est donc la distribution du t de Student à n'+n''-2 degrés de liberté.

Situer l'échantillon dans la distribution d'échantillonnage.

Pour calculer t_{obs}, nous avons besoin des moyennes, des variances corrigées et des effectifs de chaque groupe. La formule à appliquer est la suivante :

$$t_{obs} = \frac{m© - m''}{s\sqrt{\frac{1}{n©} + \frac{1}{n''}}} \text{ avec } s^2 = \frac{s©^2 (n© - 1) + s''^2 (n'' - 1)}{n© + n'' - 2}$$

Les paramètres de cette formule sont les suivants :

Groupes	*Moyennes*	*Var_{corr}*	*n*
Formation+Didacticiel	4,761	1,878	9
Contrôle	4,800	2,595	18

Comparaison	*d*	*S^2*	*t*
G1-G2	-0,039	2,365	-0,006

Comparer la proportion observée au seuil repère et formulation de la conclusion.
Nous devons consulter la table du t de Student à 9+18-2=25 degrés de liberté. Sur cette ligne de la table, la plus petite valeur qu'on peut lire est plus grande que la valeur observée. Le test est donc non significatif puisque la proportion observée est forcément supérieure au seuil unilatéral .025 (puisque notre question porte sur une différence orientée).
On ne peut donc pas dire que les sujets qui ont suivi un enseignement résolvent plus vite le premier problème que les sujets du groupe contrôle. Cela ne signifie pas que l'enseignement n'est aucun effet sur l'utilisation du logiciel. En effet, nous avons vu dans les exercices précédents que la différence entre ces groupes se traduit par une progression plus importante entre le problème 1 et le problème 2 pour les groupes ayant eu un enseignement.

Corrigé de l'exercice 19.

Analyse descriptive.
La comparaison des moyennes de ces deux groupes montre que la performance dans le premier problème est la même quelle que soit la méthode d'enseignement. Le groupe « formation » résout le problème en 4,8 minutes en moyenne, contre 4,722 minutes en moyenne pour le groupe « didacticiel.

Choix du modèle d'échantillonnage.
Le modèle d'échantillonnage est le même que dans l'exercice 18.

Détermination de la distribution d'échantillonnage.
Comme dans l'exercice 18, nous avons deux groupes indépendants et la variance parente n'est pas connue. La distribution d'échantillonnage sera donc le t de Student à 9+9-2=16 degrés de liberté.

Situer l'échantillon dans la distribution d'échantillonnage.
La formule du t de Student à utiliser est la même qu'à l'exercice 18. Les paramètres de cette formule sont les suivants :

Groupes	*Moyennes*	*Var_{corr}*	*n*

Formation	4,800	1,680	9
Didacticiel	4,722	2,307	9

Comparaison	*d*	*S^2*	*t*
G1-G2	0,078	1,993	0,017

Comparer la proportion observée au seuil repère et formulation de la conclusion.
Le seuil-repère est cette fois un seuil bilatéral puisque la question porte sur une différence non orientée à priori. Sur la ligne 16 de la table du t de Student, la plus petite valeur lue est .54. Le test est donc non significatif, le seuil étant forcément supérieur au seuil-repère. On ne peut donc pas rejeter l'hypothèse nulle, et on admettra que les deux groupes mettent le même temps en moyenne pour résoudre le problème 1.

Corrigés des exercices bis.

Voici le résultat des calculs.

