Introduction à la décision statistique, manuel du formateur

Novembre 1997

Table des matières

Chapitre 1

Introduction à l'inférence statistique

1.1 Introduction

T 1.1 ⇒Introduction à l'inférence statistique.

Texte ⇒ Ce transparent présente les différents points qui seront abordés dans le Chapitre.

T 1.2 ⇒Populations et Échantillons

Texte ⇒ On définit deux notions importantes pour la suite du Chapitre: la notion de population et celle d'échantillon.

La notion de ***population*** est fondamentale en statistique. On appelle «population» toute **collection d'objets à étudier ayant des propriétés communes**. Ces objets sont appelés des individus ou unités statistiques.

La population peut être dénombrable (finie de petite taille) ou indénombrable (infinie ou finie et de grande taille). Dans le premier cas toute la population peut être étudiée. Dans le second cas, l'on ne peut qu'étudier un sous ensemble de la population que l'on nomme ***échantillon***.

Un échantillon est donc un sous-ensemble de la population choisie pour l'étude.

T 1.3 ⇒Les deux types de démarches statistiques.

Texte ⇒ L'étude d'une population de taille finie s'appelle un ***recensement***. Tous les paramètres de cette population sont alors bien connus ; on parle de ***statistique descriptive***.

Lorsque l'on observe qu'une partie de la population (échantillon) et que l'on cherche à étendre les propriétés constatées sur l'échantillon à toute la population d'origine et à valider ou infirmer des hypothèses *a priori*, on parle de ***statistique inférentielle***.

Remarque ⇒ Le transparent présente un schéma simplifié, il peut également arriver que l'on échantillonne dans une population finie, par exemple parce qu'elle est trop nombreuse

(ex. : sondages pré-électoraux).

1.2 Notion de variable aléatoire

1.2.1 Présentation

T 1.4 ⇒Notion de phénomène aléatoire

Texte ⇒ Présenter la partie haute du transparent de façon à voir les exemples suivants:

Ex 1 mélange à parts égales d'un produit **A** et d'un produit **B** et examen du résultat du mélange: produit **C**;

Ex 2 semis de graines dans une terrine et comptage du nombre de levées après 5 jours;

Ex 3 lancement d'une pièce de monnaie;

Ex 4 jet d'un dé et examen du nombre indiqué sur la face supérieure.

Question à poser aux stagiaires : Parmi les 4 exemples, quelles sont les expériences qui, répétées, ne conduisent pas systématiquement au même résultat?

Réponse : les 3 dernières.

Les expériences dont le résultat ne peut-être connu à l'avance à partir des conditions initiales ou d'une réalisation précédente sont dites *aléatoires*.

T 1.5 ⇒Définition d'une variable

Texte ⇒ Une *variable* X est une application d'un ensemble (Ω) d'événements dans un ensemble S de valeurs numériques ou non (appelées réalisations).

Ω est un ensemble discret d'objets, d'individus, d'occasions, ..., alors que S peut être n'importe quoi.

En particulier, les valeurs de S peuvent être *numériques*, *ordinales* ou *nominales*.

Remarque ⇒ On peut si besoin utiliser l'exemple suivant pour faire «sentir» ce que sont les ensembles Ω et S.

On dispose de 2 dés que l'on jette simultanément. On peut définir :

1. l'ensemble Ω (des évènements) comme l'ensemble de tous les couples de chiffres : $(1,1), (1,2), (1,3), \ldots, (6,6)$
2. la variable X sera, par exemple, la somme des chiffres indiqués par les 2 dés.
3. l'ensemble S comme l'ensemble des valeurs $2, 3, \ldots, 12$.

T 1.6 ⇒Exemples de variables

Texte ⇒ On distingue plusieurs types de variables en fonction des caractéristiques de l'ensemble S.

1. Les valeurs de S sont numériques : on parle alors de ***variables numériques*** ou réelles.
   - si S est dénombrable, on parle de variables ***discrètes***
   - si S est continu, les variables sont dites ***continues***.
2. S contient des éléments quelconques (non numériques). Dans ce cas, on parle de variables ***qualitatives***.
   - si S ne possède pas de structure particulière, les variables sont dites ***qualitatives nominales***.
   - si S possède une structure d'ordre, alors les variables sont dites ***qualitatives ordinales***.

On peut alors demander aux stagiaires de qualifier les variables X_1, X_2, X_3.
X_1 est une variable qualitative, X_2 est une variable ordinale et X_3 est une variable continue (discrète si le poids est arrondi au kg près).

