Fiches Travaux Dirigés
Statistiques inférentielles

CQLS
cqls@upmf-grenoble.fr
http://cqls.upmf-grenoble.fr/

Remarques autour du cours

- **Points-clés de ce cours** : Voici quelques éléments qui font l'originalité de ce cours qui s'adresse tout particulièrement à des étudiants n'ayant pas nécessairement un bagage mathématique très imposant :
  1. *Le **langage mathématique** (utile pour interpréter les résultats des mathématiciens) avant les **techniques mathématiques** (utiles pour développer de nouveaux résultats mathématiques)*
  2. *Une **A**pproche **E**xpérimentale des **P**robabilités (plus intuitive) introduite pour mieux décoder les résultats de l'**A**pproche **M**athématique des **P**robabilités classiquement enseignée*
  3. *Le **cadre asymptotique** (i.e. nombre de données observées suffisamment grand) bien plus réaliste en pratique avant le **cadre Gaussien** (i.e. l'origine des données est d'un type connu mais particulier)*
  4. Outil d'aide à la décision présenté sous sa forme la plus pratique et universelle à savoir la **p-valeur**.
- **Esprit du cours** : Nous avons conscience qu'avoir fait le choix d'introduire un système de notation plutôt lourd mais avant tout précis, est un point qui peut effrayer l'étudiant lors des premiers cours et séances de T.D.. Cependant, il faut souligner qu'un cours alternatif classique (que nous avons déjà expérimenté avant la mise en place de celui-ci) s'appuie essentiellement sur l'**A**pproche **M**athématique des **P**robabilités nécessitant un niveau mathématique bien supérieur à celui requis dans ce cours. Le cours classique de Probabilités et Statistique était alors beaucoup plus orienté sur les Probabilités et la partie Statistique était principalement évoquée sur un plan méthodologique (avec pour conséquence un très grand risque de mauvaise utilisation). La raison principale est que techniquement parlant les probabilités concernant les variables aléaloires discrètes sont plus accessibles que celles concernant les variables aléatoires continues. Et pourtant, ce sont ces dernières qui sont le plus souvent intéressantes dans un contexte Statistitique. Bien connue des développeurs d'outils statistiques (que nous sommes), l'**A**pproche **E**xpérimentale des **P**robabilités n'est pas souvent enseignée alors qu'elle est accessible (même pour un étudiant "non matheux") puisqu'elle s'appuie sur le sens intuitif des probabilités dont nous semblons tous disposer (dans cette société où les jeux de hasard sont très appréciés). Dans ce contexte, il n'y a notamment aucune différence de traitement dans la façon d'aborder le comportement aléatoire d'une variable aléatoire qu'elle soit discrète ou continue (à la différence de l'approche classique). De plus, les difficultés mathématiques utilisées dans cette approche se limitent tout au plus à celles rencontrées lors du cours de Statistique Descriptive (enseigné généralement l'année précédente). Grâce à ces facilités, nous avons pu atteindre notre d'objectif de présenter la règle de décision d'un test d'hypothèses (nom donné par les matheux à l'outil d'aide à la décision) en fonction de la notion fondamentale sur un plan pratique de **p-valeur** ou valeur-p (traduction de p-value en anglais). Au niveau du langage mathématique, nous avons introduit un système de notation afin de rendre accessible cette approche aux étudiants. Si un étudiant nous fait confiance, il pourra se fixer comme objectif principal de maîtriser dans un premier temps ces nouvelles notations alourdies volontairement dans un but de précision puis dans un second temps de les simplifier (comme le font la plupart des mathématiciens) dès lors que son niveau de compréhension sera satisfaisant. Soulignons que l'**A**pproche **E**xpérimentale des **P**robabilités est spécifiquement adaptée à l'utilisation de l'ordinateur. Nous nous appuierons sur le logiciel `R` (libre, gratuit et accessible sous toutes les plateformes). Pour conclure, nous espérons que ce cours se fera malgré tout avec le *sourire* et *plein de bonne humeur*.
- **Les documents de cours** : Il y a trois documents proposés dans ce cours :
  1. Les supports de cours en amphithéatre (version courte imprimable et version longue à ne consulter qu'en mode présentation).
  2. Le document de T.D. qui s'articule avec les supports de cours.
  3. Le polycopié de cours rassemblant les informations importantes du cours.

  Tous les documents sont disponibles sur le site :

  `http://cqls.upmf-grenoble.fr`

  à l'onglet `Stat Inf`.
- **Organisation des documents** : Les séances de T.D. et les cours en amphithéatre s'enchaînent dans l'ordre décrit dans le tableau suivant (bien entendu à adapter selon ses préférences) :

Sujet	Cours	T.D.	Nbre séances
Présentation problématiques		Fiche 1	1
Introduction A.E.P.	Cours 1	Fiche 2	1
Estimation ponctuelle	Cours 2	Fiche 3	1
Intervalle de confiance	Cours 3	Fiche 3	1
Test d'hypothèses (construction)	Cours 4	Fiche 4	1
A.E.P. en graphique (p-valeur)	Cours 5	Fiche A	1
Test d'hypothèses (méthodologie)	Cours 6-8	Fiche 5	3

Fin

FICHE T.D. 1 Présentation des différents outils

Indications préliminaires

- *Proportion* : Une proportion d'individus ayant une caractéristique (d'intérêt) parmi une population de N individus est le nombre d'individus ayant la caractéristique divisé par la taille N de la population.
- *Moyenne* : La moyenne de l'ensemble des N données $\mathbf{z} := (z_1, z_2, \cdots, z_N)$ correspond à la somme des ces données divisée par le nombre total N de données. Elle est usuellement notée et définie par $\overline{z} := \frac{1}{N}\sum_{i=1}^{N} z_i$.
- *Proportion comme une moyenne* : La proportion d'individus ayant une caractéristique parmi une population de N individus peut être vue comme la moyenne $\overline{z}$ des $\mathbf{z} = (z_1, z_2, \cdots, z_N)$ où z_i vaut 1 lorsque l'individu i a la caractéristique et 0 sinon. Autrement dit, une moyenne de valeurs ne valant que 0 ou 1 est une proportion.
- *Nombre moyen* : Soit z_i un nombre (d'objets) associé à tout individu i de la population. Un nombre (d'objets) moyen (par individu de la population) est alors défini comme la moyenne $\overline{z}$ des nombres (d'objets) $\mathbf{z}$.
- *Echantillon* : Indépendament de son procédé de construction, un échantillon de taille n est un "paquet" de n individus extrait parmi l'ensemble de la population totale des N individus. Lorsqu'en particulier, on n'est intéressé que par une variable $\mathcal{Y}$ relative à la population des N indvidus, de manière un peu abusive mais simplifiée, on appelle population l'ensemble des N valeurs $\boldsymbol{\mathcal{Y}} := (\mathcal{Y}_1, \mathcal{Y}_2, \cdots, \mathcal{Y}_N)$ et un échantillon (de $\mathcal{Y}$) l'ensemble des n valeurs $\boldsymbol{y} := (y_1, \cdots, y_n)$ correspondant aux valeurs de $\mathcal{Y}$ pour les individus extraits de la population.
- *Estimation* : une estimation d'un paramètre inconnu θ sur un échantillon $\boldsymbol{y}$ est noté $\widehat{\theta}(\boldsymbol{y})$ s'exprimant littéralement par "estimation (ou plus généralement remplaçant) du paramètre (inconnu) θ calculée à partir du jeu de données $\boldsymbol{y}$" après avoir appliqué les conventions de notation suivantes :
  1. "$\widehat{\ }$" signifie (usuellement en Statistique) estimation (ou plus généralement, remplaçant) de la quantité sur laquelle il se trouve, ici le paramètre inconnu θ à estimer.
  2. "$(\boldsymbol{y})$" exprime la dépendance fonctionnelle (i.e. symbolisée dans le langage mathématique par les parenthèses qui servent à encadrer une valeur d'entrée à appliquer à une fonction afin de retourner une valeur de sortie) pouvant être traduite littéralement par "calculée à partir de l'échantillon $\boldsymbol{y}$". Le "$\widehat{\theta}$" est alors vu comme une fonction retournant en sortie l'estimation de θ lorsqu'en entrée il lui est donné un échantillon $\boldsymbol{y}$.

Fin

Exercice 1 *Entre les deux tours d'une élection présidentielle, un candidat, Max, souhaiterait "rapidement" avoir un a priori sur la proportion d'intentions de vote en sa faveur. On notera* $\boldsymbol{\mathcal{Y}}^{Max} = \left(\mathcal{Y}_1^{Max}, \ldots, \mathcal{Y}_N^{Max}\right)$ *l'ensemble des réponses des N électeurs (où* $\mathcal{Y}_i^{Max}$ *vaut 1 si l'individu i a l'intention de voter pour Max et 0 sinon).*

1. *Déterminez en fonction de* $\boldsymbol{\mathcal{Y}}^{Max}$*, le nombre puis la proportion d'intentions de vote en faveur de Max, notée respectivement* N^{Max} *et* p^{Max}*.*
2. *N étant très grand, quel serait une solution réalisable permettant d'obtenir un remplaçant (i.e. estimation) de* p^{Max}*. Proposez les notations adéquates.*
3. *Deux personnes se proposent d'interroger chacun* $n = 1000$ *électeurs. On notera* $\boldsymbol{y}_{[1]}$ *et* $\boldsymbol{y}_{[2]}$ *ces deux jeux de données recueillis. Les estimations correspondantes sont respectivement de 47% et 52%. Comment interpréter la différence des résultats qui, si on leur fait une confiance aveugle, conduit à deux conclusions différentes ?*
4. *Connaissez-vous d'autres applications nécessitant une estimation d'un paramètre inconnu ?*

Exercice 2 (Présentation des problématiques des produits A et B)

Un industriel veut lancer sur le marché deux produits que l'on nommera Produit A et Produit B . Le Produit A est acheté au plus une fois par mois tandis que le Produit B peut être acheté autant de fois que désiré. Après une étude financière, les services comptables indiquent à cet industriel que pour que le lancement de chacun de ces produits soit rentable, il faut qu'il soit vendu à plus de 300000 exemplaires par mois. La population ciblée par l'industriel est une population de taille $N = 2000000$. L'industriel se demande s'il doit ou non lancer le(s) Produit(s) A et/ou B.

Commençons par introduire quelques notations permettant de décrire le choix d'achat des individus de la population totale (ciblée par l'industriel). Les deux études des Produit A et Produit B étant plutôt similaires, nous noterons donc dans un cadre général $\bullet$ aussi bien à la place de A ou B. Ainsi $\mathcal{Y}_i^\bullet$ représente le nombre de produit(s) $\bullet$ acheté(s) par le $i^{ème}$ $(i = 1, \cdots, N)$ individu de la population totale. L'ensemble des choix d'achat des N individus $(\mathcal{Y}_i)_{i=1,\cdots,N}$ sera noté $\boldsymbol{\mathcal{Y}}^\bullet$. $N^\bullet$ désignera le nombre d'exemplaires de Produit $\bullet$ achetés par les N individus de la population.

1. *Exprimez N^A (resp. N^B) en fonction des $\boldsymbol{\mathcal{Y}}^A$ (resp. $\boldsymbol{\mathcal{Y}}^B$). Exprimez la rentabilité du Produit A (resp. Produit B) en fonction du nombre total N^A (resp. N^B) d'exemplaires du Produit A (resp. Produit B) vendus.*
2. *Même question mais en fonction du nombre moyen (par individu de la population) μ^A (resp. μ^B) d'exemplaires du Produit A (resp. Produit B) en ayant au préalable établi la relation entre μ^A et N^A (resp. μ^B et N^B) et ainsi entre μ^A et $\boldsymbol{\mathcal{Y}}^A$ (resp. μ^B et $\boldsymbol{\mathcal{Y}}^B$). Quelle relation y a-t-il donc entre μ^A et $\overline{\mathcal{Y}^A}$ (resp. entre μ^B et $\overline{\mathcal{Y}^B}$) ?*

   *Les quantités μ^A et μ^B seront appelées **paramètres d'intérêt**.*
3. *Est-il possible pour l'industriel de ne pas se tromper dans sa décision quant au lancement de chaque produit ? Si oui, comment doit-il procéder ? Cette solution est-elle réalisable ?*
4. *Est-il alors possible d'évaluer (exactement) les paramètres d'intérêt ? Comment les qualifieriez-vous par la suite ?*
5. *Une solution réalisable est alors de n'interroger qu'une sous-population de taille raisonnable $n << N$ (ex $n = 1000$). On notera alors $\boldsymbol{y}^\bullet$ le jeu de données (appelé aussi échantillon), i.e. le vecteur des n nombres d'achat $(y_i^\bullet)_{i=1,\cdots,n}$ du produit $\bullet$ des n ($n << N$) individus interrogés.*

   Comment l'industriel pourra-t-il évaluer un remplaçant de $\mu^\bullet$ à partir de son échantillon $\boldsymbol{y}^\bullet$?

   (quelle est la relation entre $\overline{y^\bullet}$, représentant la moyenne empirique des $(y_i^\bullet)_{i=1,\ldots,n}$, et l'estimation $\widehat{\mu^\bullet}\,(\boldsymbol{y}^\bullet)$?)
6. *Quelle est la nature du paramètre d'intérêt μ^A dans le cas où les données ne sont que des 0 et 1 ? Désormais cette moyenne, puisqu'elle bénéficiera d'un traitement particulier, sera notée $\boxed{p^A = \mu^A}$.*

Exercice 3 *Dans le but d'estimer un paramètre d'intérêt inconnu, on dispose d'un échantillon. Nous nous proposons maintenant de préciser plus en détail son procédé de construction.*

1. *Proposez des critères de qualité d'un tel échantillon.*
2. *A quoi correspond la notion de représentativité ?*
3. *Est-il possible de construire un échantillon représentatif d'une (ou plusieurs) caractéristique(s) donnée(s) ?*
4. *Même question sans aucun a priori (i.e. aucune caractéristique fixée).*
5. *Proposez un critère de qualité qui permettra de construire un échantillon le plus représentatif sans aucun a priori.*
6. *Fournissez un (ou plusieurs) procédé(s) d'échantillonnage satisfaisant au critère suivant de représentativité (maximale) sans a priori (RSAP) :*

   Tous les individus de la population totale ont la même chance d'être choisi dans l'échantillon.
7. *Si on répète le procédé d'échantillonnage suivant le critère RSAP et que pour chaque échantillon on évalue l'estimation du paramètre d'intérêt, pensez-vous que les résultats*

seront toujours les mêmes ? Comment qualifie-t-on alors la nature du procédé d'échantillonnage ?

Exercice 4 (Outil pour la problématique des élections)

On se propose d'estimer le paramètre d'intérêt en fournissant un intervalle (ou fourchette, encadrement) obtenu à partir des données. Cet intervalle, appelé intervalle de confiance, est centré en la valeur de l'estimation et sa largeur dépend d'un niveau de confiance que l'on se fixe (généralement plutôt grand, par exemple, 95%*).*

1. *Pensez-vous qu'il soit possible qu'une estimation* $\widehat{p}(\boldsymbol{y})$ *soit égale au paramètre estimé ? Pouvez-vous savoir l'ordre de grandeur de l'écart entre l'estimation et le paramètre inconnu ? Quel niveau de confiance accordez-vous à la valeur d'une estimation (dans notre exemple, 47% et 52% sur deux échantillons) ?*
2. *Si on vous annonce qu'un statisticien sait généralement fournir en plus de l'estimation du paramètre, l'estimation de sa fiabilité mesurée en terme de variabilité attendue, quel est la mission principale d'un intervalle de confiance ? Quelles sont les qualités souhaitées d'un intervalle de bonne confiance (95% par exemple) du paramètre d'intérêt (inconnu) ?*
3. *Compléter les phrases suivantes :*
   (a) *PLUS le niveau de confiance est fort, ... l'intervalle de confiance est petit.*
   (b) *Vue comme un intervalle de confiance de largeur 0, une estimation peut donc être associé à un niveau de confiance ...%.*
4. *Un statisticien construit les intervalles à* 95% *de confiance (via une formule d'obtention étudiée plus tard dans le cours ne faisant pas l'objet) et informe le candidat que les intervalles associés aux estimations 47% et 52% sont respectivement* $[43.90655\%, 50.09345\%]$ *et* $[48.90345\%, 55.09655\%]$. *Les élections effectuées, on évalue* $p^{Max} = 51.69\%$*, qu'en pensez-vous ?*
5. *Si vous avez des difficultés à traduire ce que signifie le niveau de confiance d'un intervalle, comparez-le avec celui que vous accorderiez à une personne qui serait censée dire la vérité avec un niveau de confiance fixé à 95%. Dans le cas de cette personne, comment traduiriez-vous (ou expliqueriez-vous) le concept de niveau de confiance ?*

   Réponse _____

   Parmi toutes les assertions énoncées par cette personne (dont on peut vérifier la véracité ou fiabilité), 95% (en moyenne) seraient censées être justes ou fiables.

_____*Fin*

Remarque

Cet exemple nous aide à appréhender la notion de niveau de confiance ou plus généralement de probabilité d'un événement en l'exprimant comme la proportion parmi toutes (en théorie, on peut imaginer en faire une infinité) réalisations de l'expérience (a priori supposée aléatoire) qui conduisent à ce que l'événement soit vérifié. Ceci nous conduit naturellement vers la notion d'Approche Expérimentale des Probabilités qui sera présentée dans la fiche de Td suivante en complément de l'Approche Mathématique des Probabilités (qui est classiquement présentée dans les cours de Probabilités et Statistique).

FICHE T.D. 2 Introduction aux probabilités

Indications préliminaires

- *Objectif* : L'originalité de ce cours réside essentiellement dans l'axe qui a été choisi pour présenter les probabilités. Dans un cours classique, les développements mathématiques (de nature plutôt technique) sont proposés en priorité en laissant peu de place à l'interprétation des concepts théoriques véhiculés. Cette approche pour introduire les concepts de probabilités sera par la suite appelée **A.M.P.** pour désigner **A**pproche **M**athématique des **P**robabilités. La Statistique (Inférentielle ou Inductive, celle présentée dans ce cours) repose largement sur la théorie des Probabilités, mais de part sa vocation à être largement utilisée par les praticiens sous une forme plutôt méthodologique, il s'ensuit souvent une difficulté pour ces utilisateurs à appréhender les conditions d'applicabilité et les points-clés des outils statistiques qui bien souvent s'expriment en fonction des concepts probabilistes pas toujours faciles à assimiler (compte tenu de leurs aspects mathématiques). Afin de remédier à cet incovénient, nous avons choisi de proposer une approche complémentaire, appelée **A.E.P.** pour désigner **A**pproche **E**xpérimentale des **P**robabilités, qui nous semble plus intuitive car basée sur l'expérimentation et dont la difficulté technique se limite aux outils de Statistique Desciptive présentés en première année (faciles à appréhender par les praticiens motivés surtout lorsqu'ils en ont l'utilité). L'objectif de cette fiche T.D. est essentiellement de faire le lien entre les deux approches **A.M.P.** et **A.E.P.**. Notamment, il sera essentiel de comprendre comment les praticiens pourrons être éclairés via l'**A.E.P.** sur les résultats techniques obtenus grâce à l'**A.M.P.** par les mathématiciens.
- *L'**A.E.P.** en complément de l'**A.M.P.*** : Soit Y une variable aléatoire réelle dont on suppose disposer (via l'**A.E.P.**) d'un vecteur $\mathbf{y}_{[m]} := (y_{[\cdot]})_m := (y_{[1]}, y_{[2]}, \cdots, y_{[m]})$ de m (a priori très grand) réalisations indépendantes entre elles. En théorie, on pourra aussi imaginer disposer du vecteur $\mathbf{y}_{[+\infty]} := (y_{[\cdot]})_{+\infty}$ qui est l'analogue de $\mathbf{y}_{[m]}$ avec $m \to +\infty$. Supposons aussi que $m = 10000$ expériences aient été réalisées et les m composantes de $\mathbf{y}_{[m]}$ aient été stockées dans `R` sous le vecteur nommé `yy`.

Quantité	**A.M.P.**		**A.E.P.** $(+\infty)$		**A.E.P.**		**Traitement** `R`
Probabilité	$\mathbb{P}(Y = a)$	$=$	$\overline{(y_{[\cdot]} = a)}_{+\infty}$	$\simeq$	$\overline{(y_{[\cdot]} = a)}_{m}$	$\overset{R}{=}$	`mean(yy==a)`
Probabilité	$\mathbb{P}(Y \in]a,b])$	$=$	$\overline{(y_{[\cdot]} \in]a,b])}_{+\infty}$	$\simeq$	$\overline{(y_{[\cdot]} \in]a,b])}_{m}$	$\overset{R}{=}$	`mean(a<yy & yy<=b)`
Moyenne	$\mathbb{E}(Y)$	$=$	$\overline{(y_{[\cdot]})}_{+\infty}$	$\simeq$	$\overline{(y_{[\cdot]})}_{m}$	$\overset{R}{=}$	`mean(yy)`
Variance	$\mathbb{V}ar(Y)$	$=$	$\overline{(y_{[\cdot]})^2}_{+\infty}$	$\simeq$	$\overline{(y_{[\cdot]})^2}_{m}$	$\overset{R}{=}$	`var(yy)=sd(yy)^2`
Quantile	$q_Y(\alpha)$	$=$	$q_\alpha\left((y_{[\cdot]})_{+\infty}\right)$	$\simeq$	$q_\alpha\left((y_{[\cdot]})_{m}\right)$	$\overset{R}{=}$	`quantile(yy,alpha)`

Les formules d'obtention des quantités ci-dessus pour les colonnes **A.M.P.** et **A.E.P.** n'ont pas été fournies. Celles concernant l'**A.M.P.** requiert un niveau plutôt avancé en mathématiques et diffèrent selon la nature (discrète ou continue) de Y. Un point fort de l'**A.E.P.** est que les formules d'obtentions ne dépendent pas de la nature de Y et sont normalement déjà connues en 1ère année dans le cours de Statistique Descriptive (pour rappel, voir polycopié de notre cours).
IMPORTANT : L'objectif principal de la fiche T.D. est l'assimilation des concepts décrits dans le tableau ci-dessus.

- *Quelques résultats sur **A.M.P.*** : Soient Y, Y_1 et Y_2 trois variables aléatoires réelles (v.a.r.) et λ un réel.
  Fonction de répartition $F_Y(y) := \mathbb{P}(Y \leq y)$: Dans l'**A.M.P.**, elle permet de calculer, pour tout $a \leq b$:
  $$\mathbb{P}(a < Y \leq b) = \mathbb{P}(Y \leq b) - \mathbb{P}(Y \leq a) = F_Y(b) - F_Y(a).$$
  Moyenne (théorique) : $\mathbb{E}(\lambda \times Y) = \lambda \times \mathbb{E}(Y)$ et $\mathbb{E}(Y_1 + Y_2) = \mathbb{E}(Y_1) + \mathbb{E}(Y_2)$
  Variance : $\mathbb{V}ar(\lambda \times Y) = \lambda^2 \times \mathbb{V}ar(Y)$ et $\mathbb{V}ar(Y_1 + Y_2) = \mathbb{V}ar(Y_1) + \mathbb{V}ar(Y_2)$
  où Y_1 et Y_2 sont en plus supposées indépendantes.

Fin

Exercice 5 (Lancer d'un dé)

1. *Proposer le Schéma de Formalisation pour la variable aléatoire correspondant à un futur*

lancer de dé.

Réponse

— ***Expérience $\mathcal{E}$:*** *Lancer un dé*

— ***Variable d'intérêt :*** *Y la face supérieure du dé*

— ***Loi de proba :*** *$\mathbb{P}(Y=k)=1/6$ avec $k=1,\cdots,6$ (si le dé est équilibré).*

Fin

2. *Quelle expérimentation mettriez-vous en oeuvre pour vérifier qu'un dé est rigoureusement non pipé (i.e. parfaitement équilibré) ? Pensez-vous qu'il existe un tel type de dé ?*
3. ***Application :*** *Un expérimentateur propose l'expérience suivante avec un dé (en théorie vendu) équilibré et un autre dont il a volontairement légèrement déséquilibré une ou plusieurs de ses faces. Les résultats des deux dés sont fournis dans un ordre arbitraire dans les tableaux ci-dessous. Sauriez-vous reconnaître les deux dés et, en particulier, déterminer les probabilités d'apparition des faces (sachant que, pour chaque dé, il n'y a en théorie pas plus de 2 choix possibles pour celles-ci) ? A partir de combien de lancers (m) êtes-vous en mesure de faire votre choix ?*

m	$\overline{(y=1)}_m$	$\overline{(y=2)}_m$	$\overline{(y=3)}_m$	$\overline{(y=4)}_m$	$\overline{(y=5)}_m$	$\overline{(y=6)}_m$	$\overline{(y)}_m$
100	21%	14%	15%	22%	16%	12%	3.34
1000	15.5%	16.8%	17.3%	17.1%	15.9%	17.4%	3.533
10000	16.46%	16.43%	16.45%	17.23%	16.46%	16.97%	3.5171
100000	16.4%	16.52%	16.28%	17.05%	16.83%	16.92%	3.5214
1000000	16.47%	16.52%	16.49%	16.85%	16.77%	16.89%	3.5161

m	$\overline{(y=1)}_m$	$\overline{(y=2)}_m$	$\overline{(y=3)}_m$	$\overline{(y=4)}_m$	$\overline{(y=5)}_m$	$\overline{(y=6)}_m$	$\overline{(y)}_m$
100	13%	13%	16%	21%	23%	14%	3.7
1000	16.1%	18.1%	15.6%	17.3%	18.6%	14.3%	3.471
10000	16.92%	17%	16.47%	16.91%	17.13%	15.57%	3.4704
100000	16.73%	16.64%	16.53%	16.59%	16.88%	16.63%	3.5015
1000000	16.68%	16.66%	16.68%	16.67%	16.71%	16.61%	3.499

4. *Fournir les instructions `R` ayant permis de déterminer les résultats des tableaux précédents.*
5. *Ayant à présent identifié (du moins nous l'espérons !) le dé équilibré, sauriez vous compléter le tableau suivant correspondant à l'éventuelle dernière ligne du tableau précédent lui correspondant :*

m	$\overline{(y=1)}_m$	$\overline{(y=2)}_m$	$\overline{(y=3)}_m$	$\overline{(y=4)}_m$	$\overline{(y=5)}_m$	$\overline{(y=6)}_m$	$\overline{(y)}_m$
∞							

Comment noteriez-vous ces quantités via l'A.M.P. ?