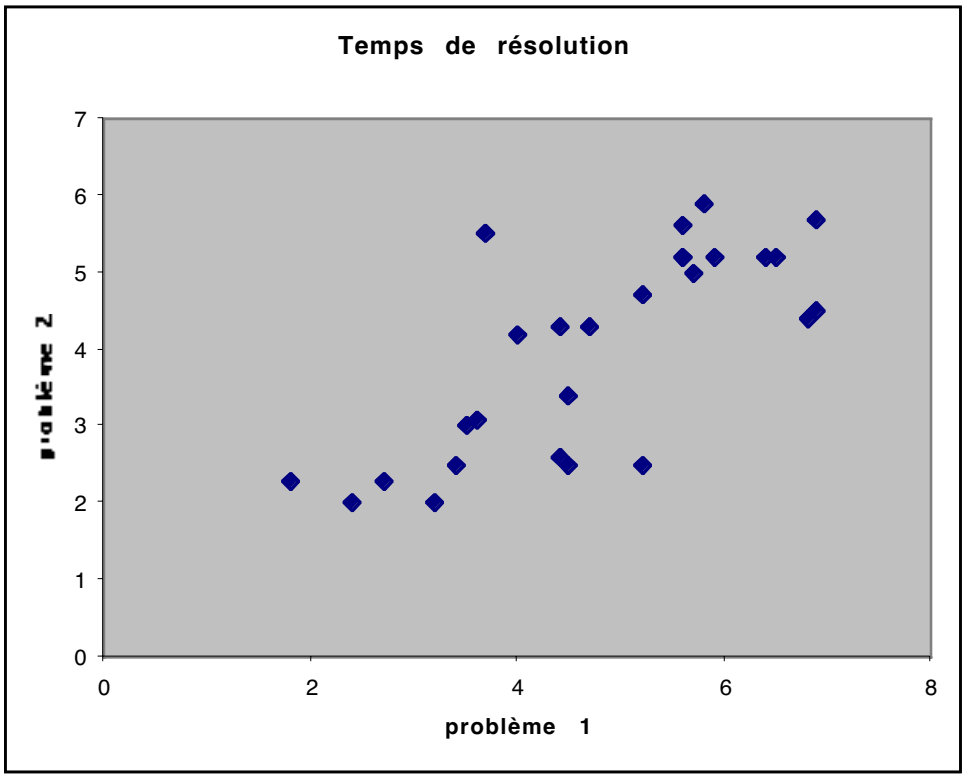
N° d'exercice	Comparaison S*T	Différence des moyennes	Nombre de degrés de liberté	t_{obs}
13 bis	E1, E2: Total	1,926	26	5,710
14 bis	E1, E2: Controle	2,556	8	3,337
15 bis	E1, E2: Didacticiel	1,556	8	3,092
16 bis	E1, E2: Formation	1,667	8	3,780

N° d'exercice	Comparaison S<G>	Différence des moyennes	Nombre de degrés de liberté	t_{obs}
18 bis	Contrôle, enseignement	4,722	25	6,110
19 bis	Didacticiel, Formation	-0,111	16	-0,194

Corrigés.

Corrigé de l'exercice 20.

Analyse descriptive.



On peut voir sur ce graphique que les points représentant les sujets s'organisent approximativement le long d'une droite ascendante. Cela montre une liaison positive

entre les variables. Le coefficient de corrélation est de 0,753, ce qui montre une corrélation positive importante.

Analyse inférentielle.

Choix du modèle d'échantillonnage.

La question posée relève de l'étude de la liaison entre deux variables. Sur le plan descriptif, on évalue une liaison entre deux variables numériques à l'aide du coefficient de corrélation de Bravais-Pearson. Dans le cadre de l'inférence combinatoire, la question revient à se demander si l'échantillon est atypique d'une distribution de r de Bravais-Pearson centrée sur 0. Dans le cadre de l'inférence fréquentiste, cela revient à tester l'hypothèse nulle, c'est-à-dire un coefficient de correlation égal à 0 dans la population parente.

Détermination de la distribution d'échantillonnage.

Choix de la statistique d'échantillonnage.

Dans ce cas, la statistique d'échantillonnage est le r de Bravais-Pearson.

Détermination de la distribution d'échantillonnage.

Comme nous ne disposons pas de la distribution exacte des coefficients de correlation, nous allons utiliser la distribution approchée du t de student à n-2 degrés de liberté en utilisant la statistique T.

Situer l'échantillon dans la distribution d'échantillonnage.

On calcule pour cela la statistique t de la manière suivante :

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,753\sqrt{27-2}}{\sqrt{1-0,753^2}} = 5,722$$

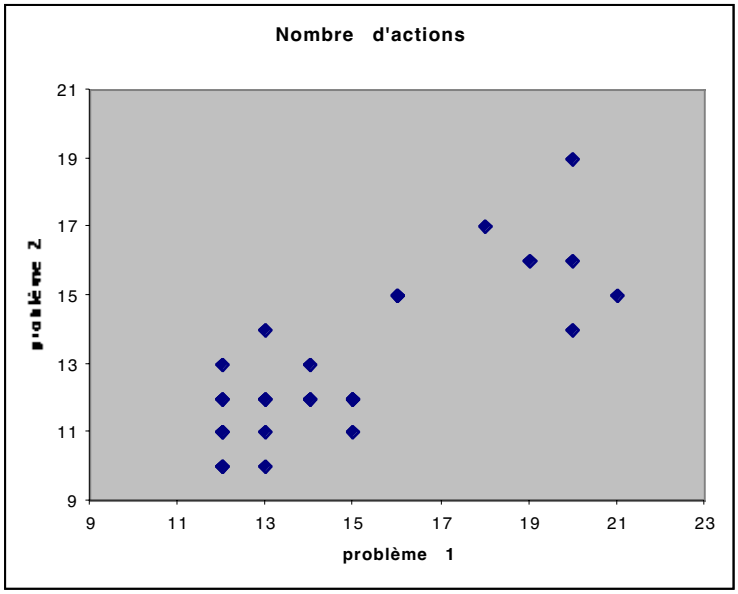
Puis on situe cette valeur de T dans la table du t de Student, à la ligne n-2. Dans notre exemple, n-2=27-2=25. Dans cette table, la valeur la plus proche inférieure à la valeur observée est 3,47. La proportion p est donc inférieure à .001 (seuil unilatérale).