T 1.7 ⇒Définition d'une variable aléatoire

Texte ⇒ Une *variable aléatoire* X est une variable associée à une expérience **aléatoire** et servant à caractériser le résultat de cette expérience. Autrement dit, à chaque réalisation (valeur de S) est associée

- une probabilité si la variable est discrète;
- une densité de probabilité si la variable est continue (Ces deux notions seront vues plus loin.)

Remarque ⇒ Traiter au tableau un exemple simple afin de faire «sentir» la notion de probabilité associée à l'ensemble S :

> Reprenons le premier exemple (lancer de 2 dés). A chaque jet des 2 dés est associé une valeur numérique dans $S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ et une probabilité.
>
> Calculons la probabilité de faire 2.
>
> $$P(X = 2) = 1/6 * 1/6 = 1/36$$
>
> Ainsi, on peut calculer la probabilité de chaque élément de S.
>
> Si le calcul des probabilités ne passe pas, on peut construire le tableau à deux entrées avec une entrée pour le dé bleu et une entrée pour le dé rouge. Dans chaque case on met le couple de chiffres tirés et la somme. La probabilité se calcule comme la proportion de cases «favorables».
>
> Par exemple: la valeur 2 n'apparaît qu'une seule fois sur l'ensemble des 36 cases, soit un proportion de 1/36.

La valeur 4 apparaît 3 fois sur les 36 cases soit une proportion de 3/36.

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
	2	3	4	5	6	7
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
	3	4	5	6	7	8
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
	4	5	6	7	8	9
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	5	6	7	8	9	10
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
	6	7	8	9	10	11
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)
	7	8	9	10	11	12

Remarque ⇒ Faire remarquer qu'à partir d'une même expérience, on peut calculer plusieurs variables aléatoires (somme des 2 dés, différences des 2 dés, ...).

Remarque ⇒ En général, on note les variables aléatoires par des lettres majuscules.

T 1.8 ⇒Types de VA

Texte ⇒ Les variables aléatoires les plus fréquentes que nous allons étudier sont les variables aléatoires discrètes (S est dénombrable et à chaque événement on peut associer une probabilité) et continues (S est continu et à chaque événement on peut associer une densité de probabilité).

Un ensemble est dénombrable si ses éléments peuvent être numérotés à l'aide d'entiers. Un ensemble peut donc être dénombrable et infini.

1.2.2 Variable aléatoire discrète

T 1.9 ⇒VA discrètes : Définitions

Texte ⇒ Elles sont caractérisées par le fait que l'ensemble S des réalisations possibles est dénombrable et que la probabilité P_x de réalisation de chacun des élément de S peut être calculée.

Lorsque les probabilités sont définies, on peut alors tracer 2 graphes qui vont caractériser la variable aléatoire étudiée

- ***la distribution de probabilité***
- ***la fonction de répartition***.

L'ensemble des valeurs x et des probabilités correspondantes P_x définit la *distribution de probabilité*.

L'ensemble des probabilités cumulées définit la *fonction de répartition*: $F(x) = P(X \leq x)$

T 1.10 ⇒VA discrètes : Exemple

Texte ⇒ Reprendre l'exemple du Transparent 1.7. Proposer et aider les stagiaires à tracer la fonction de distribution de probabilité et la fonction de répartition. La seconde colonne du tableau représente les probabilités, la troisième les probabilités cumulées.

Remarque ⇒ Faire remarquer que la somme des probabilités associées à l'ensemble S est égale à 1. Par conséquent, la fonction de répartition plafonne à 1.

1.2.3 Variable aléatoire continue

T 1.11 ⇒VA continues : Définition

Texte ⇒ Dans le cas de variables aléatoires continues, il n'est plus possible d'énumérer tous les éléments de S et de calculer leur probabilité.

On peut cependant calculer la probabilité d'obtenir une valeur appartenant à un intervalle donné ($P(a \leq X \leq b)$ ou inférieure à une valeur donnée ($P(X \leq x)$).

La fonction de répartition F d'une variable aléatoire continue est définie par $F(x) = P(X \leq x)$.

Si la fonction de répartition est dérivable, on appelle f (dérivée de F) la densité de la variable. Les valeurs de $F(x)$ correspondent donc à l'aire sous la courbe de densité. La surface totale sous la courbe de densité vaut 1.