6. *Considérons le dé (théoriquement) équilibré. Observons les expressions dans le tableau ci-dessous obtenues par le mathématicien (A.M.P.). Sauriez-vous les calculer (N.B. : c'est une question personnelle et il est donc possible de répondre NON) ? On rappelle (pour votre culture) les formules d'obtentions de la moyenne (ou espérance) de Y :*

$$\mathbb{E}(Y)=\sum_{k=1}^{6} k\times\mathbb{P}(Y=k)$$

ainsi que celle de la variance

$$\mathbb{V}ar(Y)=\sum_{k=1}^{6}(k-\mathbb{E}(Y))^2\times\mathbb{P}(Y=k)=\mathbb{E}\left(Y^2\right)-\mathbb{E}(Y)^2=\sum_{k=1}^{6}k^2\times\mathbb{P}(Y=k)-\mathbb{E}(Y)^2$$

$\mathbb{P}(Y\in[2,4[)$	$\mathbb{E}(Y)$	$\mathbb{V}ar(Y)$	$\sigma(Y)$	$q_{5\%}(Y)$	$q_{50\%}(Y)$	$q_{95\%}(Y)$
33.33%	3.5	2.9167	1.7078	1	3.5	6

Remarque (pour les amateurs) *: Puisque* $\mathbb{P}(Y=k)=\frac{1}{6}$*, les valeurs du tableau pour* $\mathbb{E}(Y)$*,* $\mathbb{V}ar(Y)$ *et* $q_p(Y)$ *(*$p=$*5%, 50% et 95%) ont simplement été obtenues en appliquant les formules de Statistique Descriptive pour la série de chiffres* $1,2,3,4,5,6$*.*

7. *Comprenons comment ces quantités peuvent être obtenues (ou intreprétées) par l'expérimentateur en les confrontant à ses résultats sur* $m=1000000$ *lancers (A.E.P.). Proposez aussi les instructions* `R` *ayant permis de les construire sachant que ces résultats ont été stockés dans le vecteur* ***yy*** *en* `R`*.*

$\overline{(y\in[2,4[)}_m$	$\overline{(y)}_m$	$\left(\overleftrightarrow{(y)}_m\right)^2$	$\overleftrightarrow{(y)}_m$	$q_{5\%}((y)_m)$	$q_{50\%}((y)_m)$	$q_{95\%}((y)_m)$
33.34%	3.499	2.9145	1.7072	1	3	6

8. *Quelle approche (A.M.P. ou A.E.P.) vous semble être la plus facile à appréhender ? Comprenez-vous les intérêts propres à chacune d'entre elles ?*

Exercice 6 (Somme de deux dés)

1. *Soient* Y_V *et* Y_R *deux variables aléatoires correspondant aux faces de 2 dés (Vert et Rouge) à lancer. Définissons* $S=Y_V+Y_R$ *correspondant à la somme de deux faces. Proposez le Schéma de Formalisation pour* S*.*

Réponse

— ***Expérience*** $\mathcal{E}$ ***:*** *Lancer de 2 dés*
— ***Variable d'intérêt :*** S *la somme des faces supérieures des 2 dés*
— ***Loi de proba :*** $\mathbb{P}(S=k)=$*??? avec* $k=2,\cdots,12$*.*

Fin

2. *Comparez* $\mathbb{P}(S=2)$*,* $\mathbb{P}(S=12)$ *et* $\mathbb{P}(S=7)$*). Sauriez-vous les évaluer ?*
3. *Que peut-on espérer en moyenne sur la valeur de* S *? (cette quantité rappelons-le est notée* $\mathbb{E}(S)$*).*
4. *Un joueur se propose de lancer* $m=5000$ *fois deux dés. A chaque lancer, il note la somme et stocke l'ensemble des informations dans un vecteur noté* `s` *en* `R`*. Voici quelques résultats d'instructions* `R` *:*

```
> s
   [1]  8  8  8  9  5  4  4  4  3  6  7  2  3 10  6  2  6  9  2  9 12  7 10 12
  [25]  3  5  9  6  6  7  7  6  7  8  9  8  7  3  4  9  8 10  5  8  7  6  8  8
...
[4969]  6 10  9  9  9 11  7  7 10  6  6 12  4  9  7  9 10  2  8  9  7  7  7  4
[4993]  8  7 12  8 10 11  6  9
> mean(s==2)
[1] 0.0314
> mean(s==12)
[1] 0.0278
> mean(s==7)
[1] 0.1698
> mean(s)
[1] 7.0062
> var(s)
[1] 5.872536
> sd(s)
[1] 2.423332
```

Pourriez-vous proposer les notations mathématiques (norme CQLS) correspondant aux résultats obtenus dans la sortie `R` *ci-dessus ?*

5. *Cette approche expérimentale confirme-t-elle le résultat du mathématicien affirmant que pour toute modalité* $k=2,\cdots,12$ *de* S*,*

$$\mathbb{P}(S=k)=\begin{cases}\frac{k-1}{36} & si\ k\leq 7\\ \frac{13-k}{36} & si\ k\geq 7\end{cases}$$

*Voici les résultats de l'****A.M.P.*** *présentés dans le tableau suivant (que vous pouvez vérifier si vous avez l'âme d'un mathématicien) :*

$\mathbb{P}(S=2)$	$\mathbb{P}(S=12)$	$\mathbb{P}(S=7)$	$\mathbb{E}(S)$	$\mathbb{V}ar(S)$
2.78%	2.78%	16.67%	7	5.8333

6. *Pourriez-vous aussi vérifier la validité des formules sur l'espérance et variance de la somme de variables aléatoires réelles fournies au début de cette fiche.*

Exercice 7 (Loi uniforme sur l'intervalle unité) *1. Soit Y_1 une variable aléatoire suivant une loi uniforme sur $[0,1]$ (en langage math., $Y_1 \rightsquigarrow \mathcal{U}([0,1])$), correspondant au choix "au hasard" d'un réel dans l'intervalle $[0,1]$. L'objectif est l'évaluation (exacte ou approximative) des probabilités suivantes $\mathbb{P}(Y_1 = 0.5)$ et $\mathbb{P}(0 < Y_1 < 0.5)$, le chiffre moyen $\mathbb{E}(Y_1)$ (espéré), l'écart-type $\sigma(Y_1)$ ainsi que la variance $\mathbb{V}ar(Y_1)$? Parmi ces quantités, lequelles sauriez-vous intuitivement (i.e. sans calcul) déterminer ?*

2. *Via **A.E.P.** : Un expérimentateur réalise cette expérience en choisissant 10000 réels au hasard (par exemple en tapant 10000 fois sur la touche RAND d'une calculatrice). Il stocke les informations dans son logiciel préféré (libre et gratuit) `R` dans un vecteur noté `y1`. Déterminez approximativement les quantités de la première question.*

```
> y1
    [1] 0.6739665526 0.7397576035 0.7916111494 0.6937727907 0.6256426109
    [6] 0.4411222513 0.8918520729 0.4331923584 0.4213763773 0.6879929998
...
 [9991] 0.3117644335 0.1422109089 0.4964213229 0.6349032705 0.3718051254
 [9996] 0.2839202243 0.7170524562 0.7066086838 0.9236146978 0.7250815830
> mean(y1)
[1] 0.4940455
> mean(y1==0.5)
[1] 0
> mean(0.25 <y1 & y1<0.5)
[1] 0.254
> var(y1)
[1] 0.08296901
> sd(y1)
[1] 0.2880434
> sd(y1)^2
[1] 0.08296901
```

3. *Via **A.M.P.** : Un mathématicien obtient par le calcul les résultats suivant pour une variable aléatoire Y représentant un chiffre au hasard dans l'intervalle $[a,b]$ (i.e. $Y \rightsquigarrow \mathcal{U}([a,b])$) :*

   (a) pour tout $a \leq t_1 \leq t_2 \leq b$, $\mathbb{P}(t_1 \leq Y \leq t_2) = \frac{t_2 - t_1}{b-a}$.

   (b) $\mathbb{E}(Y) = \frac{a+b}{2}$

   (c) $\mathbb{V}ar(Y) = \frac{(b-a)^2}{12}$

   Question optionnelle : lesquels de ces résultats sont intuitifs (i.e. déterminables sans calcul) ? Déterminez exactement les quantités de la première question.

```
> 1/12
[1] 0.08333333
> sqrt(1/12)
[1] 0.2886751
```

4. *L'**A.E.P.** confirme-t'elle les résultats théoriques de l'**A.M.P.** ?*

Exercice 8 (Somme de deux uniformes) *1. On se propose maintenant d'étudier la variable $S = Y_1 + Y_2$ où Y_1 et Y_2 sont deux variables aléatoires indépendantes suivant une loi uniforme sur $[0,1]$. Quel est l'ensemble des valeurs possibles (ou modalités) de S ? Pensez-vous que la variable S suive une loi uniforme ? Nous nous proposons d'évaluer (excatement ou approximativement) les probabilités $\mathbb{P}\left(0 < S \leq \frac{1}{2}\right)$, $\mathbb{P}\left(\frac{3}{4} < S \leq \frac{5}{4}\right)$, $\mathbb{P}\left(\frac{3}{2} < S \leq 2\right)$, la moyenne $\mathbb{E}(S)$, l'écart-type $\sigma(S)$ et la variance $\mathbb{V}ar(S)$. Lesquelles parmi ces quantités sont déterminables intuitivement ou via un simple calcul mental ? Etes-vous capable de comparer les trois probabilités précédentes ?*

2. *Via* ***A.E.P.*** *: Un expérimentateur réalise à nouveau l'expérience de choisir 1000 réels entre 0 et 1. Les informations sont stockées dans le vecteur* `y2`*. Déterminez approximativement les quantités de la premire question.*

```
> y2
    [1] 7.050965e-01 7.167117e-01 8.085787e-01 5.334738e-01 1.126156e-01
...
 [9996] 8.175774e-01 5.379471e-01 4.259207e-01 7.629429e-01 9.217997e-01
> s<-y1+y2
> mean(0<s & s <=1/2)
[1] 0.1361
> mean(3/4<s & s<=5/4)
[1] 0.4262
> mean(3/2<s & s<=2)
[1] 0.1244
> mean(s)
[1] 0.9907449
> var(s)
[1] 0.1709682
> sd(s)
[1] 0.413483
> 1/sqrt(6)
[1] 0.4082483
> 7/16
[1] 0.4375
```

3. *Via l'****A.M.P.*** *: Par des développements plutôt avancés, le mathématicien obtient pour tout réel t :*

$$\mathbb{P}(S \leq t) = \begin{cases} 0 & si\ t \leq 0 \\ \frac{t^2}{2} & si\ 0 \leq t \leq 1 \\ 2t - 1 - \frac{t^2}{2} & si\ 1 \leq t \leq 2 \\ 1 & si\ t \geq 2 \end{cases}.$$

Etes-vous en mesure de déterminer les valeurs exactes de la première question ?

4. *L'****A.E.P.*** *confirme-t'elle les résultats théoriques de l'****A.M.P.*** *?*

Exercice 9 (Loi d'une moyenne) *Cet exercice est à lire attentivement à la maison. Il permet d'appréhender via l'approche expérimentale le résultat suivant central en Statistique Inférentielle :*

> *Une moyenne d'un grand nombre de variables aléatoires i.i.d. (indépendantes et identiquement distribuées, i.e. ayant la même loi de probabilité) se comporte approximativement selon la loi Normale (qui tire son nom de ce comportement universel).*

Rappelons que les paramètres d'une loi Normale sont sa moyenne et son écart-type (les matheux préférant sa variance). Notons aussi que ce résultat s'applique dans un cadre assez général excluant tout de même le cas de moyenne de variables aléatoires n'ayant pas de variance finie (et oui, tout arrive !!!).

1. *A partir des exercices 6 et 8, pouvez-vous intuiter les comportements aléatoires des moyennes de 2 faces de dés et de 2 uniformes sur* $[0, 1]$.

   Réponse

   De manière expérimentale, il suffit de diviser par 2 les vecteurs `s` *en* `R` *pour obtenir les quantités d'intérêts désirées. Via l'A.M.P., on obtient très facilement la fonction de répartition de M_2 pour tout réel t :* $\mathbb{P}(M_2 \leq t) = \mathbb{P}(S/2 \leq t) = \mathbb{P}(S \leq 2 \times t)$.
   Les moyenne, variance et écart-type de M_2 se déduisent très facilement de ceux de S : $\mathbb{E}(M_2) = \mathbb{E}(S/2) = \mathbb{E}(S)/2$, $\mathbb{V}ar(M_2) = \mathbb{V}ar(S/2) = \mathbb{V}ar(S)/4$ *et* $\sigma(M_2) = \sigma(S)/2$.

   Fin

2. *On constate sur ces deux exemples que les modalités centrales (autour de la moyenne) sont plus probables pour la moyenne $M_2 := (Y_1 + Y_2)/2$ que sur l'une ou l'autre des variables*

aléatoires Y_1 et Y_2. Pensez-vous que ce phénomène reste vrai pour n'importe quelle paire de variables aléatoires i.i.d. selon Y ? (C'est votre avis qui est demandé !)

3. *Un expérimentateur, convaincu que ce principe est vrai, observe que la moyenne de 4 v.a. i.i.d. se décompose aussi comme une moyenne de 2 v.a. i.i.d. comme le montre la formule suivante :*

$$M_n := \frac{Y_1 + Y_2 + Y_3 + Y_4}{4} = \frac{\frac{Y_1+Y_2}{2} + \frac{Y_3+Y_4}{2}}{2}$$

Il en déduit alors que les valeurs centrales (autour de la moyenne des Y) de la moyenne de 4 v.a. i.i.d. selon Y sont plus probables que celles de la moyenne de 2 v.a. i.i.d. selon Y qui sont elles-mêmes plus probables que celles de Y. Itérant ce processus, il constate que les moyennes M_n de $n = 2^k$ (avec k un entier aussi grand qu'on le veut) v.a. i.i.d. s'écrit aussi comme une moyenne de 2 v.a. i.i.d. étant elles-mêmes des moyennes de 2^{k-1} v.a. i.i.d. elles-mêmes s'écrivant comme des moyennes de 2 v.a. i.i.d. .... En conclusion, il postule que les probabilités d'apparition des modalités centrales de Y augmentent pour la moyenne M_n de n v.a. i.i.d. selon Y lorsque n augmente. Qu'en pensez-vous au vu de son protocole expérimental suivant (les réalisations de M_n sont notées $\mu_{n,[k]}$ et correspondent aux moyennes des lancers de n dés) ?

n	$\overline{\left(\mu_{n,[\cdot]} \in [1,2[\right)}_m$	$\overline{\left(\mu_{n,[\cdot]} \in [2,3[\right)}_m$	$\overline{\left(\mu_{n,[\cdot]} \in [3,4]\right)}_m$	$\overline{\left(\mu_{n,[\cdot]} \in]4,5]\right)}_m$	$\overline{\left(\mu_{n,[\cdot]} \in]5,6]\right)}_m$
1	16.92%	17%	33.38%	17.13%	15.57%
2	8.46%	19.42%	44.63%	19.65%	7.84%
4	2.69%	20.96%	52.59%	21.05%	2.71%
8	0.4%	17.22%	64.83%	17.1%	0.45%
16	0.01%	10.76%	78.32%	10.91%	0%
32	0%	4.27%	91.45%	4.28%	0%
64	0%	0.7%	98.43%	0.87%	0%

4. *L'expérimentateur demande à son ami mathématicien s'il peut justifier sur un plan théorique (via A.M.P.) ces résultats. A sa grande surprise, le mathématicien lui annonce que ce résultat est central en statistique sous le nom de Théorème de la limite centrale (central limit theorem en anglais). Il s'énonce dans le cadre de la moyenne sous la forme suivante : pour toute v.a. Y et lorsque n est suffisamment grand (en général, $n \geq 30$)*

$$M_n := \frac{1}{n}\sum_{i=1}^{n} Y_i \overset{approx.}{\rightsquigarrow} \mathcal{N}\left(\mathbb{E}\left(M_n\right), \sqrt{\frac{\mathbb{V}ar\left(Y_1\right)}{n}}\right)$$

où $Y_1, \cdots, Y_n$ désignent n v.a. i.i.d. selon Y. La loi Normale tire son nom de ce résultat étonnant et combien important dans le sens où beaucoup de phénomènes réels peuvent être vus comme des moyennisations. Le premier paramètre d'une loi Normale correspond à l'espérance $\mathbb{E}\left(M_n\right)$ de M_n et le second à l'écart-type de M_n. Le fait marquant est que ce résultat est vrai indépendemment de la loi de Y. Afin de comparer ces résultats à ceux qu'il a déjà effectué sur la loi uniforme, il transforme toutes les réalisations des lois uniformes sur $[0,1]$ en les multipliant par 5 puis en les additionnant à 1 de sorte que toutes les nouvelles réalisations à moyenner soient celles d'une loi uniforme sur $[1,6]$. L'ensemble des modalités ainsi que celui du dés sont comprises entre 1 et 6. Ainsi, il lui semble possible de comparer les probabilités dans les deux exemples puisque les supports sont les mêmes ainsi que leurs espérances égales à 3.5.

n	$\overline{(\mu_{n,[\cdot]} \in [1,2[)}_m$	$\overline{(\mu_{n,[\cdot]} \in [2,3[)}_m$	$\overline{(\mu_{n,[\cdot]} \in [3,4])}_m$	$\overline{(\mu_{n,[\cdot]} \in]4,5])}_m$	$\overline{(\mu_{n,[\cdot]} \in]5,6])}_m$
1	16.92%	17%	33.38%	17.13%	15.57%
2	8.28%	23.81%	35.54%	23.99%	8.38%
4	1.75%	23.48%	49.27%	23.57%	1.93%
8	0.12%	16.08%	67.25%	16.33%	0.22%
16	0%	8.19%	83.46%	8.35%	0%
32	0%	2.3%	95.08%	2.62%	0%
64	0%	0.24%	99.46%	0.3%	0%

Qu'en pensez-vous ? Observez-vous à nouveau que le procédé de moyennisation concentre les probabilités vers les modalités centrales (en fait autour de l'espérance) ?

5. *Le mathématicien lui fait cependant remarquer qu'a priori les variances ne sont pas rigoureusement les mêmes (certainement assez proches) et qu'il n'est donc pas en mesure de comparer les résultats expérimentaux sur les 2 exemples. Pour comparer les résultats pour différentes v.a. Y, il faut au préalable les uniformiser (les contraindre à avoir les mêmes moyennes et variances). Une solution est de les centrer (soustraire l'espérance $\mathbb{E}(M_n)$) et les réduire (diviser ensuite par $\sqrt{\mathbb{V}ar(M_n)} = \sqrt{\frac{\mathbb{V}ar(Y_1)}{n}}$) de sorte à ce que les v.a. résultantes soient toutes d'espérances 0 et de variances 1 (et ainsi comparables). Cette transformation pourra plus tard (via une représentation graphique) être comparé au travail d'un photographe lors d'une photo de groupe qui demande d'abord à l'ensemble des photographiés de se recentrer (i.e. centrage) puis utilise son zoom (i.e. réduction ou plutôt changement d'échelle dans ce cas précis) pour bien les cadrer. Aidé par le mathémacien, il compare donc ses résultats en effectuant la dite transformation. Le mathématicien l'informe donc du nouveau résultat suivant :*

$$\Delta_n := \frac{M_n - \mathbb{E}(M_n)}{\sqrt{\mathbb{V}ar(M_n)}} = \frac{M_n - \mathbb{E}(M_n)}{\sqrt{\frac{\mathbb{V}ar(Y_1)}{n}}} \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1)$$

N.B. : *Ce résultat n'est valide que lorsque les notions d'espérance et de variance ont un sens ! Il existe en effet des v.a. (suivant une loi de Cauchy, par exemple) n'ayant pas d'espérance et variances finies !*
Voici les résultats expérimentaux pour $n = 64$ (i.e. la valeur de n la plus grande) et $m = 10000$ pour consécutivement les exemples du dé (i.e. $Y \rightsquigarrow \mathcal{U}(\{1, \cdots, 6\})$), de la loi uniforme sur $[0,1]$ (i.e. $Y \rightsquigarrow \mathcal{U}([0,1])$) et sur sa loi transformée uniforme sur $[1,6]$ (i.e. $5Y + 1 \rightsquigarrow \mathcal{U}([1,6])$). Les tableaux ci-dessous sont complétés par les résultats via l'A.M.P. correspondant (théoriquement) à $m = +\infty$.

loi de Y	$\overline{(\delta_{n,[\cdot]} < -3)}_m$	$\overline{(\delta_{n,[\cdot]} \in [-3,-1.5[)}_m$	$\overline{(\delta_{n,[\cdot]} \in [-1.5,-0.5])}_m$	$\overline{(\delta_{n,[\cdot]} \in [-0.5,0.5[)}_m$
$\mathcal{U}(\{1, \cdots, 6\})$	0.11%	6.4%	25.16%	36.76%
$\mathcal{U}([0,1])$	0.11%	6.85%	24.12%	37.63%
$\mathcal{U}([1,6])$	0.11%	6.85%	24.12%	37.63%

loi de Δ_n	$\mathbb{P}(\Delta_n < -3)$	$\mathbb{P}(\Delta_n \in [-3,-1.5[)$	$\mathbb{P}(\Delta_n \in [-1.5,-0.5])$	$\mathbb{P}(\Delta_n \in [-0.5,0.5[)$
$\mathcal{N}(0,1)$	0.13%	6.55%	24.17%	38.29%

loi de Y	$\overline{(\delta_{n,[\cdot]} \in [0.5,1.5[)}_m$	$\overline{(\delta_{n,[\cdot]} \in [1.5,3[)}_m$	$\overline{(\delta_{n,[\cdot]} \geq 3)}_m$	$\overline{(\delta_{n,[\cdot]})}_m$	$\overline{\overline{(\delta_{n,[\cdot]})}_m}$
$\mathcal{U}(\{1, \cdots, 6\})$	24.66%	6.74%	0.17%	$-8e-04$	0.9953
$\mathcal{U}([0,1])$	24.66%	6.53%	0.1%	0.0021	1.0036
$\mathcal{U}([1,6])$	24.66%	6.53%	0.1%	0.0021	1.0036

loi de Δ_n	$\mathbb{P}(\Delta_n \in [0.5,1.5[)$	$\mathbb{P}(\Delta_n \in [1.5,3[)$	$\mathbb{P}(\Delta_n \geq 3)$	$\mathbb{E}(\Delta_n)$	$\sigma(\Delta_n)$
$\mathcal{N}(0,1)$	24.17%	6.55%	0.13%	0	1

Commentez ces résultats et expliquez en particulier pourquoi les 2 lignes correspondant aux 2 exemples des lois uniformes (non transformée et transformée) sont identiques ?

6. *Fournir les instructions R permettant d'obtenir les probabilités des tableaux précédents pour $m = +\infty$.*

Réponse

Pour tout $a < b$,

$$\mathbb{P}\left(\Delta_n \in [a, b[\right) = F_{\mathcal{N}(0,1)}(b) - F_{\mathcal{N}(0,1)}(a) \overset{R}{=} \texttt{pnorm(b)} - \texttt{pnorm(a)}$$

puisque $F_{\mathcal{N}(0,1)}$ est obtenu en `R` en utilisant la fonction `pnorm`.