Comparer la proportion observée au seuil repère.

Nous avons observé p<.001 ce qui est bien inférieur au seuil repère unilatéral .025. Le test est donc significatif. On peut donc affirmer, d'un point de vue inférentiel que le coefficient de correlation est non nul (inférence fréquentiste) ou que l'échantillon est atypique d'une distribution des coefficients de correlation nuls. Il existe donc une liaison linéaire positive entre nos variables.

Corrigé de l'exercice 21.

Analyse descriptive.



L'analyse du nombre d'action fournit des résultats similaires à l'analyse du temps de résolutions.
On peut voir sur ce graphique que la répartition des sujets sur les deux variables est orientée dans le sens d'une liaison positive. Le coefficient de corrélation est de 0,802. La statistique T est alors de :

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}802\sqrt{27-2}}{\sqrt{1-0{,}802^2}} = 6{,}713$$

Cette valeur est significative au seuil .001. On peut donc dire que l'échantillon est atypique d'une distribution de coefficients de corrélation nuls. Il existe donc bien une corrélation linéaire positive entre le nombre d'actions réalisées au premier problème et le nombre d'actions réalisées au second problème.

Corrigé de l'exercice 22.

Analyse descriptive.: Dans cet exemple, les individus statistiques sont les couples mère-fille. Nous avons deux variables nominales et un protocole bivarié structuré par un croisement. L'analyse descriptive consiste dans ce cas à calculer le carré moyen de contingence.

	Filles			
Effectifs observés	Ouvrières	Employées	Cadres	Total

Mères Ouvrières	25	13	11	49
Mères Employées	10	23	12	45
Mères Cadres	11	12	13	36
Total	46	48	36	130

Effectifs théoriques	Ouvrières	Employées	Cadres	Total
Ouvrières	17,338	18,092	13,569	49
Employées	15,923	16,615	12,462	45
Cadres	12,738	13,292	9,969	36
Total	46	48	36	130

Ecarts bruts	Ouvrières	Employées	Cadres	Total
Ouvrières	**7,662**	-5,092	-2,569	0
Employées	-5,923	**6,385**	-0,462	0
Cadres	-1,738	-1,292	**3,031**	0
Total	0	0	0	0

	Ouvrières	Employées	Cadres	Total
Ouvrières	3,385	1,433	0,486	5,305
Employées	2,203	2,453	0,017	4,674
Cadres	0,237	0,126	0,921	1,284
Total	5,826	4,012	1,425	11,263

$$\phi^2 = \frac{11,263}{130} = 0,08$$

Commentaires.:L'hypothèse de la mobilité sociale correspond à l'hypothèse nulle. Selon cette hypothèse la CSP des filles ne dépend pas de la CSP des mères. À l'inverse l'hypothèse de la reproductibilité sociale suppose que les sur-représentations dans le tableau soient organisées selon la diagonale et que l'on puisse affirmer une relation entre les CSP des mères et des filles. On peut observer dans cette analyse que les sur-représentations sont toutes situées dans la diagonale (valeur en gras dans le tableau des écarts bruts). Il existe donc d'un point de vue descriptif une tendance pour les mères et les filles à appartenir à la même CSP. Le tableau de X^2 montre que ce sont surtout les catégories « ouvrières » et « employées » qui contribuent à la liaison, les contributions des « cadres étant moindres.

Bibliographie.

Plus de cours ?

Beaufils B. (1998) Statistiques appliquées à la psychologie : tome 2, Statistiques inférentielles. Bréal, collection " lexifac ".

Howell D.C. (1998) Méthodes statistiques en sciences humaines, Bruxelles : De Boeck Université, CollectionMéthodes en sciences humaines.

Rouanet H., Le Roux B., Bert M.C. (1987) Statistique en sciences humaines, procédures naturelles, Paris, Dunod.

Plus d'entrainement ?

Gueguen N. (2001) Statistique pour psychologue : Cours et exercices, Paris, Dunod, collection " Psycho sup ".

Rouanet H., Le Roux B. (1995) Statistiques en sciences humaines : Exercices et solutions, Paris, Dunod, collection " Psycho sup ".