Au tableau tracer une fonction de densité et montrer la surface correspondant à

- $P(a \leq X \leq b)$
- $P(x \leq X \leq x + \Delta x)$

Faire constater que $P(X = x) = 0$ (dans ce cas Δx tend vers 0)

1.3 Représentations d'une distribution

Les éléments descriptifs d'une population connue sont :

- les représentations graphiques : histogramme, courbe de densité et fonction de répartition ;
- les résumés numériques :
  - ∗ tendance centrale : espérance et médiane,
  - ∗ dispersion : variance (écart-type), étendue, écart interquartile (voire quantiles) ;

– les résumés semi-graphiques : “boîte à pattes” et “branchage”,

et ceci sur une variable mesurée.

T 1.12 ⇒Exemples: Tableaux des effectifs, effectifs cumulés, fréquences et fréquences cumulées. Sur le transparent, on a seulement representé les 8 premiers et les 7 derniers âges.

Texte ⇒ On débute par la présentation d'un premier exemple : la population des agents INRA en 1989.

En 1989, on avait 8 321 agents de l'INRA, répartis dans 22 centres régionaux. On connaît l'âge de tous ces agents. On s'intéresse à la répartition de l'âge dans cette population INRA.
Texte ⇒ En parcourant les âges de la population, on constate que l'âge minimum est de 19 ans et que l'âge maximum est de 65 ans.

Comme ces âges sont arrondis au nombre d'années le plus proche, on a donc $(65-19+1) = 47$ âges possibles.

L'âge est donc ici une variable discrète.

La première étape consiste donc à compter le nombre d'agents par âge possible ; on obtient un “tableau des effectifs” ou la “distribution des effectifs”.

De ce tableau, on peut aussi déduire le “tableau des fréquences”, ou la “distribution des fréquences”, dans lequel pour chaque classe, apparaît la fréquence calculée en divisant l'effectif de la classe par l'effectif total.

On peut enfin obtenir le “tableau d'effectifs cumulés” et aussi le “tableau des fréquences cumulées”.

Remarque ⇒ On a seulement représenté les 10 premiers et 8 derniers âges possibles car le but est de rappeler rapidement les définitions.

Un premier graphique possible représente le nombre (ou effectif) d'agents ayant un certain âge, en fonction des âges possibles (distribution des effectifs).

T 1.13 ⇒Types de représentations graphiques

Texte ⇒ Comment décrire le tableau précédent?

Réponse : On peut le décrire à l'aide de :

– représentations graphiques,

– résumés numériques,

– résumés semi-graphiques.

1.3.1 Représentations graphiques

T 1.14 ⇒Diagramme en bâtons : effectifs / âges possibles.

Texte ⇒ Ce diagramme offre peu d'intérêt, ayant trop de valeurs possibles (difficile à lire) et mettant trop en relief des particularités trop individuelles.

On est donc amené à regrouper ces âges possibles en classes, et à déterminer le nombre d'agents appartenant à chaque classe. Le graphique correspondant est l'histogramme des effectifs.

Définition : l'histogramme est un ensemble de rectangles ayant :

- comme largeur sur l'axe horizontal, l'amplitude de la classe, toutes les classes ayant même amplitude ;
- comme longueur sur l'axe vertical, l'effectif observé de la classe.

Remarque ⇒ On peut aussi, moins simplement, choisir des classes d'amplitudes différentes, mais il faut alors s'arranger pour que les aires soient proportionnelles aux effectifs des classes.

Remarque ⇒ Le nombre de classes est choisi de façon que l'aspect de l'histogramme soit représentatif de la distribution des âges (aspect régulier), sans en gommer les particularités intéressantes (on perd nécessairement de l'information quand on regroupe en classes).

En général ce nombre de classes optimal est de l'ordre du logarithme en base 2 de l'effectif total (voire racine carré de l'effectif total, si pas trop grand ...), un peu plus si on suspecte une dissymétrie possible ...

T 1.15 ⇒Histogramme des effectifs des âges des agents INRA.

Texte ⇒ Histogramme avec les effectifs en ordonnées. La hauteur de chaque rectangle est égale à l'effectif de la classe d'âge correspondante.

Ici, on a une même amplitude de classe, donc la somme de toutes les hauteurs est égale à l'effectif total.

On peut avoir l'impression qu'il n'y a pas d'agents de moins de 20 ans à l'INRA. En fait, il y a deux agents de 19 ans. Le rectangle de la classe 15-20 ans est trop petit pour être visible.