Fin

Quelques commentaires

- Un étudiant suivant ce cours n'est pas censé comprendre comment les résultats de l'**A.M.P.** ont été mathématiquement obtenus. Ils sont généralement proposés sans démonstration. Sa mission est en revanche de savoir comment les vérifier via l'**A.E.P.** en prenant soin de bien les interpréter. Autrement dit, l'**A.E.P.** permet à un praticien de mieux comprendre les tenants et les aboutissants des outils statistiques (qu'il utilise) développés dans le contexte de l'**A.M.P.**.
- Afin d'éviter de surcharger l'étude de l'**A.E.P.**, il a été décidé dans ce cours d'étaler son introduction en deux étapes. La première qui vous a été présentée dans cette fiche est naturellement complétée par une deuxième étape qui s'appuie sur la représentation graphique des répartitions de $\boldsymbol{y}_{[m]} := \left(y_{[\cdot]}\right)_m$ (avec m généralement très grand). Cette étape est présentée en Annexe. Un étudiant motivé pourra à sa guise choisir de compléter sa connaissance sur l'**A.E.P.** en lisant dès à présent la fiche Annexe A en Annexe consacrée à l'**A.E.P.** dans sa version "graphique". Il est toutefois important de rappeler que les 2 fiches T.D. 3 et 4 suivantes ne s'appuient que sur les outils présentées dans la fiche T.D. présentée ici.
- Dans la suite du cours (nous en avons déjà eu un aperçu dans la fiche introductive précédente), la plupart des variables aléatoires d'intérêt, appelées statistiques, seront de la forme $T := t(\boldsymbol{Y})$ où t est une fonction s'appliquant à $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ qui représente le "futur" échantillon, seule source d'aléatoire dans la variable aléatoire $t(\boldsymbol{Y})$. C'est en effet le cas pour l'estimation d'un paramètre inconnu θ qui s'écrit $\widehat{\theta}(\boldsymbol{y})$ lorsqu'il est évalué à partir de l'échantillon que l'on obtient le **Jour J** (i.e. jour d'obtention des données) et qui est la réalisation de $\widehat{\theta}(\boldsymbol{Y})$ représentant le procédé d'obtention de l'estimation à partir du "futur" échantillon $\boldsymbol{Y}$. L'étude **A.E.P.** consistera alors à construire m échantillons $\left(\boldsymbol{y}_{[\cdot]}\right)_m$ où $\boldsymbol{y}_{[k]} := (y_{1,[k]}, \cdots, y_{n,[k]})$ représente le $k^{ème}$ échantillon de taille n construit parmi les m. Le comportement aléatoire d'une statistique $T := t(\boldsymbol{Y})$ sera donc appréhendé via l'**A.E.P.** en proposant m réalisations indépendantes $\left(t_{[\cdot]}\right)_m := \left(t(\boldsymbol{y}_{[\cdot]})\right)_m$ avec $t_{[k]} := t(\boldsymbol{y}_{[k]})$ la $k^{ème}$ réalisation de T parmi les m.

Fin

FICHE T.D. 3 Estimation ponctuelle et par intervalle de confiance

Indications préliminaires

- *Objectif* : Dans la fiche d'introduction, le cadre de ce cours de Statistique Inférentielle a été posé. En question préliminaire, nous aurons, pour chaque problématique considérée, à identifier le paramètre d'intérêt (noté θ en général lorsque la problématique n'est pas encore précisée) et à bien prendre conscience que ce dernier est **inconnu**. A partir d'un échantillon $\boldsymbol{y}$ récolté le **jour J** (cette appellation sera utilisée tout au long de ce cours), nous aurons alors comme objectif de proposer une estimation, notée $\widehat{\theta}(\boldsymbol{y})$ (pour bien exprimer la dépendance en l'échantillon $\boldsymbol{y}$), afin d'avoir une idée sur l'ordre de grandeur de θ (inconnu). Dans un deuxième temps, nous réaliserons que ce type d'estimation ponctuelle (i.e. un paramètre inconnu estimé par une unique valeur estimée) n'est pas satisfaisant en termes de confiance que l'on peut apporter à l'estimation. Le statisticien se doit alors de proposer à partir du même échantillon $\boldsymbol{y}$, un niveau de qualité de l'estimation $\widehat{\theta}(\boldsymbol{y})$. L'**erreur standard** ("standard error" en anglais) est alors introduite s'exprimant comme une estimation de l'écart-type (i.e. indicateur de variabilité) de la "future" estimation $\widehat{\theta}(\boldsymbol{Y})$ (à partir du futur échantillon $\boldsymbol{Y}$) ayant pour réalisation $\widehat{\theta}(\boldsymbol{y})$ le **jour J**. En appliquant le système notation Norme CQLS (voir polycopié de cours), cette erreur standard se note $\widehat{\sigma_{\widehat{\theta}}}(\boldsymbol{y})$. La meilleure façon de proposer une estimation tenant compte du couple d'informations $(\widehat{\theta}(\boldsymbol{y}), \widehat{\sigma_{\widehat{\theta}}}(\boldsymbol{y}))$ disponible le **jour J** est de construire un intervalle de confiance $IC_{\theta,1-\alpha}(\boldsymbol{y}) := [\widetilde{\theta}_{\inf}(\boldsymbol{y}), \widetilde{\theta}_{\sup}(\boldsymbol{y})]$ à $1-\alpha$ de niveau de confiance. Grâce à la l'**A.E.P.**, nous aurons comme mission prioritaire de bien interpréter la notion de niveau de confiance.
- *Loi de probabilité de l'écart standardisé* : Les paramètres d'intérêt considérés dans ce cours sont de manière plus ou moins directe tous reliés à la moyenne. Ainsi, dans un cadre asymptotique où nous supposerons disposer d'un nombre suffisant de données, nous pourrons hériter pleinement de la puissance du Théorème de la limite centrale que nous avons étudié précédemment (notamment dans la fiche T.D. 2 mais aussi dans la fiche Annexe A consacrée aux représentations graphiques des lois de probabilité). Dans le contexte de l'estimation d'un paramètre θ traité dans ce cours, il s'exprime par (n supposé suffisamment grand) :

$$\widehat{\Theta}_n := \widehat{\theta}(\boldsymbol{Y}) \overset{approx.}{\rightsquigarrow} \mathcal{N}\left(\theta, \sigma_{\widehat{\theta}}\right) \Leftrightarrow \Delta_n := \frac{\widehat{\theta}(\boldsymbol{Y}) - \theta}{\sigma_{\widehat{\theta}}} \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1).$$

où $\sigma_{\widehat{\theta}} := \sigma\left(\widehat{\theta}(\boldsymbol{Y})\right) = \sqrt{\mathbb{V}ar\left(\widehat{\theta}(\boldsymbol{Y})\right)}$ est l'écart-type de la "future" estimation $\widehat{\theta}(\boldsymbol{Y})$. Cependant, en général, le paramètre $\sigma_{\widehat{\theta}}$ est lui-même inconnu et doit être estimé par $\widehat{\sigma_{\widehat{\theta}}}(\boldsymbol{y})$) correspondant à l'erreur standard. Un résultat applicable dans le cas où $\sigma_{\widehat{\theta}}$ est inconnu, est le suivant :

$$\Delta_{\widehat{\theta},\theta} := \delta_{\widehat{\theta},\theta}(\boldsymbol{Y}) := \frac{\widehat{\theta}(\boldsymbol{Y}) - \theta}{\widehat{\sigma_{\widehat{\theta}}}(\boldsymbol{Y})} \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1).$$

- *La probabilité comme une extension de la logique* : Nous insistons sur le fait qu'une probabilité d'un <u>événement</u> égale à **0** ou **1** signifie respectivement de manière équivalente que l'<u>événement</u> (dit **certain**) est **Faux** ou **Vrai**. C'est en ce sens que la "probabilité" est une extension de la "logique" (en tant que théorie mathématique). Un événement **incertain** a donc une probabilité strictement comprise entre 0 et 1 et exprime donc qu'il est peut-être Vrai ou peut-être Faux, la probabilité de l'événement étant d'autant plus grande (resp. petite) que l'événement a de plus en plus de chance d'être Vrai (resp. Faux). Dans le contexte statistique, un événement s'exprime à partir d'une statistique $T := t(\boldsymbol{Y})$ sous la forme $(T \in E) \Leftrightarrow (t(\boldsymbol{Y}) \in E)$ où E est un sous-ensemble de modalités de $T := t(\boldsymbol{Y})$. Ainsi, connaissant la loi de probabilité de $T := t(\boldsymbol{Y})$, nous serons en mesure d'évaluer $\mathbb{P}(T \in E) = \mathbb{P}(t(\boldsymbol{Y}) \in E)$ comprise strictement entre 0 et 1 puisque $\boldsymbol{Y}$ est intrinsèquement aléatoire. Une erreur très courante est de confondre, le **Jour J**, $\mathbb{P}(t(\boldsymbol{y}) \in E)$ avec $\mathbb{P}(t(\boldsymbol{Y}) \in E)$ alors que $\boxed{\mathbb{P}(t(\boldsymbol{y}) \in E) \in \{0,1\}} \neq \boxed{\mathbb{P}(t(\boldsymbol{Y}) \in E) \in]0,1[}$ puisque $\boldsymbol{y}$ est déterministe (i.e. strictement non aléatoire) en tant que réalisation de $\boldsymbol{Y}$.

Fin

Exercice 10 (Salaire Juste - Estimation (ponctuelle))

Une équipe de sociologues propose de réunir un comité d'experts pour la création d'un indicateur, appelé ***Salaire Juste****, mesuré pour toute personne active et qui permettra de transformer les ressources individuelles réelles (généralement mesurées par un salaire) en tenant compte de critères aussi importants que les ressources locales, le partage de ces ressources, la pénibilité du travail, le niveau d'expérience, d'expertise et bien d'autres encore... Cet indicateur est conçu de sorte qu'en théorie il devrait être équivalent (en fait égal à une valeur étalon 100) pour toute personne active dans le monde. En conséquence directe, le Salaire Juste moyen dans le monde devrait être égal à 100. Après quelques mois de travail, un premier prototype (très perfectible) du* ***Salaire Juste*** *est élaboré par la fine équipe d'experts. Les sociologues s'accordent à dire qu'un pays peut se dire non civilisé s'il vérifie aux 2 critères de discriminations suivants :*

> ***Discrimination Mondiale*** *: le Salaire Juste moyen dans le pays est très supérieur à la valeur 100 de base. Un Salaire Juste moyen excédant un seuil de 150 est considéré comme intolérable.*
>
> ***Discrimination Intérieure*** *: les Salaires Justes dans le pays sont très dispersés. La variance des Salaires Justes dans le pays supérieur à 30 est considérée comme excessive et donc anormale.*

Par la suite, $\mathcal{Y}_i$ $(i = 1, \cdots, N)$ désigne le Salaire Juste de la $i^{\text{ème}}$ personne actives du pays.

1. *Définir mathématiquement les paramètres (d'intérêt), notés μ^J et σ_J^2, permettant éventuellement d'établir des discriminations mondiale et intérieure. Quelle est la nature de ces paramètres ?*
2. *Soit Y^J la variable aléatoire (v.a.) correspondant au Salaire Juste d'un individu choisi au hasard dans la population des N personnes actives du pays. Etablir la relation entre les paramètres μ^J et σ_J^2 et la v.a. Y^J*
3. *Rappeler alors les estimateurs proposés par les mathématiciens obtenus à partir d'un "futur" échantillon $\boldsymbol{Y^J}$ (en utilisant la Norme CQLS).*
4. *Quelles sont les "bonnes" propriétés de ces estimateurs désirées par les mathématiciens ? Interrogez-vous sur comment les interpréter via l'A.M.P. ?*
5. *Proposer à présent leur interprétation via l'A.E.P. en prenant soin au préalable d'introduire les notations nécessaires (Norme CQLS). Proposer alors une description littérale pour chacune de ces "bonnes" propriétés.*
6. *Une étude est menée par un expérimentateur. Il se fixe l'ensemble des Salaires Justes sur un pays fictif de $N = 1000000$ personnes actives dont il est le seul à en connaître les valeurs. Voici les résultats présentés dans les tableaux ci-dessous :*

$\boldsymbol{Y}$	$\widehat{\mu^J}(\boldsymbol{Y})$	$\widehat{\sigma_J}(\boldsymbol{Y})$	$\widehat{\sigma_{\widehat{\mu^J}}}(\boldsymbol{Y})$	$\widehat{\mu^J}(\boldsymbol{Y}) - \mu^J$	$\delta_{\widehat{\mu}^J,\mu^J}(\boldsymbol{Y})$
$\boldsymbol{y}_{[1]}$	99.91	10.0231	0.317	-0.09	-0.29
$\boldsymbol{y}_{[2]}$	99.65	9.2615	0.2929	-0.35	-1.19
$\boldsymbol{y}_{[3]}$	100.84	10.448	0.3304	0.84	2.54
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{y}_{[9998]}$	99.58	9.3785	0.2966	-0.42	-1.43
$\boldsymbol{y}_{[9999]}$	100.2	9.9372	0.3142	0.2	0.63
$\boldsymbol{y}_{[10000]}$	99.94	9.7991	0.3099	-0.06	-0.21
Moyenne	100.0043	10.034	0.3173	0.0043	-0.0294
Ecart-type	0.3178	0.63	0.0199	0.3178	1.0058

7. *En notant $\boldsymbol{yy}$ un échantillon correspondant à une ligne du tableau ci-dessous (par exemple, la $3^{\text{ème}}$), fournir les instructions `R` qui a permis à l'expérimentateur d'obtenir les valeurs du tableau précédent (Indication : étant ici à la place de l'expérimentateur, n'oubliez pas que vous disposez exceptionnellement les valeurs de μ^J et σ_J^2).*
8. *Proposer les notations mathématiques correspondant aux 2 dernières lignes du tableau qui, nous l'espérons, permet de comprendre à quoi elles correspondent et comment elles ont été obtenues.*

9. *Quelles valeurs du tableau sont sensées mesurer (approximativement) les qualités de l'estimateur* $\widehat{\mu^J}\left(\boldsymbol{Y^J}\right)$ *? Comment les noter dans l'A.M.P. ? Sont-elles accessibles le jour J ? Mêmes questions pour l'estimateur* $\widehat{\sigma_J^2}\left(\boldsymbol{Y^J}\right)$.
10. *Comment obtient-on les estimations des qualités mesurées par les écarts-type des estimateurs* $\widehat{\mu^J}\left(\boldsymbol{Y^J}\right)$ *et* $\widehat{\sigma_J^2}\left(\boldsymbol{Y^J}\right)$. *Comment sont-elles appelées ?*
11. *A partir de maintenant, on s'imagine être le jour J. Pour cela, on suppose ne disposer que du* $3^{\text{ème}}$ *échantillon dans le tableau ci-dessus. Comment doit-on noter ce jeu de données. Proposer à partir du tableau toutes les estimations intéressantes relativement aux problèmes de discriminations mondiale et intérieure. N'en manque-t-il pas une ou plusieurs ? Retrouvez-les ou complétez-les à partir de la sortie* `R` *suivante :*

```
> length(yy)
[1] 1000
> mean(yy)
[1] 100.8388
> sd(yy)
[1] 10.44798
> var(yy)
[1] 109.1603
> seMean(yy)
[1] 0.3303941
> sd(yy)/sqrt(length(yy))
[1] 0.3303941
> seVar(yy)
[1] 9.475496
```

12. *Voici les sorties* `R`*, correspondant aux mêmes informations mais sur l'échantillon des n=100 premiers individus :*

```
> mean(yy[1:100])
[1] 101.7301
> sd(yy[1:100])
[1] 12.13053
> var(yy[1:100])
[1] 147.1498
> seMean(yy[1:100])
[1] 1.213053
> sd(yy)/sqrt(100)
[1] 1.044798
> seVar(yy[1:100])
[1] 32.54073
```

Comparer ces résultats à ceux obtenus à partir de l'échantillon initial de taille n=1000. Quelle type d'estimation vaut-il mieux préconiser lorsqu'on désire intégrer l'erreur standard ?

Exercice 11 (Salaire Juste - Estimation par intervalle de confiance)

1. *A partir de votre formulaire, rappeler les expressions des "futurs" intervalles de confiance à* 95% *(généralement noté* $1-\alpha$*) de niveau de confiance pour les paramètres* μ^J *et* σ_J^2. *Rappeler à partir de quel résultat mathématique (probabiliste) ont-ils été construits ? Evaluer la probabilité* $\mathbb{P}\left(|\delta_{\widehat{\theta},\theta}(\boldsymbol{Y^J})| \leq 1.96\right) = \mathbb{P}\left(-1.96 \leq \delta_{\widehat{\theta},\theta}(\boldsymbol{Y^J}) \leq 1.96\right)$ *où* θ *désigne indifféremment* μ^J *et* σ_J^2. *L'interpréter via l'A.E.P. notamment avec le tableau précédent.*
2. *Question optionnelle (pour ceux qui ne sont pas rebutés par de simples calculs mathématiques) : Construire mathématiquement les futurs intervalles de confiance ci-dessus.*
3. *Fournir l'instruction* `R` *permettant de les obtenir le jour J (Indication : en* `R`*,* `qnorm(.975)` $\simeq 1.96$*) et le calculer éventuellement en utilisant votre machine à calculer. Déduire un intervalle de confiance à* 95% *pour* σ_J.
4. *Voici sur les résultats expérimentaux pour les intervalles de confiance* $IC_{\mu^J}(\boldsymbol{Y^J})$ *et* $IC_{\sigma_J^2}(\boldsymbol{Y^J})$ *de* μ^J *et* σ_J^2. *Interpréter via l'approche expérimentale*

$\boldsymbol{Y}$	$IC_{\mu^J}(\boldsymbol{Y^J})$	$\mu^J \in IC_{\mu^J}(\boldsymbol{Y^J})$	$IC_{\sigma_J^2}(\boldsymbol{Y^J})$	$\sigma_J^2 \in IC_{\sigma_J^2}(\boldsymbol{Y^J})$
$\boldsymbol{y_{[1]}}$	$[99.29, 100.53]$	1	$[61.61, 139.31]$	1
$\boldsymbol{y_{[2]}}$	$[99.08, 100.23]$	1	$[71.6, 99.95]$	0
$\boldsymbol{y_{[3]}}$	$[100.19, 101.49]$	0	$[90.59, 127.73]$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{y_{[9998]}}$	$[99, 100.16]$	1	$[72.18, 103.74]$	1
$\boldsymbol{y_{[9999]}}$	$[99.58, 100.81]$	1	$[82.57, 114.93]$	1
$\boldsymbol{y_{[10000]}}$	$[99.33, 100.54]$	1	$[77.91, 114.13]$	1
Moyenne		94.86%		92.02%

5. *Evaluer les probabilités suivantes :*
$\mathbb{P}\left(\mu^J \in IC_{\mu^J}(\boldsymbol{y^J})\right) = \mathbb{P}\left(\mu^J \in [100.19, 101.49]\right)$ *et* $\mathbb{P}\left(\sigma_J^2 \in IC_{\sigma_J^2}(\boldsymbol{y^J})\right) = \mathbb{P}\left(\sigma_J^2 \in [90.59, 127.73]\right)$

Exercice 12 *(taille étudiants)*

*Pour mettre en pratique ce qu'il a appris dans son cours de Statistique Inférentielle, un étudiant souhaite utiliser l'**Approche Expérimentale** pour comprendre la notion d'intervalle de confiance. Son but est d'estimer par intervalle de confiance **la taille moyenne**, notée μ, des $N = 300$ étudiants de sa promotion.*

1) Il construit un premier échantillon (avec remise) de taille $n = 30$ (i.e. pour se placer dans le cadre asymptotique), qu'il note $\boldsymbol{y_{[1]}}$, dans la population des $N = 300$ étudiants de sa promotion :

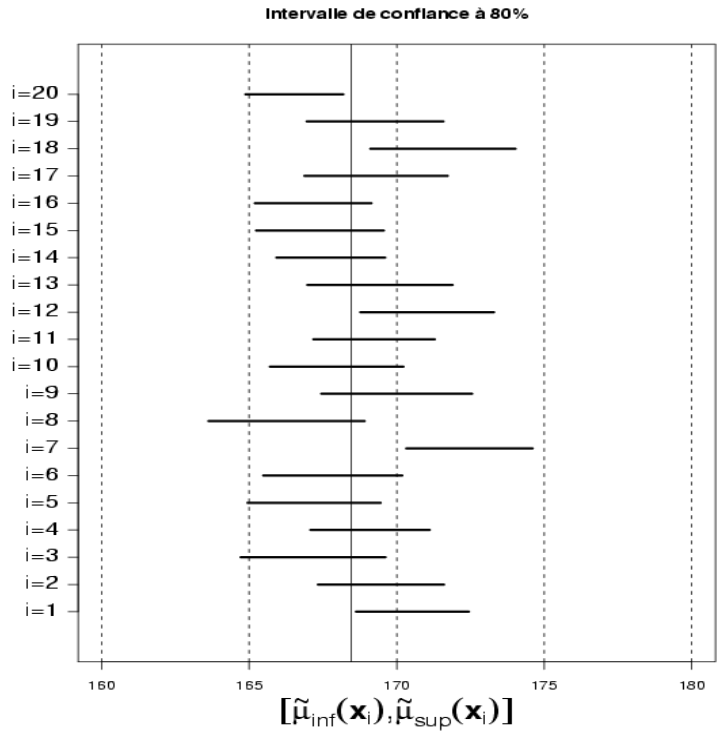
```
> y1
 [1] 165 179 171 178 171 168 166 171 182 178 177 165 174 164 175 178 167 168 185
[20] 166 162 180 167 174 159 159 184 154 172 157
```

Proposez l'instruction `R` ayant permis d'obtenir le résultat ci-dessous correspondant à un intervalle de confiance au niveau de confiance de 80% de μ :

Indication(s) R :

```
> # IC <- (instruction R à fournir dans la rédaction)
> IC
[1] 168.6308 172.4359
```

2) Ne sachant pas comment interpréter ce résultat, il construit 19 autres échantillons de taille $n = 30$ dans la population des étudiants de sa promotion que l'on notera respectivement $\boldsymbol{y_{[2]}}, \ldots, \boldsymbol{y_{[20]}}$. Il représente alors sur un même graphique ces 20 différents intervalles de confiance de μ à 80% de niveau de confiance :



Afin de confronter ses résultats expérimentaux avec la réalité, l'étudiant décide d'interroger tous les étudiants de sa promotion (notez que ceci est possible car $N = 300$*). Il peut alors calculer la valeur de* μ*, à savoir* 168.45*. Elle est représentée par le trait vertical (en trait plein) sur le graphique précédent. Sur les 20 intervalles de confiance calculés, combien contiennent* μ *? Est-ce surprenant ?*

3) Que se passerait-il si l'étudiant construisait une infinité d'intervalles de confiance de μ *à* 80% *de niveau de confiance sur des échantillons de taille* $n = 30$ *?*

Exercice 13 *(élection présidentielle) Entre les deux tours d'une élection présidentielle, on souhaite estimer par intervalle de confiance à 95% les proportions d'intentions de vote des deux candidats finalistes. Avant même d'effectuer un sondage sur une sous-population de taille* $n = 1000$*, quelle serait la plus grande longueur des deux intervalles de confiance (en utilisant la formule approchée) ?*

Indication(s) R :

```
> 2*qnorm(0.975)*sqrt(0.5*0.5/1000)
[1] 0.0619795
```

Quelle doit être la taille de l'échantillon n *pour être certain que la longueur de l'intervalle de confiance au niveau* 95% *n'excède pas* 0.04% *?*

Indication(s) R :

```
> (2*qnorm(.975)*sqrt(.5^2)/.0004)^2
[1] 24009118
> (qnorm(.975)/.0004)^2
[1] 24009118
```

Exercice 14

Un industriel s'interroge sur la proportion d'acheteurs parmi sa clientèle qui ont acheté ou ont l'intention d'acheter le produit A, proportion notée p^A*. En particulier, il souhaiterait construire un intervalle de confiance de cette proportion d'acheteurs* p^A *obtenu à partir d'un échantillon de taille* $n = 500$ *individus issus de la population de taille* $N = 2000000$*.*

1. *Proposez l'instruction* `R` *ayant permis d'obtenir le résultat ci-dessous correspondant à un intervalle de confiance au niveau de confiance de 90% de* p^A *calculé à partir du jeu de données* $\boldsymbol{y}$ *que l'on note* $\boldsymbol{y}$ *en* `R` *(cet intervalle est noté* $[\widetilde{p^A}_{inf}(\boldsymbol{y}), \widetilde{p^A}_{sup}(\boldsymbol{y})]$ *:*
Indication(s) R :

```
> # IC <- (instruction R à fournir dans la rédaction)
> IC
[1] 0.1630267 0.2209733
```

2. *Le produit A est maintenant lancé sur le marché, et il a été alors possible d'évaluer le vrai paramètre* p^A *à* 18.9%*. Pour essayer de faire comprendre à l'un de ses collègues comment il faut interpréter les intervalles de confiance (en particulier le précédent), le concurrent propose l'exercice pédagogique suivant. On construit une urne de taille* $N = 2000000$ *boules dont une proportion* $p^A = 18.9\%$ *sont numérotées 1 (les autres étant numérotées 0). On fait alors* 199 *tirages de 500 boules au hasard au sein de cette urne. Les jeux de données créés sont donc de la même nature que* $\boldsymbol{y}$*. Les* $m = 200$ *jeux de données sont notés* $\boldsymbol{y}_{[1]}$, $\boldsymbol{y}_{[2]}$, $\ldots$, $\boldsymbol{y}_{[200]}$ *(le premier* $\boldsymbol{y}_{[1]}$ *correspondant à* $\boldsymbol{y}$*). Pour chacun de ces jeux de données, on construit un intervalle de confiance au niveau de 90% du paramètre* p^A*. Voici dans l'ordre des tirages quelques uns de ces intervalles :*

```
pInf      pSup
  [1,] 0.1630267 0.2209733
  [2,] 0.1971384 0.2588616
  [3,] 0.2210000 0.2210000
...
[198,] 0.1649122 0.2230878
[199,] 0.1724662 0.2315338
[200,] 0.1573773 0.2146227
```

Parmi les $m = 200$ *intervalles de confiance, 179 contiennent le vrai paramètre* p^A*, qu'en pensez-vous ? Si l'on construisait une infinité d'intervalles de confiance, combien contiendraient le vrai paramètre* p^A *?*