Et des liens....

Encore des cours (en français) !

Intitiation aux méthodes statstistiques :
http://ibm2.cicrp.jussieu.fr/grasland/STAT98/STAT98.htm

SEL : http://www.inrialpes.fr/sel/telecharger.html

Statnet : http://www.agro-montpellier.fr/cnam-lr/statnet/

Mas11 : http://nte-serveur.univ-lyon1.fr/nte/immediato/math2002/Mass11/cours/cours_table.htm

Fiches de statistique : http://hpa.free.fr/Fiches_de_Stat.htm

Encore des cours (en anglais)!

HyperStat Online : http://davidmlane.com/hyperstat/index.html

The Statistics Homepage : http://www.statsoft.com/textbook/stathome.html

StatNotes: http://www2.chass.ncsu.edu/garson/pa765/statnote.htm

Introductory Statistics: http://www.psychstat.smsu.edu/introbook/sbk00.htm

Et si l'ordinateur vous aidait ?

Statview 5 Demos & Downloads: http://www.statview.com/product/demo.shtml

Foire aux questions.

Voici en vrac, quelques questions parmi les plus fréquemment posées. Si vous ne trouvez pas de réponses à votre question, n'hésitez pas à contacter l'enseignant. Vous pouvez également consulter la foire aux questions de l'EC "statistiques descriptives".

Comment interpréter une différence "statistiquement significative" à un seuil de .01 ?

Concrètement cela veut dire que 1 % des échantillons présentent une différence de moyennes supérieure ou égale. Cela ne veut pas dire qu'on a 99 % de chances qu'il y ait une réèlle différence entre les moyennes des deux population, puisque la distribution des différences est centrée sur 0 (on teste l'hypothèse d'absence de différence).

On peut en revanche interpréter ce résultat en disant qu'on a 1 % de chance d'observer une telle différence dans une distribution des différences centrée sur 0 (H_0). Dans ce cas, on se situe dans une approche fréquentiste et la proportion est interprétée en termes de probabilités.

Une formulation équivalente consiste à dire que s'il n'y a pas de différence entre les moyennes des populations (H0) la possibilité d'observer une telle différence serait de 1 % ou moins, puisque lorsque la différence augmente, la proportion d'échantillons correspondant diminue. Vous pouvez en effet remarquer que sur les tables statistiques comme le t de Student, lorsque les valeurs de p diminuent, les valeurs de t augmentent. Un tel résultat montre qu'on peut raisonnablement rejeter l'hypothèse nulle et admettre l'existence d'une différence. Il ne dit cependant rien de l'intérêt pratique de cette différence.

Pourquoi un test significatif à un seuil unilatéral donné, est-il significatif pour tous les seuils supérieurs, et non-significatif pour les seuils inférieurs" ?

Dire qu'un résultat est significatif à un seuil donné, cela veut effectivement dire que ce résultat est significatif pour tous les seuils supérieurs et non significatifs pour les seuils inférieurs. Pour le comprendre, il faut regarder comment est construite la table du t de Student (c'est vrai aussi pour le X^2). Si vous lisez en tête de colonne les valeurs de p, vous pouvez voir que ces valeurs sont rangées en ordre décroissant. En même temps, les valeurs de t sont rangées sur chacune des lignes en ordre croissant. Autrement dit si p diminue, t augmente. Lorsqu'on annonce qu'une valeur de t_{obs} est significative à un seuil donné, on est en train de dire que la valeur observée est supérieure à la valeur de la table pour ce seuil. Dans la mesure où les valeurs de la table sont rangées de façon croissante, cela signifie donc aussi que la valeur de t observée est significative pour toutes les valeurs de p supérieures, c'est-à-dire les valeurs de p situées à gauche de notre

seuil. Pour les valeurs de p inférieures, elle ne l'est pas puisque les valeurs de la table sont forcément supérieures au t observé (il faut se rappeler que les valeurs de t sont rangées en ordre croissant dans la table), le résultat est donc non significatif. Pour bien comprendre cela, il faut se rappeler que la distribution de t est centrée sur 0 et que les valeurs de p dans la table ne représentent rien d'autre que la proportion d'échantillons pour lesquels la différence est égale à la valeur de la table. La distribution de t est approximativement normale. Donc lorsque la différence augmente, la proportion d'échantillons correspondant diminue.

Lorsqu'on utilise le t de Student ou X2, est ce qu'on se place automatiquement dans le cadre de l'inférence fréquentiste ?