T 1.16 ⇒Polygone des fréquences, courbes de densités.

Texte ⇒ Histogramme des fréquences des âges. La proportion associée avec une classe donnée est donnée par l'aire du rectangle correspondant. L'histogramme est le même qu'au transparent précédent à l'échelle verticale près.

Texte ⇒ Sur l'histogramme la somme de toutes les aires est égale à l'unité.

On introduit ensuite le polygone des fréquences.

Définition : Le polygone des fréquences est le graphe obtenu en joignant les milieux des sommets des rectangles successifs de l'histogramme.

Si pour la variable "âge", on tenait compte non seulement du nombre d'années, mais aussi du nombre de mois, de semaines, de jours, d'heures, de minutes,... on rendrait cette variable plus continue ; donc avec un effectif total assez grand, on se rend compte que l'on finirait par

"lisser" la courbe du polygone des fréquences cumulées. On approcherait alors une courbe de densité.

Par analogie avec l'histogramme des fréquences, on constate ainsi que l'aire en dessous de la courbe de densité est égale aussi à l'unité.

T 1.17 ⇒Histogramme des âges avec les fréquences cumulées et le polygone des fréquences cumulées

Texte ⇒ Histogramme des âges avec les fréquences cumulées et le polygone des fréquences cumulées. De la même façon que pour la densité, avec un lissage du polygone des fréquences cumulées, on approche une fonction de répartition.

1.3.2 Résumés numériques

T 1.18 ⇒Tendance centrale : espérance et médiane.

Texte ⇒ On a vu au début que l'on a utilisé les valeurs minimale et maximale des âges de la population ; elles donnent les limites minimale et maximale de cette distribution ; elles ne donnent pas un résumé très précis de la distribution.

Quelles sont les valeurs qui pourraient résumer cette distribution ?

Définition mathématique de l'espérance. Sur l'histogramme, l'espérance est indiquée par la flèche.

Définition mathématique de la médiane. On a autant de chance que la variable soit au-dessus qu'en-dessous de la médiane[1]. Sur l'histogramme, la médiane est indiquée par la barre verticale en pointillée.

Si on décale une variable d'un certain facteur, alors son espérance et sa médiane sont décalées d'autant.

Si la distribution est symétrique (par ex. gaussienne), les deux paramètres de tendance centrale coïncident. Mais une (quasi-) égalité de la médiane et de l'espérance ne veut pas dire que la distribution est (quasi-) symétrique.

T 1.19 ⇒stabilité de la médiane

La médiane est un paramètre stable lorsqu'on déforme les queues d'une distribution. Sur le transparent l'histogramme du haut avec médiane et espérance indiquées correspond à la population des agents de l'INRA. Sur l'histogramme du bas avec médiane et espérance indiquées, on a déformé la distribution pour les âges plus élevés (l'effectif total est le même). La médiane reste identique après déformation ($m = 41$), par contre l'espérance est passée de 41.7 à 42.2.

1. Si on a une variable continue, la médiane est vraiment définie par

$$\Pr(X < m) = Pr(X > m).$$

Si la variable est discrète, c'est plus compliqué. Il n'y a pas alors forcément une médiane unique. Formellement, on appelle médiane tout nombre qui vérifie

$$\Pr(X < m) \leq 1/2 \leq Pr(X \leq m).$$

T 1.20 ⇒Tendance centrale: espérance – médiane.

Texte ⇒ Ces deux distributions ont même espérance et même médiane, mais des dispersions totalement différentes ; donc insuffisance des valeurs de tendance centrale pour résumer une distribution, et nécessité d'une valeur caractérisant la dispersion.

Quelles sont les mesures de dispersion dont on dispose?

Réponse : Les mesures de dispersion : variance (écart-type), étendue, écart interquartile ...

T 1.21 ⇒

La dispersion d'une variable peut être mesurée par la distribution de son écart à son espérance. Une première idée pour mesurer la dispersion d'une variable par un nombre est de prendre l'espérance de cet écart. Malheureusement, on obtient toujours 0.

Démonstration mathématique à faire éventuellement au tableau :

$$E(X - \mu) = E(X) - \mu,$$

car l'espérance de (X décalée de $-\mu$) est égale à l'espérance de X décalée de $-\mu$. Comme $E(X) = \mu$, on obtient bien 0.

Au lieu de considérer l'écart, on peut prendre son carré. On mesure alors la dispersion par l'espérance du carré de l'écart. On obtient ainsi la variance.