3. *Complétez sans justification les encadrés ci-dessous :*

$$\mathbb{P}\left(\widetilde{p^A}_{inf}(\boldsymbol{y}_{[1]}) < p^A < \widetilde{p^A}_{sup}(\boldsymbol{y}_{[1]})\right) = \boxed{}$$

$$\mathbb{P}\left(\widetilde{p^A}_{inf}(\boldsymbol{y}_{[2]}) < p^A < \widetilde{p^A}_{sup}(\boldsymbol{y}_{[2]})\right) = \boxed{}$$

$$\mathbb{P}\left(\widetilde{p^A}_{inf}(\boldsymbol{Y}) < p^A < \widetilde{p^A}_{sup}(\boldsymbol{Y})\right) \simeq \boxed{}$$

4. *Complétez sans justification les encadrés ci-dessous :*
   - *si le niveau de confiance avait été de 95% alors*

$$\mathbb{P}\left(\widetilde{p^A}_{inf}(\boldsymbol{y}_{[1]}) < p^A < \widetilde{p^A}_{sup}(\boldsymbol{y}_{[1]})\right) = \boxed{}$$

   - *si le niveau de confiance avait été de 80% alors*

$$\mathbb{P}\left(\widetilde{p^A}_{inf}(\boldsymbol{y}_{[2]}) < p^A < \widetilde{p^A}_{sup}(\boldsymbol{y}_{[2]})\right) = \boxed{}$$

5. *Complétez sans justification les encadrés ci-dessous :*
   - *si le niveau de confiance avait été de 95% alors*

$$\mathbb{P}\left(\widetilde{p^A}_{inf}(\boldsymbol{y}_{[2]}) < p^A < \widetilde{p^A}_{sup}(\boldsymbol{y}_{[2]})\right) = \boxed{}$$

   - *si le niveau de confiance avait été de 80% alors*

$$\mathbb{P}\left(\widetilde{p^A}_{inf}(\boldsymbol{y}_{[1]}) < p^A < \widetilde{p^A}_{sup}(\boldsymbol{y}_{[1]})\right) = \boxed{}$$

Exercice 15 *Avant le premier tour des élections, nous sommes souvent assaillis par de nombreux sondages. Le 13 mars 2012, deux instituts de sondages (IFOP et SOFRES) publient leurs estimations sur les intentions de votes pour deux candidats C1 et C2 :*

- *Sondage IFOP* ($n = 1638$) *:* $\widehat{p^{C1}}\left(\boldsymbol{y^I}\right) = 27\%$ *et* $\widehat{p^{C2}}\left(\boldsymbol{y^I}\right) = 28.5\%$
- *Sondage SOFRES* ($n = 1000$) *:* $\widehat{p^{C1}}\left(\boldsymbol{y^S}\right) = 30\%$ *et* $\widehat{p^{C2}}\left(\boldsymbol{y^S}\right) = 28\%$

1. *A la lumière de ce cours, nous proposons les mêmes résultats présentés à partir des intervalles de confiance à 95% de niveau de confiance :*
   - *Sondage IFOP :* $IC^{C1}(\boldsymbol{y^I}) = [24.85\%, 29.15\%]$ *et* $IC^{C2}(\boldsymbol{y^I}) = [26.31\%, 30.69\%]$
   - *Sondage SOFRES :* $IC^{C1}(\boldsymbol{y^S}) = [27.16\%, 32.84\%]$ *et* $IC^{C2}(\boldsymbol{y^S}) = [25.22\%, 30.78\%]$

   Fournir au choix :
   - *la formule mathématique (générale) permettant d'obtenir l'intervalle de confiance d'une proportion* p *s'exprimant en fonction de l'estimation* $\widehat{p}(\boldsymbol{y})$ *et de la taille d'échantillon* n*.*
   - *la vérification à la calculatrice de l'obtention de l'un des intervalles de confiance ci-dessus (détails des calculs à fournir).*
   - *la formule* `R` *d'obtention d'un intervalle de confiance en fonction de* `pEst` *et* `n` *désignant respectivement l'intention de vote pour un candidat et la taille d'échantillon.*
2. *Interpréter via l'approche expérimentale des probabilités les intervalles de confiance obtenus à la question précédente.*
3. *La plupart des commentateurs politiques ont semblé troublés par de tels résultats apparemment contradictoires. A partir de la connaissance acquise dans ce cours et en supposant (de manière un peu abusive) que tous les intervalles de confiances précédents contiennent le vrai paramètre inconnu, pensez-vous qu'on puisse savoir lequel des candidats est en tête au premier tour ? Justifiez très simplement votre réponse en envisageant deux cas de figures bien choisis.*

FICHE T.D. 4

Traitement des problématiques des produits A et B

Indications préliminaires

- *Objectif* : En pratique, on peut être spécialement intéressé par une prise de décision qui dépend de la comparaison du **paramètre d'intérêt** θ inconnu par rapport à une **valeur de référence** θ_0 (fixée selon la problématique). Cette comparaison sera par la suite appelée **assertion d'intérêt**. Ne disposant que d'une estimation $\widehat{\theta}(\boldsymbol{y})$ la décision conduisant à conclure que l'assertion d'intérêt est vraie à partir de l'échantillon $\boldsymbol{y}$ du **jour J** ne peut pas être complètement fiable. L'objectif est de construire un outil d'aide à la décision nous garantissant un risque d'erreur de se tromper dans notre décision de valider l'assertion d'intérêt n'excédant pas une valeur que nous nous sommes fixée (généralement autour des 5%).
- *Paramètre d'écart standardisé* : Comparer la paramètre d'intérêt θ à une valeur de référence θ_0 est strictement équivalent à comparer leur différence ou leur rapport à 0 ou 1. Dans le cadre asymptotique, l'assertion d'intérêt pourra toujours se réécrire en fonction d'un paramètre d'écart standardisé $\delta_{\theta,\theta_0} := \dfrac{\boldsymbol{\theta - \theta_0}}{\sigma_{\widehat{\theta}}}$. Il est important d'apprendre à lire cette expression où le numérateur $\theta - \theta_0$ a été mis en **gras** pour souligner son rôle plus important (en termes d'information pour l'utilisateur) par rapport au dénominateur $\sigma_{\widehat{\theta}}$ ayant été introduit principalement pour des raisons techniques (mais toutefois indispensables dans la construction de l'outil d'aide à la décision). Il est alors direct de voir que :

$$\textbf{Assertion d'intérêt} \iff \left\{ \begin{array}{lllll} \theta < \theta_0 & \iff & \theta - \theta_0 < 0 & \iff & \delta_{\theta,\theta_0} < 0 \\ \theta > \theta_0 & \iff & \theta - \theta_0 > 0 & \iff & \delta_{\theta,\theta_0} > 0 \\ \theta \neq \theta_0 & \iff & \theta - \theta_0 \neq 0 & \iff & \delta_{\theta,\theta_0} \neq 0 \end{array} \right\}$$

  En commentaire non prioritaire, on peut tout de même remarquer que l'interprétation du $\sigma_{\widehat{\theta}}$ dans l'expression de δ_{θ,θ_0} est assez naturelle : plus l'estimation de θ est fiable, se traduisant par un $\sigma_{\widehat{\theta}}$ d'autant plus faible, plus le paramètre d'écart standardisé δ_{θ,θ_0} est grand et ainsi plus facile à comparer à 0.
- *Estimation du paramètre d'écart standardisé* : Dépendant du paramètre d'intérêt θ inconnu, le paramètre d'écart standardisé δ_{θ,θ_0} est lui-même inconnu (en fait doublement inconnu puisque dépendant aussi de $\sigma_{\widehat{\theta}}$ inconnu). Il est facilement estimable à partir de l'échantillon $\boldsymbol{y}$ du **jour J**. Nous l'exprimons ci-dessous à partir du "futur" échantillon $\boldsymbol{Y}$:

$$\widehat{\delta_{\theta,\theta_0}}(\boldsymbol{Y}) := \frac{\widehat{\theta}(\boldsymbol{Y}) - \theta_0}{\widehat{\sigma_{\widehat{\theta}}}(\boldsymbol{Y})}$$

  Pour mesurer les risques d'erreur de décision, nous serons tout particulièrement intéressés par la loi de probabilité de $\widehat{\delta_{\theta,\theta_0}}(\boldsymbol{Y})$ lorsque $\theta = \theta_0$. Dans ce cas très particulier, nous remarquons que lorsque n est suffisamment grand :

$$\widehat{\delta_{\theta,\theta_0}}(\boldsymbol{Y}) := \frac{\widehat{\theta}(\boldsymbol{Y}) - \boxed{\theta_0}}{\widehat{\sigma_{\widehat{\theta}}}(\boldsymbol{Y})} = \frac{\widehat{\theta}(\boldsymbol{Y}) - \boxed{\theta}}{\widehat{\sigma_{\widehat{\theta}}}(\boldsymbol{Y})} =: \delta_{\widehat{\theta},\theta}(\boldsymbol{Y}) \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1)$$

  où $\delta_{\widehat{\theta},\theta}(\boldsymbol{Y})$ a été introduit au début de la fiche T.D. 3.

Fin

Exercice 16 (Forme des Règles de Décision pour produits A et B)

1. *Exprimer les assertions d'intérêt pour les produits A et B correspondant aux lancements des produits sur le marché en fonction des paramètres d'intérêts.*
2. *Reécrire ces assertions d'intérêt à partir des paramètres d'écart standardisé fournis dans le formulaire de cours.*

3. *Proposer les formes des Règles de décision associées aux expressions précédentes des assertions d'intérêts (via paramètres d'intérêt et d'écart standardisé).*
4. *Selon ces Règles de Décision, est-il possible pour l'industriel de ne pas se tromper dans sa décision quant au lancement de chaque produit ?*
5. *Exprimez les erreurs de décision éventuelles en les illustrant par des exemples de situations réelles envisageables.*

   Il en ressort qu'il y a 2 risques d'erreurs de décision :
   — *Risque d'erreur I (ou première espèce) : risque de décider à tort l'assertion d'intérêt.*
   — *Risque d'erreur II (ou deuxième expèce) : risque de ne pas décider à tort l'assertion d'intérêt.*

   Comment les reformuleriez-vous dans les problématiques de l'industriel ? Lequel parmi ces 2 types d'erreurs est-il plus important de controler ?
6. *Un statisticien informe l'industriel que la règle de décision (connue de tous les statisticiens) est de lancer le produit A si* $\widehat{p^A}\left(\boldsymbol{y^A}\right) > 16.8573\%$. *L'industriel pensant que son produit est tel que* $p^A \geq 20\%$ *se dit que cette règle de décision sera donc toujours acceptée. Qu'en pensez-vous ?*

Exercice 17 (Etudes expérimentales pour produits A et B)

L'expérimentateur désire mener une étude sur les outils d'aide à la décision pour les problématiques des produits A et B. Il se propose alors de construire 6 urnes U_p^A *(avec* $p = 10\%$, 14%, 15%, 15.1%, 16% *et* 20%*) et 6 urnes* U_μ^B *(avec* $\mu = 0.1$, 0.14, 0.15, 0.151, 0.16 *et* 0.2*). Pour chacune des 12 urnes il construit* $m = 10000$ *échantillons (notés* $\boldsymbol{y_{[k]}}$, $k = 1, \cdots, m$*) de taille* $n = 1000$. *Voici les caractéristiques de ces urnes : une urne* $U_\mu^\bullet$ *(*$\mu \geq 0$ *et* $\bullet$*=A ou B) contient N=2000000 boules numérotées de 0 à 3.* $N_i^\bullet$ *désignant le nombre de boules dont le numéro est i (*$i = 0, \cdots, 3$*), les répartitions de ces urnes sont fixées de la manière suivante :*
— U_p^A *(*$\mu = p \in [0,1]$*) :* $N_1^A = N \times p$ *et* $N_2^A = N_3^A = 0$ *de sorte qu'il y a une proportion p de boules numérotées 1. La moyenne et la variance des numéros sont respectivement égaux à* $\mu = p$ *et* $\sigma^2 = p(1-p)$
— U_μ^B *(*$\mu \geq 0$*) :* $N_1^B = N \times \mu - 100000$, $N_2^B = N_3^B = 20000$ *de sorte que la moyenne des numéros est égale à* μ. *La variance est égale à* $\sigma^2 = \mu(1-\mu) + \frac{2}{25}$.

Notons que $N_0^\bullet = N - (N_1^\bullet + N_2^\bullet + N_3^\bullet)$.

Résultats expérimentaux pour le produit A : *L'expérimentateur décide d'éprouver les Règles de Décision suivantes sur tous les* $m = 10000$ *échantillons des 6 urnes* U_p^A *:*

$$\textit{Décider de lancer le produit A si } \boxed{\widehat{p^A}\left(\boldsymbol{y}\right) > p_{lim}^+}$$

avec p_{lim}^+ *pouvant prendre les 4 valeurs du tableau ci-dessous choisies sous le conseil du mathématicien. Voici les résultats fournis via la quantité* $\gamma_m(p) = \overline{\left(\widehat{p^A}\left(\boldsymbol{y}_{[.]}\right) > p_{lim}^+\right)}_m$ *correspondant à la proportion d'échantillons parmi les* $m = 10000$ *conduisant au lancement du produit. Les valeurs entre parenthèses dans le tableau, fournies par le mathématicien, correspondent aux différentes valeurs de* $\gamma_{+\infty}(p)$ *(i.e.* $m = +\infty$*) qui, via la relation entre l'A.E.P. et l'A.M.P, est égal à* $\gamma(p) := \mathbb{P}\left(\widehat{p^A}\left(\boldsymbol{Y}\right) > p_{lim}^+\right)$.

p	p_{lim}^+							
	15%		16.4471%		16.8573%		17.6268%	
10%	0%	*(≃ 0%)*	0%	*(≃ 0%)*	0%	*(≃ 0%)*	0%	*(≃ 0%)*
14%	16.57%	*(≃ 18.11%)*	1.45%	*(≃ 1.29%)*	0.52%	*(≃ 0.46%)*	0.05%	*(≃ 0.05%)*
15%	48.06%	*(≃ 50%)*	10.07%	*(≃ 10%)*	5.17%	*(≃ 5%)*	0.86%	*(≃ 1%)*
15.1%	51.52%	*(≃ 53.52%)*	11.93%	*(≃ 11.71%)*	6.18%	*(≃ 6.03%)*	1.37%	*(≃ 1.28%)*
16%	78.95%	*(≃ 80.58%)*	34.42%	*(≃ 34.99%)*	22.89%	*(≃ 22.98%)*	7.75%	*(≃ 8.03%)*
20%	99.99%	*(≃ 100%)*	99.84%	*(≃ 99.75%)*	99.35%	*(≃ 99.35%)*	96.67%	*(≃ 96.97%)*

Le tableau ci-dessous est l'équivalent du précédent pour les Règles de Décision de la forme :

$$\textit{Décider de lancer le produit A si } \boxed{\widehat{\delta_{p^A,15\%}}\left(\boldsymbol{y}\right) > \delta_{lim}^+}.$$

Les valeurs (resp. entre parenthèses) du tableaux correspondent à $\gamma'_m(p) = \overline{\left(\widehat{\delta_{p^A,15\%}}\left(\boldsymbol{y}_{[.]}\right) > \delta^+_{lim}\right)_m}$ *(resp. à* $\gamma'_{+\infty}(p) = \gamma'(p) := \mathbb{P}\left(\widehat{\delta_{p^A,15\%}}\left(\boldsymbol{Y}\right) > \delta^+_{lim}\right)$*).*

p	δ^+_{lim}							
	0		1.281552		1.644854		2.326348	
10%	0%	*(≃ 0%)*	0%	*(≃ 0%)*	0%	*(≃ 0%)*	0%	*(≃ 0%)*
14%	16.57%	*(≃ 18.11%)*	1.45%	*(≃ 1.29%)*	0.52%	*(≃ 0.46%)*	0.05%	*(≃ 0.05%)*
15%	48.06%	*(≃ 50%)*	10.07%	*(≃ 10%)*	5.17%	*(≃ 5%)*	0.86%	*(≃ 1%)*
15.1%	51.52%	*(≃ 53.52%)*	11.93%	*(≃ 11.71%)*	6.18%	*(≃ 6.03%)*	1.37%	*(≃ 1.28%)*
16%	78.95%	*(≃ 80.58%)*	34.42%	*(≃ 34.99%)*	22.89%	*(≃ 22.98%)*	7.75%	*(≃ 8.03%)*
20%	99.99%	*(≃ 100%)*	99.84%	*(≃ 99.75%)*	99.35%	*(≃ 99.35%)*	96.67%	*(≃ 96.97%)*

Résultats expérimentaux pour le produit B : *il expérimente les Règles de Décision suivantes sur tous les* $m = 10000$ *échantillons des 6 urnes* U^B_μ *:*

$$\textit{Décider de lancer le produit B si } \boxed{\widehat{\mu^B}\left(\boldsymbol{y}\right) > \mu^+_{lim}}$$

avec μ^+_{lim} *pouvant prendre les 4 valeurs du tableau ci-dessous fournissant les différentes quantités* $\gamma_m(\mu) = \overline{\left(\widehat{\mu^B}\left(\boldsymbol{y}_{[.]}\right) > \mu^+_{lim}\right)_m}$ *(et* $\gamma_{+\infty}(\mu) = \gamma(\mu) := \overline{\left(\widehat{\mu^B}\left(\boldsymbol{Y}\right) > \mu^+_{lim}\right)_m}$*).*

μ	μ^+_{lim}							
	0.15		0.168461		0.173694		0.183511	
0.1	0.01%	*(≃ 0.01%)*	0%	*(≃ 0%)*	0%	*(≃ 0%)*	0%	*(≃ 0%)*
0.14	23.39%	*(≃ 24%)*	2.55%	*(≃ 2.22%)*	1.13%	*(≃ 0.87%)*	0.14%	*(≃ 0.11%)*
0.15	47.92%	*(≃ 50%)*	9.95%	*(≃ 10%)*	5.35%	*(≃ 5%)*	1.29%	*(≃ 1%)*
0.151	50.72%	*(≃ 52.77%)*	10.95%	*(≃ 11.27%)*	5.61%	*(≃ 5.76%)*	1.25%	*(≃ 1.2%)*
0.16	74.24%	*(≃ 75.27%)*	27.85%	*(≃ 28.17%)*	17.56%	*(≃ 17.48%)*	5.99%	*(≃ 5.42%)*
0.2	99.96%	*(≃ 99.94%)*	98.43%	*(≃ 97.91%)*	96.45%	*(≃ 95.53%)*	86.35%	*(≃ 85.64%)*

Le tableau ci-dessous est l'équivalent du précédent pour les Règles de Décision de la forme :

$$\textit{Décider de lancer le produit B si } \boxed{\widehat{\delta_{\mu^B,0.15}}\left(\boldsymbol{y}\right) > \delta^+_{lim}}.$$

Il fournit toutes les quantités $\gamma'_m(\mu) = \overline{\left(\widehat{\delta_{\mu^A,0.15}}\left(\boldsymbol{y}_{[.]}\right) > \delta^+_{lim}\right)_m}$ *(ainsi que* $\gamma'_{+\infty}(\mu) = \gamma'(\mu)$ *:* $\overline{\left(\widehat{\delta_{\mu^A,0.15}}\left(\boldsymbol{Y}\right) > \delta^+_{lim}\right)_m}$*).*

μ	δ^+_{lim}							
	0		1.281552		1.644854		2.326348	
0.1	0.01%	*(≃ ? ? ? ? ?)*	0%	*(≃ ? ? ? ? ?)*	0%	*(≃ ? ? ? ? ?)*	0%	*(≃ ? ? ? ? ?)*
0.14	23.39%	*(≃ ? ? ? ? ?)*	12.86%	*(≃ ? ? ? ? ?)*	10.15%	*(≃ ? ? ? ? ?)*	6.71%	*(≃ ? ? ? ? ?)*
0.15	47.92%	*(≃ 50%)*	8.31%	*(≃ 10%)*	3.66%	*(≃ 5%)*	0.51%	*(≃ 1%)*
0.151	50.72%	*(≃ ? ? ? ? ?)*	9.02%	*(≃ ? ? ? ? ?)*	3.81%	*(≃ ? ? ? ? ?)*	0.54%	*(≃ ? ? ? ? ?)*
0.16	74.24%	*(≃ ? ? ? ? ?)*	59.05%	*(≃ ? ? ? ? ?)*	53.52%	*(≃ ? ? ? ? ?)*	45.56%	*(≃ ? ? ? ? ?)*
0.2	99.96%	*(≃ ? ? ? ? ?)*	98.57%	*(≃ ? ? ? ? ?)*	96.46%	*(≃ ? ? ? ? ?)*	85.37%	*(≃ ? ? ? ? ?)*

1. *En vous rappelant que pour un paramètre de moyenne* μ*, la loi de probabilité de son estimateur* $\widehat{\mu}\left(\boldsymbol{Y}\right)$ *est* $N(\mu, \frac{\sigma}{\sqrt{n}})$*, donner les instructions* `R` *qui ont permis au mathématicien de déterminer les valeurs de* $\gamma(p)$ *et* $\gamma(\mu)$ *dans les tableaux relatifs aux paramètres d'intérêt* p^A *et* μ^B*.*
2. *A quoi correspondent les instructions suivantes :*

```
> p<-c(.1,.14,.15,.151,.16,.2)
> 100*pnorm(.169,p,sqrt(p*(1-p)/1000))
[1] 100.0000000  99.5890332  95.3780337  94.4054974  78.1221075   0.7127646
> mu<-p
> 100*pnorm(.169,mu,sqrt((mu*(1-mu)+2/25)/1000))
[1] 99.999994 97.974752 90.641532 89.388901 73.060810  2.269396
```

3. *En vous appuyant sur les résultats du formulaire de cours, indiquer les instructions* `R` *permettant de déterminer les valeurs de* $\gamma(15\%)$ *et* $\gamma(0.15)$ *pour les tableaux relatifs aux paramètres d'écart standardisé* $\delta_{p^A,15\%}$ *et* $\delta_{\mu^B,0.15}$.
4. *A quoi correspond l'instruction suivante :*

```
> 100*pnorm(1.645)
[1] 95.00151
```

Exercice 18 (Finalisation des Règles de Décision)

A partir des résultats expérimentaux de l'exercice précédent, on se propose de finaliser les Règles de Décision (μ *pouvant être remplacé par p pour le produit A). Les quantités* p^+_{lim}, μ^+_{lim} *et* δ^+_{lim} *sont ici appelés seuils limites.*

1. *Evaluer (approximativement en fonction de* $\gamma_m(\mu)$*) les risques d'erreurs de décision de type I (notés* $\alpha(\mu)$*) et de type II (notés* $\beta(\mu)$*). Exprimez-les en fonction de* $\gamma(\mu)$ *ou* $\gamma'(\mu)$.
2. *Quelle est la plus grande valeur possible de la somme des deux risques de type I et II ? Peut-on proposer une Règle de Décision permettant de controler simultanément tous les risques d'erreurs I et II ? Quel risque sera alors à controler ?*
3. *Quelles sont les situations (i.e. valeurs de* μ*) à envisager pour générer une erreur de décision de type I ? Elles seront appelées* ***Mauvaises situations*** *en opposition aux* ***Bonnes situations*** *qui correspondent aux valeurs de* μ *pour lesquelles l'assertion d'intérêt est vraie.*
4. *Pour quelle valeur de* μ *le risque de type I est-il maximal ? Quelle est alors la* ***pire des (mauvaises) situations*** *qui permet de controler simultanément tous les risques de type I ? Les urnes* $U^\bullet_\mu$ *correspondant à cette pire des situations sont-elles uniques ?*
5. *Proposer des Règles de Décision associées à un risque maximal de type I, noté* α*, fixé à (ou n'excédant pas)* 5%*. Même question pour* $\alpha = 1\%$, $\alpha = 10\%$.
6. *Dans la pire des situations, combien en proportion d'estimations (du paramètre d'intérêt ou du paramètre d'écart standardisé selon le cas étudié) sont plus petit que les seuils limites. Comment peut-on alors définir directement ces seuils limites ? En déduire les instructions* `R` *permettant de les obtenir.*
7. *Dans le cas où la pire des situations correspond à plusieurs urnes, est-il possible de finaliser une unique Règle de Décision associée à* $\alpha = 5\%$ *basée sur le paramètre d'intérêt ?Même question pour la Règle de Décision basée sur le paramètre d'écart standardisé. Quelles conclusions en tirez-vous sur les différentes Règles de Décision proposées précédemment ?*

Exercice 19 (Rédaction standard et abrégée)

L'industriel est disposé à acheter deux échantillons $\boldsymbol{y^A}$ *et* $\boldsymbol{y^B}$ *pour lesquels il obtient* `mean(yA)=0.204`, `mean(yB)=0.172`, `sd(yB)=0.5610087`. *Nous proposons les* ***rédactions standard*** *des corrections pour les questions :*

Est-ce que le produit est rentable au risque maximal de type I fixé à $\alpha = 5\%$ *?*

En vous appuyant sur la construction des outils d'aide à la décision proposés dans les exercices précédents, expliquer les différents ingrédients de ces rédactions standard. Notamment, à quoi correspondent $\mathbf{H_1}$*, non* $\mathbf{H_1}$ *et* $\mathbf{H_0}$ *? Ces rédactions représentent-elles de bons résumés des principales informations relatives aux outils d'aide à la décision pour les produits A et B ?*

Rédaction Standard pour Produit A

Hypothèses de test *:* $\mathbf{H}_0 : p^A = 15\%$ *vs* $\mathbf{H}_1 : p^A > 15\%$
Statistique de test sous $\mathbf{H}_0$ *:*

$$\widehat{\delta_{p^A,15\%}}\left(\boldsymbol{Y^A}\right) = \frac{\widehat{p^A}\left(\boldsymbol{Y^A}\right) - 15\%}{\sqrt{\frac{15\%\times(1-15\%)}{1000}}} \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1)$$

Règle de décision *: Accepter* $\mathbf{H}_1$ *si* $\widehat{\delta_{p^A,15\%}}\left(\boldsymbol{y^A}\right) > \delta^+_{lim,5\%}$
Conclusion *: puisqu'au vu des données,*

$$\widehat{\delta_{p^A,15\%}}\left(\boldsymbol{y^A}\right) \overset{R}{=} (\texttt{mean(yA)} - 0.15)/\texttt{sqrt}(0.15 * (1 - 0.15)/\texttt{length(yA)}) \simeq 4.78232$$
$$> \delta^+_{lim,5\%} \overset{R}{=} \texttt{qnorm}(1 - .05) \simeq 1.644854$$

on peut plutôt penser (avec un risque de 5%) que le produit A est rentable.