Non, entre l'approche combinatoire et l'approche fréquentiste, il n'y a pas de différences de méthodes, mais des différences de présupposés théoriques dont vont dépendre les conclusions possibles. Le t de Student et le X2 sont des tests approchés. On les utilise lorsqu'on ne connaît pas la population de référence (et donc qu'on ne peut construire l'espace des échantillons). Dans ces deux tests, la distribution d'échantillonnage nous est donnée avec la table statistique.

Dans l'approche fréquentiste, en quoi les conclusions sur le rejet ou non de l'hypothèse nulle sont-elles assymétriques?

Dans le cadre de l'inférence fréquentiste, les conclusions ne sont effectivement pas symétriques. On peut interpréter un résultat significatif au seuil .01 en disant qu'on a 1 % de chance d'observer une telle différence dans une distribution des différences centrée sur 0 (H_0). Dans ce cas, la proportion est interprétée en termes de probabilités. Cette probabilité n'est cependant jamais nulle. Il reste donc toujours une chance (même infime) que l'hypothèse nulle soit vraie.

Pourquoi dit-on qu'il faut utiliser le t de Student quand la variance est inconnue puisqu'on peut toujours trouver une variance à partir d'une distribution ?

Quand on dit que le t s'utilise lorsqu'on ne connaît pas la variance, on parle de la variance de la population parente et non de la variance de la distribution. Si vous connaissez la variance de la population parente, vous utilisez la statistique Z, sinon vous utilisez t pour lequel la variance de la population parente est remplacée par s.

Consulte-t-on la table du t de Student de la même façon pour un t positif et un t négatif?

Oui, le t de Student étant distribué selon une loi normale, la distribution est symétrique par rapport à la moyenne 0, et donc la table des différences négatives est la même que la table des différences positives. Pour lire la table, on ne considèrera que la valeur absolue de t.

Formulaires

Statistiques	Formules
Nombre de permutation de *n* objets	$nbre.permutations = n!$
Nombre de combinaisons de n objets dans un ensemble de N éléments	$\left(\frac{N}{n}\right) = \frac{N!}{n!(N-n)!}$
Nombre d'arrangement de n objets dans un ensemble de N éléments	$\left(\frac{N}{n}\right) = \frac{N!n!}{n!(N-n)!} = \frac{N!}{(N-n)!}$
Comparaison à une moyenne connue	$z_{obs} = \frac{(m-\mu)}{\sigma/\sqrt{n}}$
Comparaison de groupes indépendants	$z_{obs} = \frac{m© - m''}{\sigma_0\sqrt{\frac{1}{n©} + \frac{1}{n''}}}$
Comparaison à une moyenne connue	$t_{obs} = \frac{(m_{obs} - \mu_0)}{s/\sqrt{n}}$
Comparaison de groupes appariés	$t_{obs} = \frac{(d-\delta_0)}{s/\sqrt{n}}$
Comparaison de groupes indépendants	$t_{obs} = \frac{m© - m''}{s\sqrt{\frac{1}{n©} + \frac{1}{n''}}}$
Khi-deux corrigé	$\chi^2 = \sum \frac{(\lvert obs - théo\rvert - 0{,}5)^2}{théo}$
Khi-deux	$\chi^2 = \sum \frac{(obs - théo)^2}{théo}$
Phi-deux	$\Phi^2 = \chi^2 / n$
Test de Mac Nemar	$\chi^2_{corr} = \frac{(\lvert A - D\rvert - 1)2}{A + D}$
Coefficient de corrélation de Bravais-Pearson	$r = \frac{\sum (x_i - m_x)(y_i - m_y)}{\sqrt{\sum (x_i - m_x)^2 \sum (y_i - m_y)^2}}$ $= \frac{\sum xiyi - \frac{\sum x \sum y}{n}}{\sqrt{(\sum x_i^2 - (\sum x)^2/n)(\sum x_i^2 - (\sum y_i^2 - (\sum y)^2/n)}}$
Inférence sur un coefficient de corrélation.	$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

Distribution hypergéométrique.

	-	+	Total
Echantillon	k	n - k	n
Complément	A - k	N - A - (n - k)	N - n
Population	A	N-A	N

Proportion d'échantillons de k éléments de la première catégorie.

$$pk = \frac{\binom{A}{k}\binom{N-A}{n-k}}{\binom{N}{n}} = \frac{A!(N-A)!n!(N-n)!}{k!(n-k)!(A-k)!(N-A-(n-k))!N!}$$

Distribution binomiale.

P est la proportion d'éléments de la catégorie considérée dans la population et Q la proportion complémentaire.

Proportion d'échantillons de n éléments ayant k éléments de la catégorie considérée :

$$p_k = \binom{n}{k} P^k Q^{n-k}$$