Rédaction Standard pour Produit B

Hypothèses de test *:* $\mathbf{H}_0 : \mu^B = 0.15$ *vs* $\mathbf{H}_1 : \mu^B > 0.15$
Statistique de test sous $\mathbf{H}_0$ *:*

$$\widehat{\delta_{\mu^B,0.15}}\left(\boldsymbol{Y^B}\right) = \frac{\widehat{\mu^B}\left(\boldsymbol{Y^B}\right) - 0.15}{\widehat{\sigma_{\widehat{\mu^B}}}\left(\boldsymbol{Y^B}\right)} \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1)$$

Règle de décision *: Accepter* $\mathbf{H}_1$ *si* $\widehat{\delta_{\mu^B,0.15}}\left(\boldsymbol{y^B}\right) > \delta^+_{lim,5\%}$
Conclusion *: puisqu'au vu des données,*

$$\begin{aligned}\widehat{\delta_{\mu^B,0.15}}\left(\boldsymbol{y^B}\right) &\overset{R}{=} (\texttt{mean(yB)} - \texttt{0.15})/\texttt{seMean(yB)} \simeq 1.24009\\ &\not> \delta^+_{lim,5\%} \overset{R}{=} \texttt{qnorm(1 - .05)} \simeq 1.644854\end{aligned}$$

on ne peut pas plutôt penser (avec un risque de 5%) que le produit B est rentable.

Rédactions Abrégées : *commenter les rédactions abrégées plus axées sur la pratique :*

Rédaction Abrégée pour Produit A

Assertion d'intérêt *:* $\mathbf{H}_1 : p^A > 15\%$
Application numérique *: puisqu'au vu des données,*

$$\begin{aligned}\widehat{\delta_{p^A,15\%}}\left(\boldsymbol{y^A}\right) &\overset{R}{=} (\texttt{mean(yA)} - \texttt{0.15})/\texttt{sqrt(0.15} * \texttt{(1} - \texttt{0.15)/length(yA))} \simeq 4.78232\\ &> \delta^+_{lim,5\%} \overset{R}{=} \texttt{qnorm(1 - .05)} \simeq 1.644854\end{aligned}$$

on peut plutôt penser (avec un risque de 5%) que le produit A est rentable.

Rédaction Abrégée pour Produit B

Assertion d'intérêt *:* $\mathbf{H}_1 : \mu^B > 0.15$
Application numérique *: puisqu'au vu des données,*

$$\begin{aligned}\widehat{\delta_{\mu^B,0.15}}\left(\boldsymbol{y^B}\right) &\overset{R}{=} (\texttt{mean(yB)} - \texttt{0.15})/\texttt{seMean(yB)} \simeq 1.24009\\ &\not> \delta^+_{lim,5\%} \overset{R}{=} \texttt{qnorm(1 - .05)} \simeq 1.644854\end{aligned}$$

on ne peut pas plutôt penser (avec un risque de 5%) que le produit B est rentable.

FICHE T.D. 5 Tests d'hypothèses

Avertissement : Dans la plupart des exercices ci-dessous traitant des tests d'hypothèses, les indications `R` sont fournies à la fois pour le quantile et la p-valeur. Dans le cadre d'un examen, notez qu'un seul type d'indication `R` est généralement fourni.

Quelques quantiles pour la loi $\mathcal{N}(0,1)$:

```
> qnorm(c(0.8,0.9,0.95,0.975,0.99,0.995))
[1] 0.8416212 1.2815516 1.6448536 1.9599640 2.3263479 2.5758293
```

Exercice 20 *Une certaine agence pour l'emploi affirme que le taux de chômage en France serait cette année inférieur à* 10%. *Une enquête auprès d'un échantillon de* 200 *personnes choisies au hasard dans la population active donne* 16 *chômeurs.*

1. *Si on envisage un risque d'erreur de première espèce (maximal) fixé à 5%, peut-on confirmer l'assertion avancée par l'agence.*
   Indication(s) R :

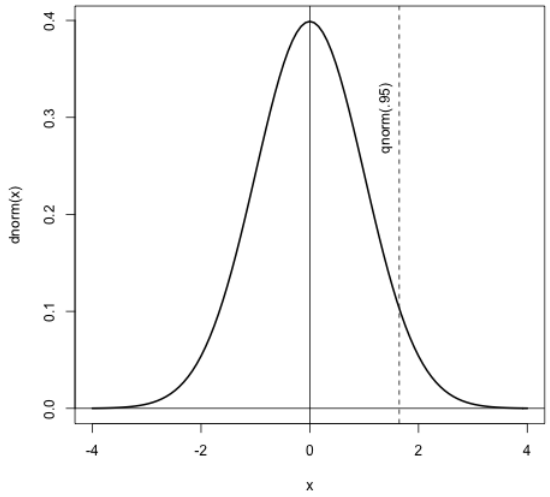
```
> 16/200
[1] 0.08
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -0.942809
> pnorm(deltaEst.H0)
[1] 0.1728893
```

2. *Peut-on pour autant montrer que l'agence pour l'emploi a effectué une mauvaise analyse ?*

Exercice 21 *(suite du produit A - Juin 2003)*

Sur vos conseils, l'industriel a lancé le produit A sur le marché (sa contrainte financière, rappelons-le, était de vendre mensuellement au moins 300000 exemplaires). Il est particulièrement satisfait car dès les deux premiers mois, il vend 350000 et 342000 exemplaires. Avec l'ambition qui le caractérise, l'industriel souhaite améliorer la qualité de son produit car selon lui beaucoup de gens le connaissent alors que finalement peu l'achètent. Il estime que s'il parvient à montrer que plus de 80% *(strictement) des individus parmi la population ciblée de taille* $N = 2000000$ *connaissent le produit A (sans pour autant l'avoir acheté) alors il pourra investir dans l'amélioration de son produit. Pour montrer ceci, il décide d'interroger au hasard* $n = 500$ *individus (la question est moins "grave" que pour le lancement du produit A d'où le choix d'un n plus petit) issus de la population mère. Les informations sont stockées dans `R` dans le vecteur `yCA` (voir ci-après). Sur cet échantillon 420 personnes (autrement dit 84% des individus interrogés) lui ont répondu connaître effectivement le produit A.*

1. *Pourriez-vous décrire littéralement les deux erreurs mises en jeu dans cette problématique ?*
2. *Peut-on montrer avec un risque d'erreur de première espèce préfixé à 5% que plus de 80% (de gens parmi la population ciblée) connaissent le produit A ? Reportez sur le graphique ci-après (représentant la densité d'une loi $\mathcal{N}(0,1)$) le quantile d'ordre 95% ainsi que la p−valeur.*



Indication(s) R :

```
> yCA
  [1] 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1
 [38] 0 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1
...
[445] 1 0 1 1 1 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1
[482] 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1
> mean(yCA)
[1] 0.84
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 2.236068
> pnorm(deltaEst.H0)
[1] 0.9873263
```

3. *Même question avec un risque d'erreur de première espèce préfixé à 1% ?*

Exercice 22 *(prime encouragement pour la qualité du produit A - contrôle continu 2003)*

Comme chaque année l'industriel afin d'encourager ses employés décide de leur verser une prime de fin d'année si la clientèle (potentielle) juge le produit A de ***bonne qualité****. Sur une échelle de valeurs entre 0 et 10 mesurant le degré de satisfaction, le produit est jugé de* ***bonne qualité*** *si la note moyenne (sur les $N = 2000000$ acheteurs potentiels), notée μ^Q, est strictement supérieure à 6 prouvant une qualité plus qu'honorable de son produit.*

Partie I (du point de vue de l'industriel) (25 min.)

1. *a) Exprimez l'assertion d'intérêt conduisant à récompenser les employés en fonction du paramètre d'intérêt μ^Q.*

   b) De quelles informations doit-on disposer pour évaluer le paramètre μ^Q ? Est-ce envisageable en pratique ?

   Pour apporter un élément de réponse, on décide d'interroger un échantillon de 200 personnes. Chacun d'eux se prononce donc sur la qualité du produit A à travers une note de 0 à 10. En `R`*, on stocke le jeu de données $\boldsymbol{y}^Q$ dans le vecteur* `yQ` *:*

```
> yQ
  [1]  9  4  3  7  6  8  2  7  3  9  9  4  8  9  4  8 10  4  9  5  2  7  2  3  2
 [26]  4  9  6 10  8  5  5  5  5 10  7  4  4  4  4  6  8  2  8  9  5  7  8  6  4
...
[151]  7 10  6  5  4  9  5  6  4  2  6  7  5  6 10  8  6  5  9  7  2  2  2  8  9
[176]  6  3  8  7  6  3  8 10  2  2  8  9 10  9  8  2  7  7 10  3  3  2  9  7  6
```

2. *a) Décrivez littéralement les deux erreurs de décision.*
   b) Précisez pour quelle(s) valeur(s) du paramètre d'intérêt μ^Q inconnu chacune de ces deux erreurs intervient.
   c) Quelle vous semble être la plus grave de ces deux erreurs ***du point de vue de l'industriel*** *?*
   d) A votre avis autour de quelles valeurs de μ^Q la décision vous semblera-t-elle difficile à prendre (cette valeur constituera la pire des situations) ?
3. *Dans la pire des situations (i.e. lorsque $\mu^Q = 6$), la mesure d'écart standardisée de test s'écrit :*
$$\widehat{\delta_{\mu^Q,6}}\left(\boldsymbol{Y^Q}\right) = \frac{\widehat{\mu^Q}\left(\boldsymbol{Y^Q}\right) - 6}{\sqrt{\frac{\widehat{\sigma^2_Q}\left(\boldsymbol{Y^Q}\right)}{200}}} \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1)$$
   Rappeler très brièvement comment on peut interpréter ce résultat via l'Approche Expérimentale.
4. *a) Si on ne souhaite excéder 5% de risque de décider à tort de récompenser les employés, établir la règle de décision et conclure quant à la bonne qualité du produit en vous aidant des quelques instructions* `R`*.*
   b) Concrètement, quelle décision doit prendre l'industriel quant au versement de la prime de fin d'année.

   Indication(s) R :

```
> mean(yQ)
[1] 6.245
> sd(yQ)
[1] 2.458699
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 1.40921
> pnorm(deltaEst.H0)
[1] 0.9206135
```

Partie II (du point de vue des employés) (25 min.)

Indépendamment des réponses précédentes un employé ayant de solides connaissances en statistiques s'interroge sur le test précédent et notamment sur les deux risques d'erreur de décision mis en jeu.

1. ***Du point de vue des employés****, quelle est la plus grave des deux erreurs de décision dans le test précédent ?*
2. *En tant que délégué, un employé exprime alors au nom de ses collègues la revendication suivante : "nous acceptons de ne pas recevoir de prime de fin d'année si vous pouvez prouver au vu du même jeu de données précédent que le produit n'est pas de bonne qualité (i.e. $\mu^Q < 6$)". Peut-on prouver au vu des données que la note moyenne du degré de satisfaction est strictement inférieure à 6 ? (Indication : rédaction standard et conclure en fournissant la valeur de la $p-$valeur)*
3. *Ce délégué désireux d'approfondir son argumentation, pose la question suivante à l'industriel : "si on avait obtenu $\widehat{\mu^Q}\left(\boldsymbol{y^Q}\right)$ et $\widehat{\sigma_Q}\left(\boldsymbol{y^Q}\right)$ de l'ordre de 6.245 et 2.459 respectivement non pas avec la taille d'échantillon actuelle de $n = 200$ mais avec une taille plus grande, pensez-vous que vous auriez pu montrer que le produit était de bonne qualité ?"*
   Et vous qu'en pensez-vous en vous aidant des instructions suivantes (en particulier si la taille d'échantillon avait été $n = 300$, $n = 500$ et $n = 1000$) ?

```
> (mean(yQ)-6)/sqrt(var(yQ)/300)
[1] 1.725923
> pnorm((mean(yQ)-6)/sqrt(var(yQ)/1000))
[1] 0.9991867
```

Exercice 23 *(compétence) Un service de contrôle d'une entreprise de métallurgie s'intéresse à savoir si un technicien est suffisamment précis au niveau des mesures quotidiennes qu'il effectue*

sur des minerais de fer (les mesures effectuées sont des mesures de diamètre). Une compétence "suffisante" que nous allons définir) sera récompensée par une prime. Le technicien sera d'autant plus précis que l'écart entre deux mesures d'un même minerai (sans qu'il le sache) est faible. Fort de constater que l'écart moyen est théoriquement nul (à justifier), la précision sera naturellement mesurée par la variance des écarts.

Soit Y^C un futur écart entre deux mesures d'un même minerai. C'est une variable aléatoire qui est caractérisée (via l'approche expérimentale) par la donnée d'une infinité d'écarts virtuels (entre deux mesures) $y^C_{[1]}, \ldots, y^C_{[m]}, \ldots$. On notera σ^2_C la variance de l'infinité de ces écarts de mesure qui constitue l'indicateur du niveau de précision du technicien. Cet indicateur constitue le paramètre d'intérêt de l'étude. Le service de contrôle décide que le technicien est compétent (et pourra ainsi lui verser une prime) si le paramètre d'intérêt est inférieur à 0.1.

1. *Au vu des données peut-on montrer que le technicien est compétent avec un risque d'erreur de première espèce fixé à 5%.*

   Indication(s) R :

```
> yC
 [1]  0.34520624  0.36187714  0.28326083 -0.04273267  0.07897429 -0.57583456
 [7] -0.71994324  0.18188198  0.04438047 -0.01951828 -0.34820050  0.26067820
...
[25]  0.31212572  0.12692537 -0.17803505 -0.17861634 -0.48143225 -0.07185419
> var(yC)
[1] 0.08843689
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -0.5474076
> pnorm(deltaEst.H0)
[1] 0.2920494
```

2. *Le service de contrôle (à qui on a fait comprendre que le risque α est généralement fixé à 5%) décide au vu de la p-valeur du précédent test de renvoyer sur le champ le technicien. Ce dernier bien plus au courant des techniques statistiques assigne son employeur aux prud'hommes et gagne le procès haut la main. L'argument avancé par le technicien est le suivant : "Le service de contrôle n'a en aucun cas prouvé que je n'étais pas compétent. Il n'a seulement pas pu prouver au vu des données que j'étais compétent : soit il prouve que je ne suis pas compétent, soit il me soumet à un nombre plus élevé d'échantillons de manière à ce que l'estimation de la variance soit significative.". Essayez de traduire mathématiquement l'argument du technicien.*

3. *Le service de contrôle décide alors de soumettre deux fois le technicien à $n = 500$ mesures sur les échantillons de minerai. Le vecteur des 500 écarts de mesures est encore noté $\mathbf{y}^C$. Avec ces nouvelles observations que concluez-vous au seuil de 5% ?*

   Indication(s) R :

```
> yC
  [1]  0.345206239  0.361877141  0.283260829 -0.042732674  0.078974292
  [6] -0.575834563 -0.719943239  0.181881984  0.044380473 -0.019518279
...
[491] -0.070185269  0.134088213  0.277789013 -0.079484256 -0.107560964
[496] -0.265767699 -0.124663757 -0.026684338  0.494449721 -0.304572345
> var(yC)
[1] 0.08299474
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -3.299262
> pnorm(deltaEst.H0)
[1] 0.0004846965
```

Exercice 24 *(diététicien) Un diététicien affirme que son régime alimentaire permet une perte de poids rapide. On observe la répartition d'un échantillon de 10 femmes ayant suivi ce régime*

pendant 2 *semaines.*

Poids avant AV	64	67	68	76	72	69	62	65	64	73
Poids après AP	65	61	64	69	65	66	60	59	61	68
Perte de poids Y	-1	6	4	7	7	3	2	6	3	5

Voici les données préliminairement saisies et placées dans deux vecteurs ***AV*** *et* ***AP****. Le vecteur* $\boldsymbol{y}$ *est obtenu en faisant une différence élément par élément :*

Indication(s) R :

```
> AV-AP
 [1] -1  6  4  7  7  3  2  6  3  5
> mean(AV-AP)
[1] 4.2
> sd(AV-AP)
[1] 2.529822
> seMean(AV-AP)
[1] 0.8
```

1. En supposant que $Y \rightsquigarrow N(\mu, \sigma)$*, éprouver l'affirmation du dié téticien au seuil de signification* 5%*.*

Indication(s) R :

```
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 5.25
> pt(deltaEst.H0,9)
[1] 0.9997362
```

2. Fort de ce constat, le diététicien aimerait un titre plus accrocheur et souhaiterait montrer avec le même jeu de données $\boldsymbol{y}$ *que son régime permet une perte de deux kilos par semaine. Éprouvez cette nouvelle affirmation au seuil de* 5%*.*

Indication(s) R :

```
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 0.25
> pt(deltaEst.H0,9)
[1] 0.5958998
```

3. Pas désappointé pour autant par la précédente analyse, le diététicien décide de soumettre 40 femmes supplémentaires à son régime, et complète ainsi son précédent vecteur $\boldsymbol{y}$ *correspondant aux pertes de poids. Il obtient finalement le jeu de données suivant :*

Indication(s) R :

```
> yD
 [1] -1  6  4  7  7  3  2  6  3  5  5  7  4  4  2  4  6  6  5  3  5  5  2  7  4
[26]  5  4  3  6  7  4  6  4  5  2  6  4  6  5  5  6  4  3  2  3  6  5  7  2  4
> mean(yD)
[1] 4.5
> sd(yD)
[1] 1.729103
> seMean(yD)
[1] 0.244532
```

a) que pensez-vous alors de l'hypothèse faite précédemment affirmant : $Y \rightsquigarrow N(\mu, \sigma)$ *?*

b) pensez-vous avec ce nouveau jeu de données que la nouvelle affirmation du diététicien est vraie (à un risque d'erreur de première espèce fixé à 5%*) ?*

Indication(s) R :

```
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 2.044722
> pnorm(deltaEst.H0)
[1] 0.9795589
```

c) Qu'expriment les erreurs standard (`seMean(y)`) pour $n = 10$ et $n = 50$? Expliquez alors pourquoi on a pu accepter l'assertion d'intérêt du diététicien pour $n = 50$.

Exercice 25 *Un pilote de course en Formule 1 hésite entre deux équipes. Il fait alors des essais dans chacune des deux équipes pour savoir laquelle est la plus performante.*

1) Pour l'équipe 1, le pilote commence par faire 20 premiers tours. Les données des temps effectués par tour sont exprimées en secondes et stockées dans le vecteur ***y1****. Au vu des données peut-on montrer que le temps moyen de la voiture de l'équipe 1 est inférieur à 51 secondes avec un risque d'erreur de première espèce fixé à 5% ?*

Indication(s) R :

```
> mean(y1)
[1] 50.21973
> sd(y1)
[1] 2.276776
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -1.532634
> pt(deltaEst.H0,19)
[1] 0.07092418
```

2) Le pilote effectue 20 tours supplémentaires (les données sont toujours stockées dans le vecteur ***y1****). Même question que précédemment avec ce nouveau jeu de données complétées.*

Indication(s) R :

```
> y1
 [1] 47.89674 50.04087 54.53240 54.36718 48.80645 51.44077 49.72669 44.81843
 [9] 50.37910 48.32037 53.55150 50.38611 50.37899 49.44585 50.47262 50.66881
...
[33] 48.83423 51.93978 49.19886 52.67034 49.21360 48.35678 49.43116 48.95199
> mean(y1)
[1] 50.39458
> sd(y1)
[1] 1.965069
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -1.948534
> pnorm(deltaEst.H0)
[1] 0.02567557
```

3) Avec la voiture de l'équipe 2, le pilote effectue 50 tours. Les données des temps effectués par tour sont exprimées en secondes et stockées dans le vecteur ***y2****. Au vu des données peut-on montrer que le temps moyen de la voiture de l'équipe 2 est inférieur à 51 secondes avec un risque d'erreur de première espèce fixé à 5% ?*

Indication(s) R :

```
> y2
 [1] 51.89371 51.35814 52.16305 51.83228 52.97653 51.43513 50.89370 51.50756
 [9] 51.54468 52.22917 51.21122 52.96252 51.61797 52.40225 50.21097 51.73468
...
[49] 53.55974 52.44708
> mean(y2)
[1] 52.02422
> sd(y2)
[1] 0.8670206
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 8.353126
```

```
> pnorm(deltaEst.H0)
[1] 1
```

4) Quel est l'ordre de grandeur de la p−valeur du précédent test ? Donc à la vue des sorties R précédentes, les données permettent-elles de laisser penser qu'une certaine assertion d'intérêt (à préciser) est vraie ?
5) Au vu des données peut-on montrer que le temps moyen de l'équipe 1 est inférieur à celui de l'équipe 2 avec un risque d'erreur de première espèce fixé à 5% ?
Cela est-il surprenant au regard le la conclusion de la question précédente ?
Indication(s) R :

```
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -4.878811
```

6) Au vu des données peut-on montrer que la variance des temps de l'équipe 2 est inférieure à celle de l'équipe 1 avec un risque d'erreur de première espèce fixé à 5% ?
Indication(s) R :

```
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 3.161661
```

7) Exprimez littéralement les deux conclusions des deux tests précédents. Si vous étiez le pilote, quelle équipe choisiriez-vous ?

Exercice 26 *(1h - environ 10 pts)*
Partie I : compétence d'un technicien

Un service de contrôle d'une entreprise de métallurgie s'intéresse à savoir si un technicien (Alfred) est suffisamment précis au niveau des mesures qu'il effectue sur des minerais de fer. Le technicien sera d'autant plus précis que l'écart entre deux mesures d'un même minerai (sans qu'il le sache) est faible. La précision serait théoriquement mesurée par la variance d'une infinité d'écarts entre deux mesures. On notera σ_A^2 (comme Alfred) cette variance. Le service de contrôle décide qu'un technicien est compétent si ce paramètre d'intérêt est inférieur à 0.1.

1. *On commence par soumettre Alfred à $n^A = 20$ échantillons de minerai de fer. Le jeu de données est stocké dans le vecteur* **yA** *dans* **R**. *Peut-on montrer au seuil de* 5% *qu'Alfred est compétent ? (précisez l'hypothèse mathématique faite sur la nature des données)*
   Indication(s) R :

```
> yA
 [1]  0.14467956  0.30839102  0.16507184  0.08100885 -0.15048984 -0.02163446
 [7] -0.25558794  0.09871536  0.59275135 -0.22962500 -0.21676732 -0.09707208
[13]  0.17050529 -0.05732366  0.65337514  0.17802469  0.29278735 -0.16514972
[19] -0.30080221  0.32129752
> var(yA)
[1] 0.07325571
> sd(yA)
[1] 0.2706579
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 13.91858
> qchisq(c(0.01,0.025,0.05,0.1,0.9,0.95,0.975,0.99),19)
[1]  7.632730  8.906516 10.117013 11.650910 27.203571 30.143527 32.852327
[8] 36.190869
> pchisq(deltaEst.H0,19)
[1] 0.2115835
```

2. *On complète le jeu de données précédent en soumettant le technicien à 30* ***nouveaux*** *échantillons. Le jeu de données est toujours stocké dans le vecteur* **yA**. *Peut-on maintenant montrer au seuil de* 5% *que le technicien est compétent ?*
   Indication(s) R :

```
> yA
 [1]  0.144679564  0.308391020  0.165071844  0.081008851 -0.150489836
 [6] -0.021634458 -0.255587943  0.098715356  0.592751352 -0.229624997
...
[41]  0.262084534 -0.215753300  0.173626815 -0.200160681 -0.255138748
[46]  0.125329351 -0.326049545  0.207517750  0.070438920 -0.303221493
> var(yA)
[1] 0.06362229
> sd(yA)
[1] 0.2522346
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -2.762438
> pnorm(deltaEst.H0)
[1] 0.002868572
```

Partie II : comparaison de deux techniciens

1. *On s'intéresse à un second technicien (Bernard) dont on cherche à montrer qu'il est compétent. Bernard a été soumis à $n^B = 20$ échantillons de minerai ; les données relatives à ses écarts de mesure sont stockées en `R` dans le vecteur `yB`. Peut-on montrer qu'il est compétent au seuil de 5% ?*

   Indication(s) R :

```
> yB
 [1]  0.024830162  0.093791329  0.145188006 -0.049699994 -0.153214255
 [6]  0.120875740 -0.112933043 -0.345291716 -0.007106278  0.122016115
[11] -0.191976511 -0.368424436  0.188209329 -0.119061948 -0.202052804
[16]  0.249518927 -0.301396393  0.112303313  0.216480111  0.239335084
> var(yB)
[1] 0.03921612
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 7.451062
```

2. *Peut-on montrer (toujours au seuil de 5%) qu'Alfred est moins précis que Bernard ?*

   Indication(s) R :

```
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 1.622351
> qf(c(0.01,0.025,0.05,0.1,0.9,0.95,0.975,0.99),49,19)
[1] 0.4350385 0.4954826 0.5546675 0.6325832 1.7129041 2.0008646 2.2983466
[8] 2.7135032
```

3. *Alfred et Bernard critiquent le précédent résultat et proposent de refaire le même test à taille d'échantillon identique (i.e. $n = 20$). Que peut-on dire à la vue de l'instruction ci-dessous ?*

   Indication(s) R :

```
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pf(deltaEst.H0,19,19)
[1] 0.9088193
```

4. *On complète alors le jeu de données du second technicien, en soumettant Bernard à 40* ***nouveaux*** *échantillons (les données sont toujours stockées en `R` dans le vecteur `yB`). Peut-on cette fois-ci montrer qu'Alfred est moins précis que Bernard ?*

   Indication(s) R :

```
> yB
 [1]  0.024830162  0.093791329  0.145188006 -0.049699994 -0.153214255
 [6]  0.120875740 -0.112933043 -0.345291716 -0.007106278  0.122016115
```

```
...
[51]  0.296275243 -0.162141374  0.060792874  0.090976915  0.119856496
[56]  0.310565975 -0.146785494  0.061968269 -0.182127778 -0.187890277
> var(yB)
[1] 0.03442929
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 2.037317
```

Exercice 27 *(conduite)*

Un expérimentateur veut savoir si les femmes conduisent mieux que les hommes au vu des notes de conduite suivantes :

Hommes *: 24,28,29,29,34,36,40,41 et 60.*

Femmes *: 21,31,34,37,38,39,42,43,44,50 et 51.*

Nous supposerons que $Y^H \rightsquigarrow N\left(\mu^H, \sigma_H\right)$ *et* $Y^F \rightsquigarrow N\left(\mu^F, \sigma_F\right)$*, et nous choisirons un seuil de signification de* 5%*.*

Indication(s) R :

```
> yH
[1] 24 28 29 29 34 36 40 41 60

> yF
 [1] 21 31 34 37 38 39 42 43 44 50 51
> mean(yH)
[1] 35.66667
> mean(yF)
[1] 39.09091
> sd(yH)
[1] 10.75872
> sd(yF)
[1] 8.561011
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -0.7935825
> pt(deltaEst.H0,18)
[1] 0.2188878
```

Traitement hors exercice pour vérifier si l'hypothèse sur l'égalité des variances de X *et* Y *n'était pas abusive :*

Indication(s) R :

```
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 1.579323
> pf(deltaEst.H0,8,10)
[1] 0.7550565
```

Exercice 28 *Situons le contexte de cette étude pour laquelle toute ressemblance avec une situation ou des personnages ayant réellement existé serait purement fortuite : nous sommes un peu* ***avant le premier tour*** *des élections présidentielles de 2001. Parmi l'ensemble des candidats, nous nous intéresserons aux trois principaux : Racchi, Pinjos et Penle. On notera dans l'ordre* p^{C1}*,* p^{C2} *et* p^{C3} *les proportions d'électeurs (parmi tout l'électorat) ayant voté pour ces trois candidats.*

Un journaliste interroge ces trois candidats séparément et leur demande le score que chacun pense faire au premier tour. Racchi pense réaliser un score autour de 20%, Pinjos autour de 19% et Penle 18%. Le journaliste n'ayant pas confiance sur la façon dont sont construits les sondages se crée un échantillon de $n = 1000$ *(choisis au hasard dans l'électorat français) et obtient les trois estimations* $\widehat{p^{C1}}(\boldsymbol{y}) = 20\%$*,* $\widehat{p^{C2}}(\boldsymbol{y}) = 15.4\%$ *et* $\widehat{p^{C3}}(\boldsymbol{y}) = 18.3\%$ *stockés en* **R** *respectivement dans les variables* `pC1Est`*,* `pC2Est` *et* `pC3Est`*.*

1. *Au vu des a priori des trois candidats et de ce que semblent penser les médias et la population, le journaliste veut alors savoir si la proportion d'électeurs votant pour*

a) Racchi est au moins de 17.5%.
b) Pinjos est au moins de 17.5 %.
c) Penle est inférieure à 17.5 %.
Formez les trois tests d'hypothèses répondant à ces questions pour un risque d'erreur de première espèce qui n'excède pas 5%. Rédigez sous forme standard le premier test.

Indication(s) R :

```
> 200/1000
[1] 0.2
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 2.080626
> pnorm(deltaEst.H0)
[1] 0.9812659
```

```
> 154/1000
[1] 0.154
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -1.747726
> pnorm(deltaEst.H0)
[1] 0.04025576
```

```
> 183/1000
[1] 0.183
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 0.6658003
> pnorm(deltaEst.H0)
[1] 0.7472306
```

2. *Pour les candidats Pinjos et Penle (qui ne sont pas sur un bateau) peut-on montrer le contraire des assertions respectives (justifiez en rédigeant succinctement) ?*
3. *La situation semblant plus préoccupante pour Pinjos et Penle, le journaliste tente de les départager en proposant de vérifier à partir du même jeu de données les assertions suivantes mais au seuil de $\alpha = 20\%$ (réponse sous la forme de rédaction abrégée) :*
   a) le score de Pinjos est inférieur à 16.5%
   b) le score de Penle est supérieur à 16.5%

   Indication(s) R :

```
> 154/1000
[1] 0.154
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -0.9371465
```

```
> 183/1000
[1] 0.183
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 1.533512
```

Exercice 29 *(rapport adjectifs-verbe)*

Un psychologue est intéressé par le rapport adjectif-verbe pour caractériser le style de discours d'un individu. Il veut alors savoir s'il y a une différence de style entre les étudiants. Un échantillon de 10 étudiants de chaque formation est choisi. Chque étudiant écrit un ensemble de textes libres. Le rapport entre le nombre de verbes et le nombre d'adjectifs utilisés par chaque est étudiant est donné dans le tableau suivant :

scientifique	*1.04*	*0.93*	*0.75*	*0.33*	*1.62*	*0.76*	*0.97*	*1.21*	*0.8*	*1.18*
litteraire	*1.32*	*2.3*	*1.98*	*0.59*	*1.02*	*0.88*	*0.92*	*1.39*	*1.95*	*1.25*

1) Peut-on plutôt penser que les scientifiques d'une part et les littéraires d'autre part utilisent plus de deux fois plus d'adjectifs que de verbes ?

Indication(s) R :

```
> sc
 [1] 1.04 0.93 0.75 0.33 1.62 0.76 0.97 1.21 0.80 1.18
> mean(sc)
[1] 0.959
> sd(sc)
[1] 0.343267
> deltaEst.H0
[1] 4.228445
> pt(deltaEst.H0,9)
[1] 0.9988942
```

```
> litt
 [1] 1.32 2.30 1.98 0.59 1.02 0.88 0.92 1.39 1.95 1.25
> mean(litt)
[1] 1.36
> sd(litt)
[1] 0.5540959
> deltaEst.H0
[1] 4.908102
> pt(deltaEst.H0,9)
[1] 0.9995809
```

2) Peut-on penser que le discours des littéraires est plus littéraire que celui des scientifiques avec un risque d'erreur de 5% ?

Indication(s) R :

```
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -1.945469
> pt(deltaEst.H0,18)
[1] 0.03375338
```

Traitement hors exercice pour vérifier si l'hypothèse sur l'égalité des variances n'était pas abusive :

Indication(s) R :

```
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 0.3837905
> pf(deltaEst.H0,9,9)
[1] 0.08494635
```

Exercice 30 (Dictée) *En 2000, une dictée (de niveau 3ème) a été proposée à un très grand nombre de futurs candidats au baccalauréat. A l'époque la note moyenne (calculée à partir de l'ensemble des candidats présents) obtenue était de* 6.3*. Un professeur de français pense (et voudrait le vérifier rapidement en soumettant 30 lycéens choisis au hasard à cette dictée) que les nouvelles méthodes d'enseignement, les nouveaux programmes, les nouvelles préoccupations des lycéens,... ont un effet sur le niveau en orthographe des bacheliers actuels.*

1. *On notera μ^D la note moyenne des bacheliers actuels soumis à la même dictée. Avec un risque d'erreur de première espèce (maximal) préfixé à 5% ? Peut-on penser, au vu des données, que l'assertion d'intérêt du professeur de français est plutôt vraie ? (rédaction standard)*

   Indication(s) R :

```
> yD
 [1]  9 10  0  1  0  5  6 10  8  1 13  9  8  3  0  0  1  0  0  0  6  9  6  8  3
[26]  5 11  5  0  0
> mean(yD)
[1] 4.566667
```

```
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -2.278765
```

2. *Avec le même risque d'erreur de première espèce, que peut-on dire de plus (pertinent) ? (une rédaction abrégée suffit)*
3. *Le professeur de français s'interroge alors sur l'hétérogénéité des étudiants. En particulier, il souhaite montrer, au vu des données, que la variance des notes (notée σ_D^2) des bacheliers actuels est supérieure à celle qui avait été obtenue en 2000 (par le très grand nombre de futurs bacheliers de l'époque) et qui valait* 10.8. *Rédigez sous forme standard un test d'hypothèses et conclure avec un risque d'erreur de première espèce (maximal) fixé à 5% ?*

Indication(s) R :

```
> var(yD)
[1] 17.35747
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 2.482891
```

4. *Même question (en ne donnant que la conclusion) avec $\alpha = 1\%$.*
5. *A la vue de l'instruction ci-dessous, quelle(s) assertion(s) peut-on confirmer en tolérant un risque d'erreur de première espèce (maximal) fixé à 5% ?*

```
> pnorm((var(yD)-12)/seVar(yD))
[1] 0.9787468
```

Exercice 31 *(machine)*

Un chef d'entreprise désire changer son ancienne machine produisant des pièces d'un certain type au rythme moyen de 1214 *pièces par jour avec un écart-type de* 35.4 *pièces par jour. Via l'Approche Expérimentale, ces caractéristiques correspondent respectivement à la moyenne et à l'écart-type des productions quotidiennes obtenues* ***si on pouvait faire tourner cette machine pendant une infinité de jours*** *(il est sous-entendu que le chef d'entreprise ne l'a constaté que pour une période de m jours avec m très très grand).*

Partie 1

Un représentant lui propose une nouvelle machine (que l'on appellera ***machine 1****) produisant des pièces du même type. Le chef d'entreprise souhaite l'acheter à condition qu'il soit assuré que cette machine 1 est d'une part plus performante (c'est à dire, de moyenne théorique une fois et demi plus grande que son ancienne machine de référence) et d'autre part plus régulière (c'est à dire, d'écart-type théorique deux fois plus petit que l'ancienne). Soient μ^{M1} et σ_{M1}, les caractéristiques de la machine 1 correspondant respectivement à la moyenne et à l'écart-type des productions quotidiennes obtenues* ***si on pouvait faire tourner la machine 1 pendant une infinité de jours****.*

1. Exprimer à partir de μ^{M1} et σ_{M1}^2 (ou σ_{M1}) les deux conditions d'achat de la nouvelle machine exprimées par le chef d'entreprise.

Pour espérer connaître les ordres de grandeur de ces quantités inconnues, le chef d'entreprise demande au représentant une période d'essai de 100 jours. Le vecteur $\boldsymbol{y^{M1}} = (y_1^{M1}, y_2^{M1}, \ldots, y_{100}^{M1})$ représente les nombres de pièces fabriquées pour chaque jour de cette période d'essai. Ce vecteur de données a été préalablement saisi dans le logiciel R sous le nom $yM1$.

2. Peut-on conseiller au chef d'entreprise d'acheter la machine 1 (lorsqu'on sait qu'il accepte que tout test statistique est généralement traité à un seuil de signification $\alpha = 5\%$) ?

Indication(s) R :

```
> yM1
  [1] 1844 1828 1837 1833 1831 1818 1836 1837 1840 1820 1845 1815 1831 1839 1824
 [16] 1839 1836 1840 1822 1824 1820 1839 1849 1846 1817 1822 1832 1846 1832 1834
```

```
...
 [76] 1810 1838 1844 1830 1830 1829 1807 1797 1814 1807 1844 1834 1827 1841 1830
 [91] 1830 1834 1840 1832 1844 1815 1825 1821 1840 1821
```

Performance

```
> mean(yM1)
[1] 1830.13
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 8.197877
> pnorm(deltaEst.H0)
[1] 1
```

Régularité

```
> var(yM1)
[1] 124.0334
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -11.41626
> pnorm(deltaEst.H0)
[1] 1.734226e-30
```

Partie 2 Le chef d'entreprise prêt à acheter la machine 1 reçoit la visite d'un second représentant vantant les mérites de sa machine (que l'on appellera bien évidemment ***machine 2****) par rapport à toutes les autres existant sur le marché. Les machines 1 et 2 étant de prix équivalents, le chef d'entreprise convaincu par l'éloquence du représentant décide de tester cette seconde machine sur une période de* $n^{M2} = 50$ *(*$< n^{M1} = 100$ *de par les réalités du calendrier que doit tenir le chef d'entreprise) jours afin de la comparer à la machine 1.*
Définissons par μ^{M2} *et* σ_{M2}*, les caractéristiques de la machine 2. On stockera les productions quotidiennes sur 50 jours dans le vecteur* y^{M2}*, préalablement saisi en* `R` *sous le nom* `yM2` *:*

```
> yM2
 [1] 2025 2045 2017 2024 2016 2025 2023 2020 2008 2025 2017 2014 2024 2028 2009
[16] 2023 2024 2034 2023 2024 2029 2032 2013 2017 2019 2022 2023 2005 2031 2012
...
[31] 2014 2032 2018 2022 2035 2024 2034 2012 2017 2015 2020 2015 2018 2020 2033
[46] 2025 2026 2026 2023 2014
```

1. a) Peut-on montrer que les productions moyennes des deux machines sont différentes ?
b) Peut-on montrer que la seconde machine produit plus (en moyenne) et plus régulièrement que la machine 1 ?
Indication(s) R : *Performance*

```
> mean(yM2)
[1] 2021.88
> mean(yM1)-mean(yM2)
[1] -191.75
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -122.3457
> pnorm(deltaEst.H0)
[1] 0
```

Régularité

```
> var(yM2)
[1] 60.80163
> var(yM1)-var(yM2)
[1] 63.2318
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
```

```
[1] 2.992739
> pnorm(deltaEst.H0)
[1] 0.9986176
```

2. a) Peut-on montrer que la seconde machine produit 190 pièces de plus par jour (en moyenne) que la première ?

b) La seconde machine est-elle 1.5 fois plus régulière (i.e. de variance 1.5 fois plus petite) que la première ?

```
> (mean(yM1)-mean(yM2)+190)/seDMean(yM1,yM2)
[1] -1.116584
> pnorm( (mean(yM1)-mean(yM2)+190)/seDMean(yM1,yM2) )
[1] 0.1320861
> (var(yM1)/var(yM2)-1.5)/seRVar(yM1,yM2)
[1] 1.044032
> pnorm((var(yM1)/var(yM2)-1.5)/seRVar(yM1,yM2))
[1] 0.8517646
> pnorm((var(yM2)/var(yM1)-1/1.5)/seRVar(yM2,yM1))
[1] 0.07782401
```

Exercice 32 *(contentement du menu d'un restaurant - Juin 2003)*

Un restaurateur s'intéresse au contentement de sa carte auprès de ses clients s'exprimant par une note entre 0 et 10. Il considère que le contentement est satisfaisant si la note moyenne (que l'on notera μ^{AV}) de l'ensemble de ses clients potentiels est strictement supérieure à 6. Pour appuyer son analyse, il interroge 40 individus et stocke les informations dans un vecteur ***yAV*** *(**les traitements R sont fournis en fin de document**).*

1. *Avec un risque d'erreur de première espèce préfixé à 5%, le restaurateur parvient-il à montrer que le niveau de satisfaction de sa carte est satisfaisant ?*

   Indication(s) R :

```
> yAV
 [1] 8 7 6 7 9 7 6 4 7 5 8 7 6 6 6 6 7 5 7 6 7 7 8 5 4 7 6 5 7 6 8 6 7 7 7 8 5 8
[39] 5 5
> mean(yAV)
[1] 6.45
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pnorm(deltaEst.H0)
[1] 0.9922594
```

   Afin d'améliorer la qualité de son établissement, le restaurateur envisage une modification de sa carte. Il décidera de maintenir cette nouvelle carte si les clients sont bien plus satisfaits qu'auparavant. Pour appuyer son analyse, il interroge rapidement 30 ***nouveaux*** *individus et stocke les informations dans le vecteur* ***yAP1*** *(voir fin exercice). On notera μ^{AP1} la note moyenne de satisfaction des clients de la nouvelle carte.*

2. *Le restaurateur juge sa nouvelle carte plus attractive si la note moyenne de satisfaction de sa clientèle a augmenté de 1 (de l'ancienne à la nouvelle carte). Peut-on le prouver au seuil de 5% ?*

   Indication(s) R :

```
> yAP1
 [1]  9  9  9  6  7  7  9  6  8  6 10 11  6  9  8  8  6 10  7 10  6  8  7  8  9
[26]  6  9  9  6  7
> mean(yAP1)
[1] 7.866667
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pnorm(deltaEst.H0)
[1] 0.8957027
```

Dans l'expectative, le restaurateur décide d'interroger les 40 clients qui s'étaient déjà prononcés sur la première carte, pensant qu'ils seraient plus à même de juger la nouvelle carte. Il reprend contact avec ces 40 clients et les invitent à se prononcer sur sa nouvelle carte. On stocke les notes associées à la nouvelle carte dans le vecteur ***yAP2*** *(voir fin exercice).*

3. *Avec un risque d'erreur de première espèce préfixé à 5%, le restaurateur parvient-il cette fois-ci à montrer que le niveau de satisfaction de sa carte a augmenté de 1 ? (Indication : faites attention au fait que les* ***mêmes*** *individus ont été interrogés)*

   Indication(s) R :

```
> yAP2
 [1]  8 10  8  8  9  8  9  5 10  6 10  7  7  6  6  7 11  7 12  6  7  7 10  6  7
[26]  7  7  5  8  8 10  6  9  8  8  9  5 10  7  8
> mean(yAP2)
[1] 7.8
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pnorm(deltaEst.H0)
[1] 0.9615129
```

Exercice 33 *(notes étudiants - Mai 2003)*

Les deux parties peuvent être traitées indépendamment l'une de l'autre. Tous les tests s'effectueront à un risque d'erreur de première espèce (maximal) ***fixé à 5%****.*

Partie I (comparaison entre contrôle continu et examen final)

A l'université, il est commun de penser que les examens sont plus difficiles que le contrôle continu et donc que la note à l'examen est en moyenne plus faible que celle du contrôle continu. Un professeur de statistiques souhaite rassurer ses étudiants, en leur montrant que certes l'examen est plus difficile mais que cette différence n'excède pas 2 points sur 20. Il interroge 50 étudiants (n'ayant pas réussi à contacter l'ensemble de l'ancienne promotion) de la section C ayant suivi son cours l'année dernière et leur demande respectivement leur note au contrôle continu ainsi qu'à l'examen. On saisit sous `R` *ces notes dans deux vecteurs notés* ***yContC*** *et* ***yExamC*** *(voir la partie indications ci-après).*

1. *Peut-on penser au vu des données que la note moyenne de l'examen (des étudiants de la section C) est strictement supérieure à 12 ?*

   Indication(s) R :

```
> yExamC
 [1] 14 17 15 13 13 13 12 16 13 15 12 14 13 15 17 15 17 13 13 12 15 14 12 13  9
[26] 10 16 11 16 13 13 14 13 15 11 16 14 10  8 15 10 12 12 12 15 10 15  9 13 11
> mean(yExamC)
[1] 13.18
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 3.806977
```

2. *Formez le test statistique laissant penser que la différence de notes entre contrôle continu et examen final (des étudiants de la section C est en moyenne strictement positive.*

   Indication(s) R :

```
> yContC-yExamC
 [1]   2  -1  -3   3  -2  -3   3   3  -1  -1   5   2  -1  -5   0   1  -5   1  -2
[20]   4   3   6   1   1   0   0   2  -4 -10  -3   4   4  -2   2   3   0  -2   2
[39]   5  -2   2   4  -1   3   4   4   1   4   4   7
> mean(yContC-yExamC)
[1] 0.84
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 1.808212
```

3. *Peut-on penser (au vu des données) que la différence moyenne entre les deux notes n'excède pas deux points ?*

   Indication(s) R :

```
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -2.497055
```

Partie II (comparaison entre deux sections d'une même promotion)

A présent, on cherche à comparer les notes d'examen de statistiques entre deux sections (les sections C et D) d'une même promotion de deuxième année de sciences économiques (toute ressemblance avec des personnages connus ou ayant existé serait purement fortuite). Ayant déjà l'information des notes de 50 étudiants de la section C (information stockée dans le vecteur ***yExamC****), on décide d'interroger (au hasard et avec remise) 50 étudiants de la section D. On range alors les notes des étudiants interrogés dans le vecteur* ***yExamD****. On décide de noter μ^C (resp. μ^D) et σ_C (resp. σ_D) les moyennes et écart-types des notes de l'ensemble de la section C (resp. D).*

1. *Montrez que la note moyenne de la section C est de plus d'un point supérieure à la note moyenne de la section D.*

   Indication(s) R :

```
> yExamD
 [1] 13 13 11 10 13 11 12 11  9 13 10 11 12 14 11 11 10 17 10  7 17 11  9 10 14
[26] 11  9 11 10 10 12 11 12 12 10  9 12 12 10 11  8  5 14  9 12 11 11  9 11 11
> mean(yExamD)
[1] 11.06
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 2.606969
```

2. *Peut-on montrer que les niveaux des étudiants en section C et D sont hétérogènes, i.e. que les variances des notes des deux sections sont* ***différentes*** *?*

   Indication(s) R :

```
> var(yExamC)
[1] 4.803673
> var(yExamD)
[1] 4.424898
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 0.2558515
```

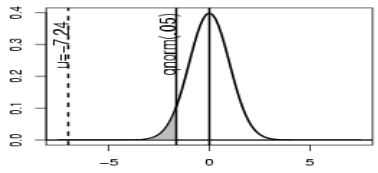
3. *Le professeur de statistiques tape dans son logiciel préféré (à préciser) les quatre instructions* `R` *suivantes calculant des* $\boldsymbol{p-}$***valeurs*** *associés à une question particulière qu'il se pose. A la seule lecture de ces quatre instructions complétez le tableau suivant (sans justification).*

```
> pnorm((mean(yExamC)-mean(yExamD)-2.5)/seDMean(yExamC,yExamD))      ## test 1
[1] 0.1882112
> pnorm((mean(yExamC)/mean(yExamD)-1.5)/seRMean(yExamC,yExamD))      ## test 2
[1] 2.220347e-13
> 1-pnorm((var(yExamC)/var(yExamD)-0.75)/seRVar(yExamC,yExamD))      ## test 3
[1] 0.1718079
> pnorm((var(yExamC)-var(yExamD)-2)/seDVar(yExamC,yExamD))           ## test 4
[1] 0.1367389
```

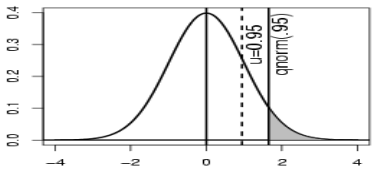
Test	*hypothèse* H_1	*Expression littérale de* H_1	*Acceptation de* H_1
			Oui
test 3			
	$H_1 : \sigma_C^2 - \sigma_D^2 < 2$		
		la différence de note moy. entre les sections C et D est-elle inférieure à 2.5 points ?	

4. *Associez à chacun des tests (1 à 4) précédents le graphique représentant la règle de décision tracée au seuil de 5% basée sur* $\widehat{\delta_{\theta,\theta_0}}(\boldsymbol{y})$ *notée* `u` *en* `R`.

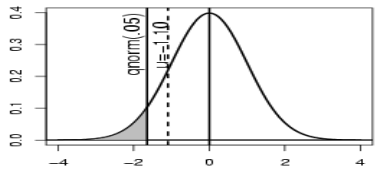
Graphique 1



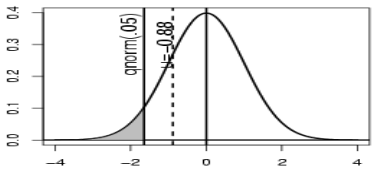
Graphique 2



Graphique 3



Graphique 4



Exercice 34 *(Effet des vacances sur la connaissance des étudiants)*

Un enseignant veut savoir si les vacances nuisent au suivi des connaissances. Il commence tout d'abord par évaluer le niveau moyen avant les vacances. Le niveau de compréhension d'un étudiant est évalué par un devoir dont la note est comprise entre 0 et 10. Il considère que le niveau moyen de compréhension est satisfaisant si la note moyenne (que l'on notera μ^{AV}*) de l'ensemble des étudiants de la promotion est strictement supérieure à 6. Pour appuyer son analyse, il n'interroge que 40 individus et stocke les informations dans un vecteur* `yAV` *(**les traitements R sont fournis en fin de document**).*

1. *Avec un risque d'erreur de première espèce préfixé à 5%, l'enseignant parvient-il à montrer (rédaction standard) que le niveau moyen de compréhension est satisfaisant ?*

   Indication(s) R :

```
> yAV
 [1] 6 7 8 7 6 9 7 7 8 9 9 5 9 8 8 9 8 5 9 5 8 7 8 8 5 8 6 8 9 7 6 9 5 6 9 9 8 5
[39] 6 5
> mean(yAV)
[1] 7.275
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pnorm(deltaEst.H0)
[1] 1
```

2. *Proposez l'instruction* `R`*, permettant d'obtenir l'intervalle de confiance au niveau de confiance 95% de la note moyenne.*

```
> # IC <- (Instruction R à fournir dans la rédaction)
> IC
[1] 6.82571 7.72429
```

3. *Avec un risque d'erreur de première espèce préfixé à 5%, l'enseignant parvient-il à montrer (rédaction standard) que l'hétérogénéité des notes est faible c'est-à-dire que la variance des notes est inférieure à 3 ?*

   Indication(s) R :

```
> var(yAV)
[1] 2.101923
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] -3.147746
```

4. ***Après les vacances****, 30* ***nouveaux*** *étudiants sont invités à passer le même devoir. Les informations sont stockées dans le vecteur* `yAP1` *(voir fin exercice). On notera* μ^{AP1} *la note moyenne obtenue à ce devoir par l'ensemble des étudiants après les vacances. L'enseignant peut-il penser (rédaction standard) , au seuil de 5%, que les vacances sont nuisibles dans le sens que le niveau moyen de compréhension a diminué de plus de deux points ?*

   Indication(s) R :

```
> yAP1
 [1] 5 5 4 7 5 7 6 5 5 4 7 7 4 4 5 5 5 5 5 6 4 6 5 5 6 5 6 5 7 5
> mean(yAP1)
[1] 5.333333
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pnorm(deltaEst.H0)
[1] 0.4198664
```

5. *Peut-on plutôt penser (rédaction abrégée) avec un risque d'erreur de 5% que la variance des notes après les vacances est plus de 2 fois plus grande qu'avant les vacances ?*

   Indication(s) R :

```
> var(yAP1)
[1] 0.9195402
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pnorm(deltaEst.H0)
[1] 0.9991686
```

6. *Pensant que le problème précédent est mal posé (puisque ce ne sont pas les mêmes étudiants qui ont passé le devoir avant et après les vacances), il demande aux 40 étudiants ayant passé le devoir* ***avant les vacances*** *de le repasser. On stocke les notes associées à dans le vecteur* `yAP2` *(voir fin exercice). Avec un risque d'erreur de première espèce préfixé à 5%, l'enseignant parvient-il à prouver (rédaction abrégée) que les vacances sont nuisibles dans le sens que le niveau moyen de compréhension a diminué de plus de deux points (et qu'il faudrait donc les supprimer) ?*

   Indication(s) R :

```
> yAP2
 [1] 2 3 5 4 2 6 4 6 7 7 6 2 6 7 6 5 4 2 5 3 7 3 7 7 3 6 3 4 5 5 4 7 2 4 8 7 7 3
[39] 5 4
> mean(yAP2)
[1] 4.825
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pnorm(deltaEst.H0)
[1] 0.9948867
```

Exercice 35 (Chiffres d'affaires) *Dans cet exercice, on va s'intéresser aux performances de l'ensemble des petites et moyennes entreprises (PME) de deux pays fictifs notés P1 et P2 en 2004 et 2005 en analysant leurs chiffres d'affaires (exprimés dans une même unité).*

Partie I :

Dans cette partie, on s'intéresse uniquement au chiffre d'affaires moyen des PME du **pays P1** *et plus précisément à son évolution au cours des années 2004 et 2005. Ne pouvant pas interroger l'ensemble des PME, on ne pourra disposer que des chiffres d'affaires sur un échantillon de PME. On commence par recueillir pour* **2004 et 2005**, *les chiffres d'affaires de 20 PME (ce sont les mêmes PME que l'on suit de 2004 à 2005). Les données sont stockées dans les vecteurs* `y04` *et* `y05`. *On propose de noter* μ^{04} *et* μ^{05} *les chiffres d'affaires annuels moyens de l'ensemble des PME du pays P1 en 2004 et 2005.*

1. *Peut-on penser que le chiffre d'affaires annuel moyen des PME en 2004 est supérieur à 80 unités ?*

   Indication(s) R :

```
> y04
 [1] 84.03 95.47 88.89 93.09 87.24 90.00 86.85 86.61 73.24 73.88 97.20 96.47
[13] 85.61 64.47 67.98 78.20 86.76 81.73 74.35 83.55
> mean(y04)
[1] 83.781
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 1.827336
> qt(c(0.9,0.95,0.975,0.99),19)
[1] 1.327728 1.729133 2.093024 2.539483
```

2. *Peut-on penser que le chiffre d'affaires annuel moyen a augmenté entre 2004 à 2005 de 10 unités ?*

   Indication(s) R :

```
> y05
 [1]  98.83  96.56  86.08  84.08  93.68 106.74  93.42 104.04  99.24  87.47
[11] 117.65 115.26 109.33  92.71 105.48  93.09 106.59  82.92  96.31  87.99
> mean(y05)
[1] 97.8735
> mean(y04-y05)
[1] -14.0925
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pt(deltaEst.H0,19)
[1] 0.06435257
```

3. *On décide de compléter les précédents jeux de données en interrogeant 20 PME supplémentaires et en recueillant leur chiffres d'affaires en 2004 et 2005. Les jeux de données sont toujours notés* `y04` *et* `y05`. *Que peut-on dire de l'assertion d'intérêt de la question précédente ?*

   Indication(s) R :

```
> y04
 [1]  84.03  95.47  88.89  93.09  87.24  90.00  86.85  86.61  73.24  73.88
[11]  97.20  96.47  85.61  64.47  67.98  78.20  86.76  81.73  74.35  83.55
[21]  85.15  76.67  87.75  84.52 104.08  72.72 101.80  87.52  86.61  89.96
[31]  76.96  95.11  70.88  89.79  87.29  83.36  73.73  79.94  91.97 100.07
> y05
 [1]  98.83  96.56  86.08  84.08  93.68 106.74  93.42 104.04  99.24  87.47
[11] 117.65 115.26 109.33  92.71 105.48  93.09 106.59  82.92  96.31  87.99
[21]  99.77 111.22 106.49 100.80 109.97  96.91  83.39 101.57 100.10 110.07
[31]  94.03 114.85 105.30 106.50  88.68 100.94  98.40 101.98 112.11  79.68
> mean(y04)
[1] 85.0375
> mean(y05)
[1] 99.50575
> mean(y04-y05)
[1] -14.46825
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
```

```
> pnorm(deltaEst.H0)
[1] 0.01285186
```

Partie II :

Dans cette partie, on va comparer les chiffres d'affaires des PME en 2005. Pour simplifier les notations, on propose de noter par μ^{P1} et σ^2_{P1} (respectivement par μ^{P2} et σ^2_{P2}) la moyenne et la variance des chiffres d'affaires de l'ensemble des PME du pays P_1 (respectivement du pays P_2) en 2005.

1. *Peut-on penser que le chiffre d'affaires annuel moyen des PME du pays P1 est de plus de 20 unités supérieur à celui du pays P2 ?*

   Indication(s) R :

```
> yP2
 [1] 63.89 72.36 88.48 74.28 71.63 82.45 67.42 76.01 74.33 77.81 71.67 72.38
[13] 80.33 77.67 67.29 73.98 65.97 76.65 74.02 88.96 71.19 81.90 75.03 80.35
[25] 86.16 73.15 73.94 63.95 79.94 59.04 67.50 77.15 74.01 77.45 78.13 74.46
[37] 96.59 80.00 78.19 72.97
> mean(yP2)
[1] 75.467
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pnorm(deltaEst.H0)
[1] 0.9825057
```

2. *Peut-on penser que les hétérogénéités (mesurées par les variances) des chiffres d'affaires annuels des PME des deux pays diffèrent ?*

   Indication(s) R :

```
> var(yP1)
[1] 94.55306
> var(yP2)
[1] 52.20786
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pnorm(deltaEst.H0)
[1] 0.9748944
```

3. *En analysant plus précisément les résultats de la question précédente, ne peut-on pas affirmer une certaine assertion d'intérêt ? (Rédaction abrégée)*

Exercice 36 (Qualité d'évaluation de correction de copies) *Dans cet exercice, on s'intéresse à la qualité d'évaluation de correction de copies.*

Partie I :

On commence par s'intéresser à l'effet de deux correcteurs différents. L'examen d'une même épreuve s'est déroulé dans deux amphis différents que l'on notera A_1 et A_2. Deux correcteurs C_1 et C_2 prennent en charge respectivement l'amphi A_1 et l'amphi A_2. Pour se faire une idée sur leur différence d'évaluation, ils commencent chacun par corriger trente copies. Les jeux de données sont stockés en `R` dans les vecteurs `yA1` et `yA2`.

1. *On notera μ^{A_1} et μ^{A_2} les moyennes des notes données respectivement par C_1 à A_1 et par C_2 à A_2. Dans le passé, C_1 est souvent passé pour un correcteur plus souple que C_2. Ceci peut-il être confirmé au vu des données en montrant que μ^{A_1} est strictement supérieure à μ^{A_2} ?*

   Indication(s) R :

```
> yA1
 [1]  5 19 15 12 10 15 10  9 12  6  8  6  2  8 10 17 12 13 11  6  9  9 11 12  8
[26]  8  9  2 10  6

> yA2
 [1] 12  8 11  6  5  5 11  5 10  7  4  5  7  3  3  6  9  0  6  0 10  6 14  5  8
[26]  7 11  5  0  3
> mean(yA1)
[1] 9.666667
```

```
> mean(yA2)
[1] 6.4
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pnorm(deltaEst.H0)
[1] 0.9996505
```

2. *Après réflexion, les deux correcteurs comprennent que la question précédente ne permet pas de répondre à la plus grande souplesse du correcteur C_1. Les correcteurs s'accordent donc sur le fait qu'ils doivent corriger les mêmes copies. Le correcteur C_2 corrige alors les trente copies déjà corrigées par C_1. Ce nouveau jeu de données sera noté* **`yA1C2`**. *Pour simplifier les notations, on notera μ^{C_1} et μ^{C_2} les moyennes des notes données par les correcteurs C_1 et C_2 respectivement. Peut-on cette fois penser que C_1 est plus souple que C_2 ?*

   Indication(s) R :

```
> yA1C2
 [1]  3 19 15 12 10 15  9  8 12  6  8  6  1  8  9 17 11 13 10  5  8  9 10 11  8
[26]  8  9  2 10  5
> mean(yA1C2)
[1] 9.233333
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pnorm(deltaEst.H0)
[1] 0.9999852
```

Partie II :

On s'interroge maintenant sur un éventuel changement d'évaluation d'une deuxième correction (par un même correcteur). Après avoir corrigé une première fois l'ensemble des copies, le correcteur C_1 décide de ***recorriger*** *les trente premières copies, ce nouveau jeu de données sera noté, en* `R`, **`yAP`**. *On renomme le jeu de données* **`yA1`** *par* **`yAV`**. *On note $Y^D = Y^{AV} - Y^{AP}$ une future différence de notes entre la première et la seconde correction ; le jeu de données associé est stocké dans le vecteur* **`yD`** *en* `R`. *On se propose de réaliser deux tests l'un portant sur la moyenne, l'autre sur la variance de la variable différence de notes*

1. *Peut-on plutôt penser que la moyenne de la différence de notes est différente de zéro (ce qui traduirait un effet sur la moyenne d'une deuxième correction) ?*

   Indication(s) R :

```
> yAP
 [1]  5 20 16 13  7 13 12  7 13  6  8  6  2  7 10 17 12 11 12  6  9  8 10 12  7
[26]  7  9  1 10  5
> yD<-yAV-yAP

> yD
 [1]  0 -1 -1 -1  3  2 -2  2 -1  0  0  0  0  1  0  0  0  2 -1  0  0  1  1  0  1
[26]  1  0  1  0  1
> mean(yD)
[1] 0.3
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pnorm(deltaEst.H0)
[1] 0.9345922
```

2. *Peut-on plutôt penser que la variance de la différence de notes est supérieure à 0.25 ?*

   Indication(s) R :

```
> var(yD)
[1] 1.182759
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pnorm(deltaEst.H0)
[1] 0.998733
```

3. *Que concluez-vous sur cette partie II ?*

Exercice 37 (Autour du produit B) *L'industriel après avoir décidé de lancer le produit B (le jeu de données de taille 1000 ayant conduit à plutôt penser que $\mu^B > 0.15$) en janvier 2004 décide d'approfondir ses connaissances en statistiques. Tour à tour, il se concentre sur trois petits problèmes.*

Partie I

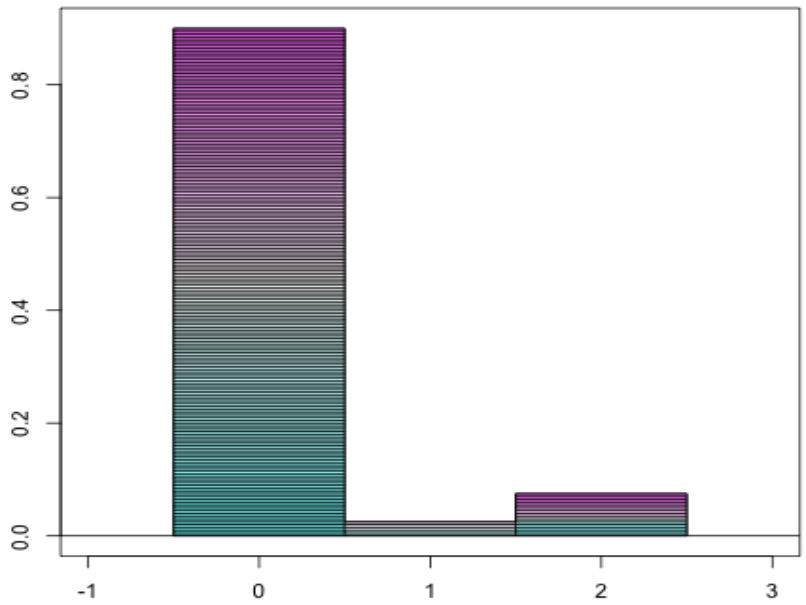
L'industriel s'interroge sur l'importance de la taille d'échantillon et se demande ce qu'il aurait conclu s'il n'avait pu disposer le jour J (avant janvier 2004) que d'un échantillon de taille bien inférieure à celle qu'il a eu en réalité à savoir 1000.

1. *Il s'imagine alors ne disposer que des nombres de produit(s) acheté(s) pour les 10 premières personnes de son échantillon de taille* 1000. *A partir des indications suivantes, quelle décision aurait-été prise par l'industriel quant au lancement de son produit ?* ***Rappelez auparavant la contrainte à imposer pour pouvoir répondre à ce type de question****.*

   Indication(s) R :

```
> y10
 [1] 1 0 0 0 0 0 0 0 0 0
> mean(y10)
[1] 0.1
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pt(deltaEst.H0,9)
[1] 0.3145356
```

2. *La contrainte à imposer vous semble-t-elle raisonnable dans cette problématique ? Justifiez brièvement votre réponse (3 lignes maximum). Le graphique ci-dessous représentant la répartition en histogramme discret des réponses des 1000 individus confirme-t-elle vos propos ?*



3. *Il convient alors d'augmenter la taille d'échantillon et de sélectionner dans son nouveau échantillon les 50 premiers individus de son échantillon original. Qu'aurait-il pu en conclure (rédaction standard) ?*

   Indication(s) R :

```
> y50
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 2 2 2 0 0 0 0 0 0 0
[39] 0 0 1 0 0 0 0 0 0 0 0 0
```

```
> mean(y50)
[1] 0.18
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> deltaEst.H0
[1] 0.3786397
```

Partie II

En Janvier 2005, cela fait un an que le produit B a été lancé. L'industriel pense que son produit mieux connu des consommateurs est plus vendu. Il dispose déjà de l'échantillon de taille $n = 1000$ *obtenu en 2004 (que l'on note ici* `yB04`*) ayant notamment servi à estimer le nombre moyen de produit B vendus en 2004, désormais noté* μ^{B04}*. En 2005, il décide d'interroger les mêmes* $n = 1000$ *individus qu'en 2004. Le jeu de données associé est noté* `yB05` *servant notamment à estimer le nombre moyen* μ^{B05} *de produit B vendus en 2005.*

1. *Peut-on penser au vu de ces données qu'il y a eu une augmentation moyenne de 2004 à 2005 d'au moins* 0.02 *en nombre de produit(s) B vendu(s) (rédaction standard) ?*

   Indication(s) R :

```
> yB05-yB04
   [1]  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  1  0  0  0  0 -2  0  0
  [25]  0  1  1  0 -2 -2 -2  0  0  0  0  0  0  0  0  0 -1  0  0  3  0  0  0  0
...
 [961]  0  0  0 -1  0  0  0  0  0  0  1  0  0  0  0  0  0  0 -1 -2  0  1  0  0
 [985]  0  0  0  0  3  1  0  0  0  0  0  0  0  0  0  0
> mean(yB05-yB04)
[1] 0.085
> # deltaEst.H0 <- (instruction R à fournir dans la rédaction)
> pnorm(deltaEst.H0)
[1] 0.991839
```

2. *En fait, après réflexion, il réalise qu'il dispose en 2005 de la valeur de* μ^{B04} *égale à* 0.19*. Reformulez l'assertion d'intérêt et proposez l'instruction* `R` *fournissant la p-valeur associée.*

Partie III

Souhaitant exporter le produit B en Allemagne, l'industriel se crée un échantillon de taille $n = 500$ *individus, consommateurs potentiels du produit B allemands. Il souhaite alors comparer les deux pays. Ce nouveau jeu de données est noté* `yAll`

On notera μ^{All} *et* σ^2_{All} *les moyennes et variances des réponses des consommateurs potentiels allemands. Pour simplifier, on notera* μ^{Fr} *et* σ^2_{Fr} *les moyennes et variances des réponses des consommateurs potentiels français, et l'échantillon de taille 1000 est noté* `yFr`*.*

A partir des instructions `R` *ci-dessous, combien d'assertions d'intérêt différentes (en précisant lesquelles) peut-on confirmer avec les données ?*

```
> (mean(yFr)-mean(yAll)-0.045)/seDMean(yFr,yAll)
[1] 1.821887
> pnorm((var(yFr)/var(yAll)-2)/seRVar(yFr,yAll))
[1] 0.996465
> pnorm((var(yAll)/var(yFr)-.5)/seRVar(yAll,yFr))
[1] 1.967895e-05
```

Annexe

A Représentations graphiques dans l'A.E.P.

Indications préliminaires

- *Objectif* : Comme nous l'avons vu à la fiche T.D. 2, l'**A.E.P.** s'appuie sur l'étude descriptive de m (grand) réalisations indépendantes $\boldsymbol{y}_{[m]} := \left(y_{[\cdot]}\right)_m$ d'une variable aléatoire d'intérêt Y. Afin d'alléger l'introduction de l'**A.E.P.**, cette étude descriptive a été volontairement limitée à une étude quantitative n'utilisant aucune représentation graphique issue de la Statistique Descriptive. Les graphiques étant d'une grande aide pour représenter les répartitions de séries de données, ils vont donc nous aider à mieux appréhender le comportement aléatoire des variables aléatoires d'intérêt.
- *Représentations graphiques usuelles* : Les représentations graphiques des répartitions diffèrent selon la nature des variables. Ainsi, lorsque la variable est de nature discrète (i.e. les modalités ou valeurs possibles sont dénombrables), on utilise un diagramme en bâton, et lorsqu'elle est de nature continue (i.e. à valeurs dans un continuum qui est non dénombrable), un histogramme est utilisé. Ce choix pose problème lorsqu'une étude expérimentale nous amène à comparer sur un même graphique des répartitions de plusieurs variables n'ayant pas la même nature. Il n'est pas possible de représenter une variable continue par un diagramme en bâton mais il en est tout autrement pour une variable discrète qui peut se représenter via un histogramme discret que nous allons introduire très prochainement.
- *Histogramme (continu)* : Rappelons les règles générales pour construire un histogramme représentant la répartition de la série $z_1, z_2, \cdots, z_m$:
  1. L'ensemble des modalités est découpé en une partition d'intervalles pas forcément de même largeur.
  2. Chaque intervalle de la partition est représenté par un rectangle ayant pour base l'intervalle et de surface égale à la proportion des $z_1, z_2, \cdots, z_m$ appartenant au dit intervalle.
  3. La somme de tous les rectangles est donc égale à 100%=1.

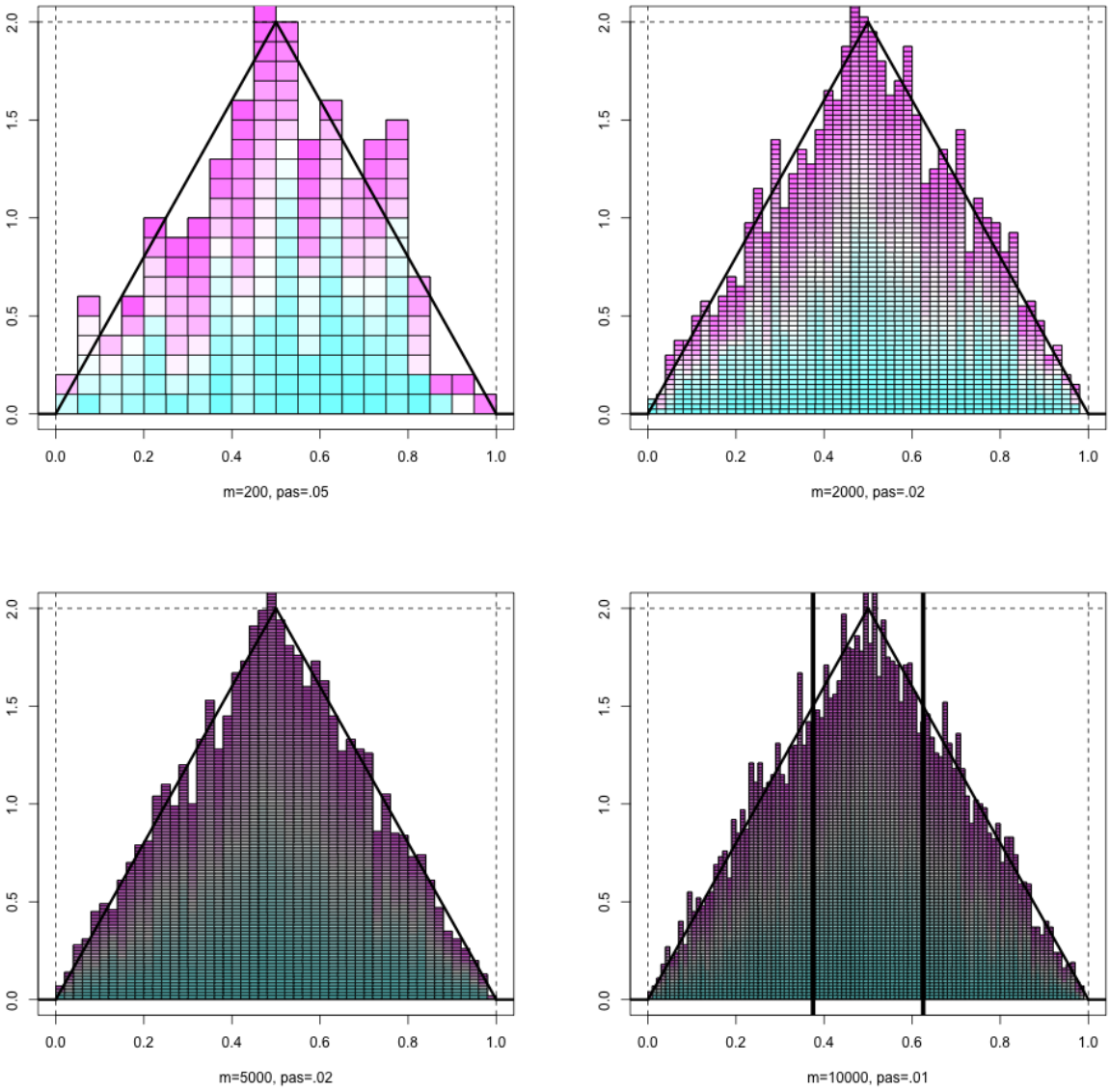
  Pour construire pratiquement un histogramme, il est conseillé au préalable de trier les $z_1, z_2, \cdots, z_m$ afin de les regrouper et ainsi de les affecter plus facilement à leurs intervalles d'appartenance. Au lieu de directement construire les rectangles associés aux intervalles de la partition, nous avons choisi d'associer à chaque donnée z_i un i$^{\text{e}}$ rectangle, que l'on appellera brique (l'histogramme étant un vu de manière imagée comme un "mur") :
  1. Toute brique associée à z_i a pour base l'intervalle d'appartenance de z_i et a une surface égale à $\frac{1}{m}$.
  2. La somme de toutes les briques associées aux $z_1, z_2, \cdots, z_m$ est donc égale à 1.
  3. Un rectangle associé à un intervalle est l'empilement de toutes les briques associés aux z_i de l'intervalle.

  Cette version de l'histogramme est équivalente à la version originale et propose en plus une représentation individualisée de toutes les données $z_1, z_2, \cdots, z_m$. Ce point particulier sera un atout pour comprendre la représentation graphique usuelle de loi de probabilités de variable aléatoire continue.
- *Histogramme discret* : Dans le même esprit, nous allons maintenant présenter la notion d'histogramme discret en l'adaptant à des données $z_1, z_2, \cdots, z_m$ à valeurs dans un espace dénombrable (donc pas dans un continuum) :
  1. Une brique associée à z_i a pour surface $\frac{1}{m}$ et sa base est centrée en z_i (à la différence d'une brique d'un histogramme continu où z_i doit appartenir à l'intervalle représentant la base de la brique).
  2. Les bases des briques sont fixées de sorte que le mur (i.e. l'histogramme) constitué de l'ensemble des briques ait le moins d'espace (trou) possible. Les briques voisines doivent se toucher dès que possible.
  3. La somme des briques est toujours égale à 100%=1 et l'empilement des briques associées aux données ayant même modalité forme un rectangle dont la surface est égale à la proportion des données égales à la modalité associée.

  IMPORTANT : Nous remarquons que dans la représentation d'histogramme (discret ou continu) les valeurs en ordonnée n'ont pas d'unité réellement interprétable. En revanche, lorsque les bases des rectangles sont toutes les mêmes, les hauteurs pourront donc être comparées entre elles pour nous éclairer sur la répartition des données puisque les aires des surfaces des rectangles (vues comme des empilements de briques) représentent naturellement des proportions de données.

Fin

Exercice 38 (Histogramme continu) *Cet exercice fait suite à l'exercice 8 mais dans l'esprit de l'exercice 9 puisqu'on s'intéresse à la moyenne plutôt que la somme. Voici 4 graphiques représentant les histogrammes continus des* $m = 200$*, 2000, 5000, 10000 premières réalisations de la variable* $M_2 := \frac{Y_1+Y_2}{2} = \frac{S}{2}$*. Nous rappelons que les* $m = 10000$ *réalisations de* S *avaient été stockées dans le vecteur* `s` *en* `R`*. Les* $m = 10000$ *réalisations* $\left(m_{2,[\cdot]}\right)_m$ *de* M_2 *sont donc accessibles en* `R` *via l'instruction* `s/2`*.*



*Dans le contexte de l'**A.E.P.**, les intervalles d'un histogramme continu peuvent sans restriction être de même largeur (car* m *est censé être suffisamment grand). Le "pas" d'un histogramme nomme en général la largeur du plus petit de ces intervalles.*

1. *Pour chaque graphique, indiquez quelle est la surface d'une brique.*
2. *Lequel de ces graphiques est le plus informatif ? A partir de ce dernier, êtes-vous en mesure de déterminer les valeurs de* M_2 *qui sont les plus probables ?*
3. *Identifiez les briques associées aux valeurs comprises entre* $\frac{3}{8}$ *et* $\frac{5}{8}$ *incluses. Quelle est la valeur de l'aire de la surface occupée par ces briques en vous rappelant que la proportion des* $\left(m_{2,[\cdot]}\right)_m$ *comprises entre* $\frac{3}{8}$ *et* $\frac{5}{8}$ *est fourni par :*

```
> mean(3/8<=s/2 &  s/2<=5/8)
[1] 0.4262
```

4. *Sauriez-vous alors évaluer approximativement la probabilité* $\mathbb{P}\left(M_2 \in [\frac{3}{8}, \frac{5}{8}]\right)$ *?*
5. *Est-il possible d'évaluer approximativement* $\mathbb{P}\left(M_2 = \frac{1}{2}\right)$ *qui via l'**A.E.P.** est approchée grâce à :*

```
> mean(s/2==1/2)
[1] 0
```

Que faudrait-il faire pour pouvoir y arriver ?

6. *En s'imaginant que le pas* $\to 0$ *au fur et à mesure que* $m \to +\infty$*, pouvez-vous décrire à quoi ressemblera une brique ? Même question pour le mur de briques ? Représentez-le sur le graphique en ne dessinant que le "dessus" (i.e. contour supérieur) du mur. Vu comme une fonction, comment interprèteriez-vous le contour supérieur du mur ?*
7. *Un mathématicien, sollicité pour nous assister dans l'étude de l'**A.M.P.**, nous apprend qu'il est classique de caractériser le comportement aléatoire de* M_2 *en fournissant la densité de probabilité (qui porte bien son nom !) s'exprimant ici mathématiquement par :*

$$f_{M_2}(t) = \begin{cases} 4t & si\ t \in [0, \frac{1}{2}] \\ 4 - 4t & si\ t \in [\frac{1}{2}, 1] \\ 0 & sinon \end{cases}$$

Représentez cette fonction sur le dernier graphique et comparez-la avec les histogrammes continus. Sont-ils très différents de la fonction ?

8. *Le mathématicien nous annonce que* $\mathbb{P}\left(M_2 \in [\frac{3}{8}, \frac{5}{8}]\right) = \int_{\frac{3}{8}}^{\frac{5}{8}} f_{M_2}(t)dt$ *qui est représentée graphiquement par la surface des points sous la courbe* $f_{M_2}(t)$ *et dont les abscisses sont compris entre* $\frac{3}{8}$ *et* $\frac{5}{8}$*. Représentez alors* $\mathbb{P}\left(M_2 \in [\frac{3}{8}, \frac{5}{8}]\right)$ *sur le graphique. Cela ne vous rappelle pas quelquechose ? Sachant qu'il n'est pas difficile de montrer que* $\mathbb{P}\left(M_2 \in [\frac{3}{8}, \frac{5}{8}]\right) = \mathbb{P}\left(S \in [\frac{3}{4}, \frac{5}{4}]\right) = \frac{7}{16} \simeq 43.75\%$ *(déjà évaluée à l'exercice 8), évaluez l'aire de la surface représentant cette probabilité.*
9. *Quelle est la valeur exacte de* $\mathbb{P}\left(M_2 = \frac{1}{2}\right)$ *?*
10. *Sélectionnez la bonne réponse parmi les réponses (proposées en suivant entre parenthèses) : La modalité* $\frac{1}{2}$ *est le mode de la loi de* M_2 *car* $\frac{1}{2}$ *est la valeur qui maximise la fonction _____ (* $\mathbf{p_{M_2}}(\mathbf{x}) := \mathbb{P}(M_2 = x)$ *ou* $\mathbf{f_{M_2}}(\mathbf{x})$ *ou* $\mathbf{F_{M_2}}(\mathbf{x}) := \mathbb{P}(M_2 \leq x)$ *). Cela se traduit littéralement par :* $\frac{1}{2}$ *est la valeur de plus grande _____(**probabilité** ou **densité de probabilité** ou **fonction de répartition**).*

A retenir

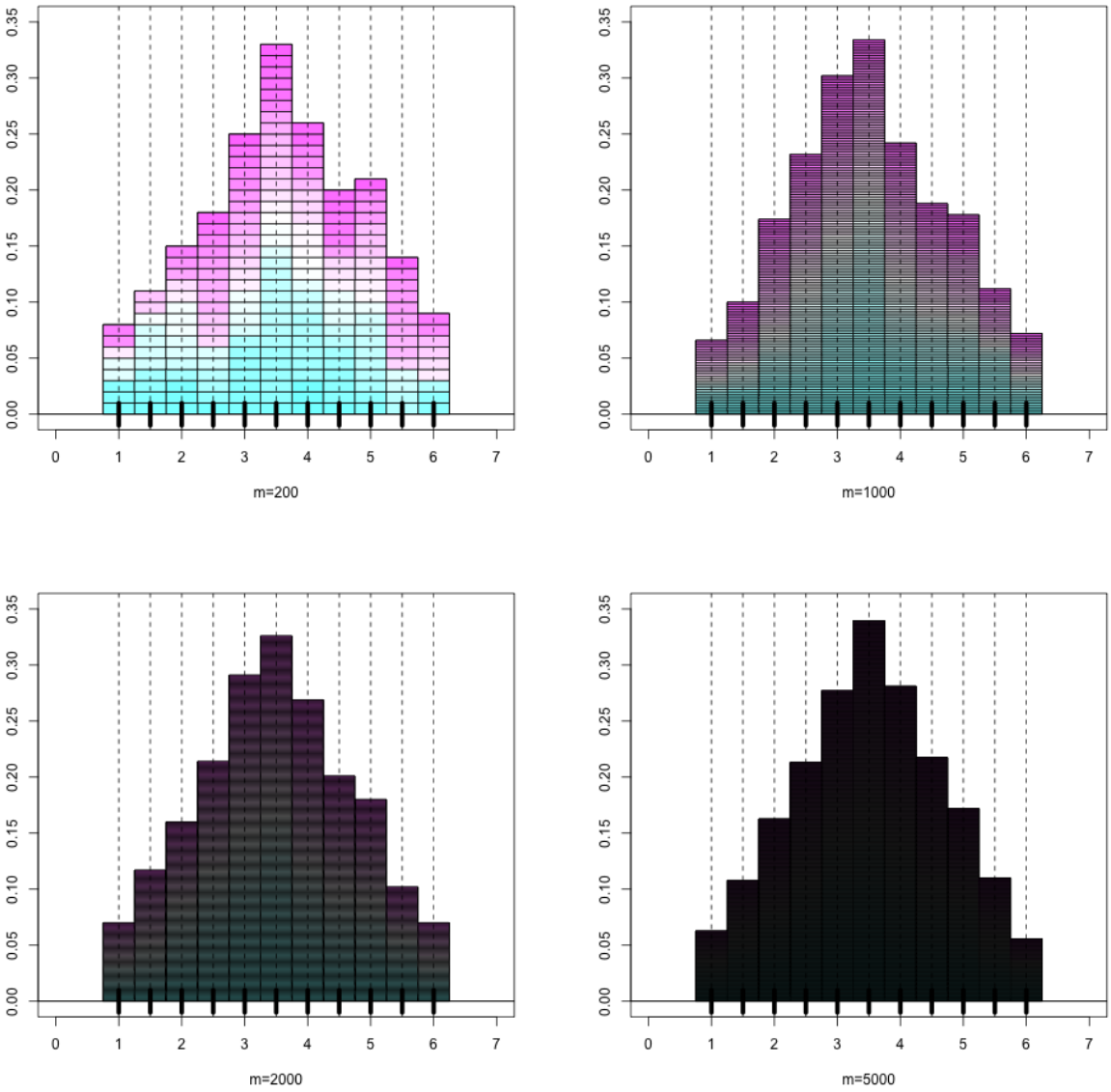
→ *La densité de probabilité caractérisant la loi de probabilité d'une v.a. continue* Y *est vue via l'**A.E.P.** comme le* `contour supérieur` *de **l'histogramme à "pas zéro" d'une infinité de ses réalisations** (i.e.* $\boldsymbol{y}_{[+\infty]} := \left(y_{[\cdot]}\right)_{+\infty}$*). De manière plus imagée, cet histogramme se décrit comme **un mur de briques devenues points** ou comme **un "tas de points"** (pour traduire la notion d'empilement) où chaque point est associé à une des réalisations. Autrement dit, tous ces objets permettent de décrire de manière très synthétique l'ensemble de "tous" les résultats possibles (représentés par les composantes de* $\boldsymbol{y}_{[+\infty]}$*) de la variable aléatoire* Y*.*

→ *La probabilité* $\mathbb{P}(Y \in [a, b]) = \int_a^b f_Y(t)dt$ *dans le contexte de l'**A.M.P.** correspond via l'**A.E.P.** à la proportion des composantes de* $\boldsymbol{y}_{[+\infty]}$ *appartenant à* $[a, b]$*. Du points de vue de l'**A.M.P.** ou de celui de l'**A.E.P.**, elle se représente par la surface occupée par les points sous la courbe* f_Y *et d'abscisses appartenant à* $[a, b]$*.*

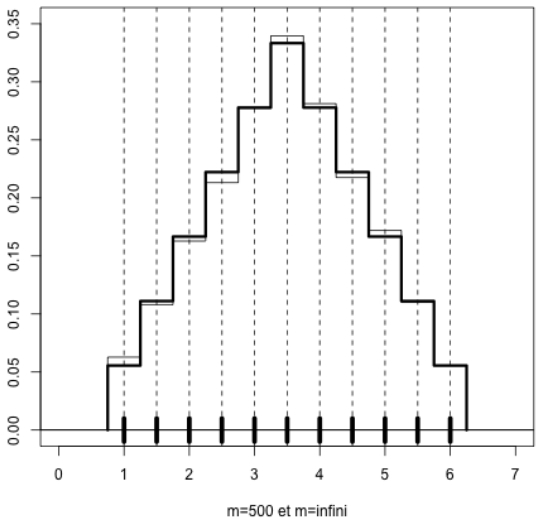
Exercice 39 (Histogramme discret) *On s'intéresse à la loi de probabilité de la moyenne* $M_2 := \frac{Y_1+Y_2}{2} = \frac{S}{2}$ *où* S *représente la somme de deux dés (introduite auparavant à l'exercice 6). Le but est ici d'introduire la notion d'histogramme discret qui n'est pas la représentation graphique la plus usuelle pour représenter une variable aléatoire discrète. Voici 4 histogrammes discrets pour*

les $m = 200$, 1000, 2000 et 5000 premières somme des deux dés. Remarquons que dans le cadre de l'exercice, nous aurions pu nous limiter à la représentation graphique usuelle en diagramme en bâton puisque nous n'aurons aucune intention de comparer la loi de probabilité de M_2 avec celle d'une loi de probabilité associée à une variable aléatoire continue. Ce sera en revanche le cas dans l'exercice 40.

Afin de mettre en avant les caractéristiques les distinguant des histogrammes continus, nous avons complété les histogrammes en identifiant les modalités de M_2 par des petits traits en gras sur l'axe des abscisses prolongés par des lignes verticales en trait pointillé.



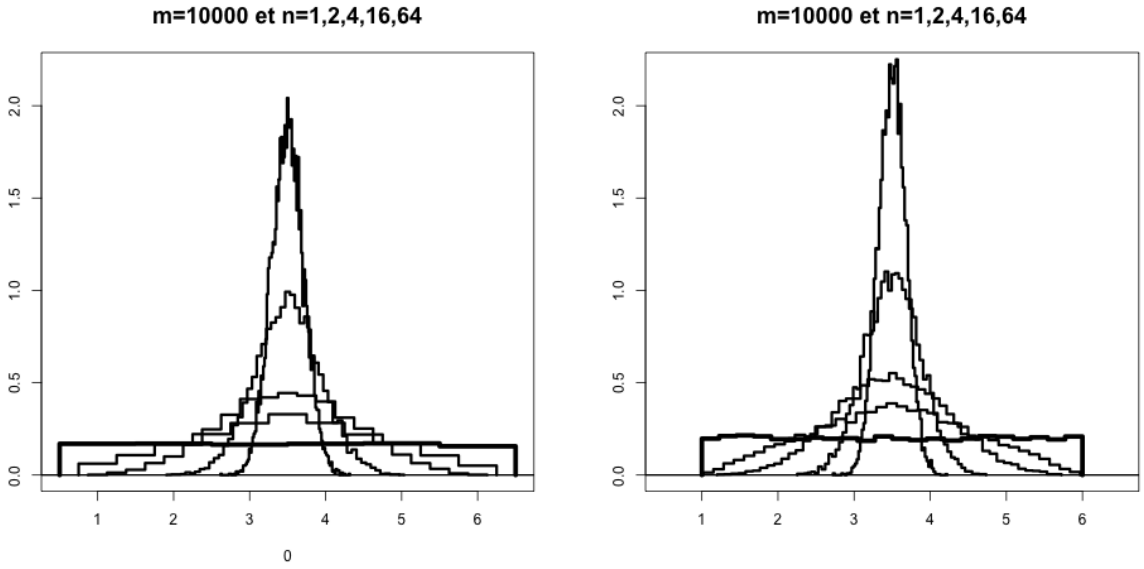
1. *Quelles sont les largeurs des briques utilisées dans ces histogrammes discrets ? Changent-elles lorsque m augmente (justifier votre réponse) ? Sur chacun des graphiques, indiquez l'aire de la surface de chaque brique.*
2. *A partir du premier graphique, donnez un ordre de grandeur de la probabilité $\mathbb{P}(M_2 = 1)$. Quelle est l'aire de la surface occupée par les briques associées aux $m_{2,[k]} = 1$? A-t'on vraiment besoin de voir les briques individuellement pour évaluer $\mathbb{P}(M_2 = 1)$? Si vous avez répondu non, appliquez cela sur le dernier graphique puisque les briques ne sont pas distinguables individuellement tellement elles sont plates.*
3. *Le graphique suivant fournit deux histogrammes discrets l'un correspondant à $m = 5000$ et l'autre à celui $m \to +\infty$. Deux types de trait (simple et en gras) ont été utilisés pour les représenter. Sauriez-vous les identifier ? Evaluez (le plus précisément possible) $\mathbb{P}(M_2 = 1)$ ainsi que $\mathbb{P}(M_2 \in \{1, 1.5, 6\})$.*



A retenir

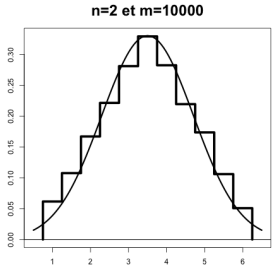
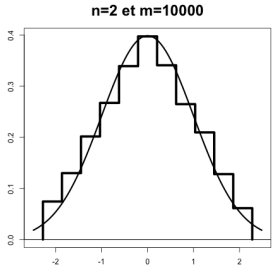
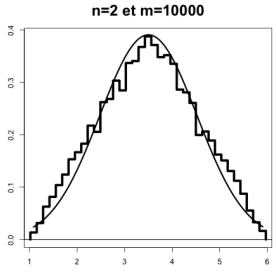
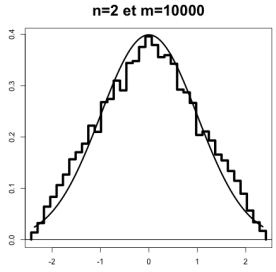
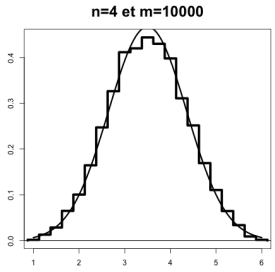
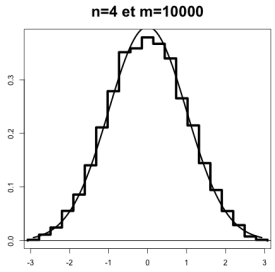
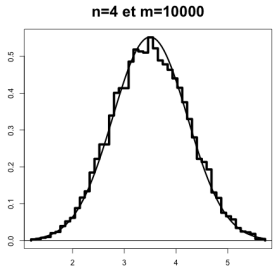
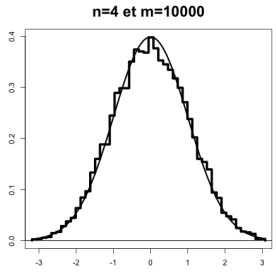
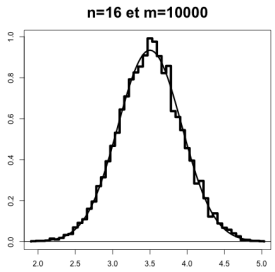
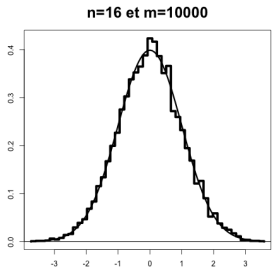
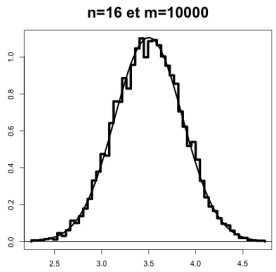
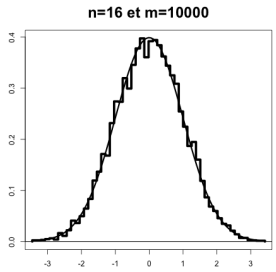
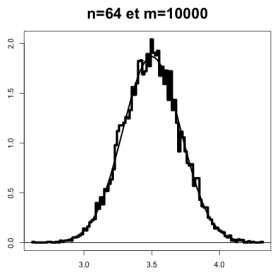
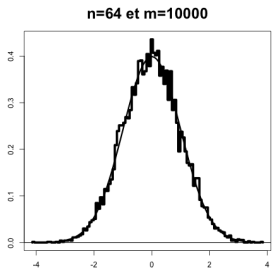
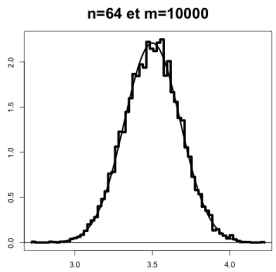
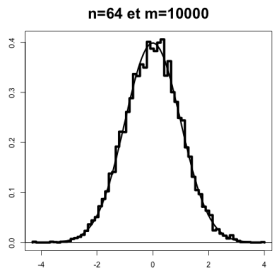
→ L'histogramme discret s'interprète de la même manière qu'un histogramme continu où les probabilités (ou proportions) sont mesurées via des aires de surfaces.
→ A la différence d'un histogramme continu, dans un histogramme discret :
- les briques sont de largeur fixe (même lorsque m varie)
- la base d'une brique n'a pas vraiment de sens
- seul le centre de la base d'une brique a un sens puisqu'il indique la valeur associée qui se lit en abscisse (surtout si la brique n'est pas la première à avoir été empilée).

Exercice 40 (Histogramme de moyenne) *L'étude menée est la suite de l'exercice 9. Ayant introduit les notions d'histogrammes discret et continu, nous allons pouvoir apprécier de visualiser le théorème de la limite centrale notamment dans le cadre de l'exemple de la moyenne de dés. Voici pour commencer, 2 graphiques représentant les contours supérieurs des histogrammes (discrets pour l'exemple du dé à gauche et continu pour l'exemple de la loi uniforme sur* $[1, 6]$ *à droite) des lois de probabilité* M_n *pour* $n = 1, 2, 4, 16, 64$. *Les échelles sont identiques pour les 2 graphiques.*



1. *Pour chaque graphique, quelle est l'histogramme qui représente via l'**A.E.P.** la loi de probabilité approximative de* Y_1 *?*

2. *Ces représentations graphiques expriment-elles le résultat que nous avions décrit sur le procédé de moyennisation qui concentre les modalités ?*
3. *Comparez les 2 graphiques. Pour quelle étude (dé ou uniforme), la moyenne est de plus grande variance ?*
4. *Sauriez-vous anticiper les histogrammes pour le cas où $n \to +\infty$ avec $m \to +\infty$?*
5. *Comme il n'est pas possible d'observer la forme de l'histogramme dans le cas précédent, il est naturel de faire comme un photographe en rezoomant le graphique de sorte à pouvoir mieux cadrer l'histogramme sur le graphique. C'est aussi ce que fait automatiquement le logiciel `R` comme on peut le voir dans la série de graphiques suivants :*

Exemple du dé : $Y_i \rightsquigarrow \mathcal{U}(\{1,2,3,4,5,6\})$		*Exemple de l'uniforme : $Y_i \rightsquigarrow \mathcal{U}([1,6])$*	
loi de M_n	*loi de Δ_n*	*loi de M_n*	*loi de Δ_n*
			
			
			
			

6. *Pour les 2 exemples et pour chaque n, comparez la forme de l'histogramme (en trait le plus épais) des réalisations de M_n avec celle de l'histogramme (en trait le plus épais) des réalisations Δ_n ? Expliquez pourquoi il en est ainsi ?*
7. *Pourquoi ces histogrammes sont-ils de plus en plus irréguliers lorsque n augmente ? Qu'aurait dû faire l'expérimentateur pour qu'il n'en soit pas ainsi ? Pouvez-vous tout de même imaginer ce qui ce serait passé lorsque $m \to +\infty$?*
8. *D'après le Théorème de la limite centrale, on peut mathématiquement affirmer, lorsque n*

est suffisamment grand (convention simplifiée appliquée dans ce cours : $n \geq 30$) :

$$M_n \overset{approx.}{\rightsquigarrow} \mathcal{N}\left(\mathbb{E}\left(Y_1\right), \sqrt{\frac{\mathbb{V}ar\left(Y_1\right)}{n}}\right) \Leftrightarrow \Delta_n := \frac{M_n - \mathbb{E}\left(Y_1\right)}{\sqrt{\frac{\mathbb{V}ar(Y_1)}{n}}} \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1).$$

Aussi, on rappelle que, pour l'exemple du dé, $\mathbb{V}ar\left(Y_1\right) = 2.9167$ et que, pour l'exemple de la loi uniforme sur $[1,6]$, $\mathbb{V}ar\left(Y_1\right) = \frac{25}{12}$.
Pour chaque graphique, que représente la courbe en trait le plus fin ? Est-elle de plus en plus ressemblante à l'histogramme en trait le plus épais lorsque n augmente (Indication : éviter de tenir compte du caractère irrégulier de l'histogramme quand n augmente uniquement dû au fait que m aurait dû être augmenté en même temps que n) ?

9. *Dans le contexte de l'**A.E.P.**, comment décririez-vous ces courbes ? Dans l'exemple du dé, les deux histogrammes représentées sur chaque graphique sont-ils de la même nature ? Avez-vous une idée sur comment illustrer graphiquement le Théorème de la limite centrale sans l'utilisation de l'histogramme discret (Indication : Une réponse très courte est bien venue) ?*
10. *Le Théorème de la limite centrale s'appliquant pour tout Y_1 ayant n'importe quelle loi de probabilité admettant une moyenne et une variance finies, imaginez la même série de graphiques que précédemment mais pour d'autres exemples que ceux (uniformes) choisis dans cette étude.*