D'autres paramètres de dispersion sont bien sûr envisageables : l'espérance de la valeur absolue de l'écart par exemple. Mais la variance est le paramètre de dispersion le plus classique pour des raisons mathématiques (les mathématiciens préfèrent la fonction "carré" à la fonction "valeur absolue").

Un paramètre de dispersion étroitement relié à la variance est l'écart-type défini comme la racine carrée de la variance. L'écart-type a même dimension que la variable.

T 1.22 ⇒Dispersion : Étendue, quartiles et IRQ

Texte ⇒ **Définition :** L'étendue d'une série statistique est la différence entre les valeurs maximale et minimale de la série.

L'étendue dépend des valeurs extrêmes d'une variable. Or souvent ces valeurs extrêmes sont peu probables et peu donc intéressantes d'un point de vue statistique. Il est donc difficile de considérer l'étendue comme une mesure stable de la dispersion.

Texte ⇒ En vue de diminuer l'influence des valeurs extrêmes, on se propose de tenir compte d'autres valeurs plus stables de la distribution ; c'est-à-dire que l'on calcule un écart en supprimant le même pourcentage d'observations de part et d'autre de la série.

Pour obtenir la médiane, on a cherché la valeur qui sépare la distribution en 2. On peut donc reprendre la même technique que pour les deux parties obtenues, soit séparer chaque moitié en deux parties de même probabilité.

On obtient ainsi les quartiles représentés sur l'histogramme du bas par des flèches. On les note : Q_1, Q_2 (Q_2 est la médiane), et Q_3.

On peut alors définir une autre mesure de dispersion : *l'intervalle interquartile* (ou IQR).

Définition : L'intervalle interquartile (noté IQR), est l'intervalle compris entre Q_1, le premier quartile et Q_3 le troisième, il contient la moitié de la distribution.

Formule :

$$IQR = Q_3 - Q_1$$

Remarque ⇒ Cela amène à définir les quantiles en général :

T 1.23 ⇒Dispersion : les quantiles.

Définition : les quantiles d'ordre k (k nombre entier) sont les valeurs caractéristiques : $Q_1, Q_2, \ldots, Q_{k-1}$ qui divisent la distribution en k parties équiprobables.

Texte ⇒ Donc la probabilité que la variable soit inférieure à Q_i est égale à i/k.

EXEMPLES : pour les quartiles ($k = 4$) la probabilité que $X < Q_1$ est égale à $1/4$, que $X < Q_2$ à $1/2$ et que $X < Q_3$ à $3/4$. On définit de la même façon les déciles ($k = 10$), les centiles ($k = 100$) ...

Les segments ne sont pas de longueur égale (le segment i contient autant d'individus que le segment 3, tout en étant plus grand : il est moins «dense»).

De même Q_p, le p-quantile (p nombre réel compris entre 0 et 1), est la valeur telle que la probabilité que $X < Q_p$ est égale à p.

Remarque ⇒ Exemple d'utilisation de la fonction de répartition pour déterminer ces quartiles (et plus généralement les quantiles).

T 1.24 ⇒Dispersion : Les quantiles

Texte ⇒ On introduit ensuite sur ce transparent la notion de fonction de répartition inverse, notée F^{-1}, à partir des valeurs précédentes, soit (en le montrant sur le graphe) :

⇒ on a : $0,25 = F(Q_1)$ donc $Q_1 = F^{-1}(0,25)$

⇒ et de même : $0,75 = F(Q_3)$ donc $Q_3 = F^{-1}(0,75)$

On a donc une façon d'exprimer les valeurs des quartiles (et en généralisant, les valeurs des quantiles), en utilisant la fonction inverse de la fonction de répartition :

$$Q_i = F^{-1}(i/4)$$

1.3.3 Représentations semi-graphiques

T 1.25 ⇒Le *boxplot* ou "boîte à moustaches" ou "boîte à pattes".

Texte ⇒ On ne présente que 2 types : la "boîte à pattes" et le "branchage". Mentionner le nom anglais : *Boxplot* pour boîte à pattes ou à moustaches; *stem and leaf* pour branchage.

La boîte à pattes est un résumé semi-graphique qui représente la tendance centrale d'une distribution (mesurée par la médiane et/ou l'espérance) et des indices de dispersion (étendue, quartiles, IQR).

On a représenté tous ces indices sur le schéma du haut. En dessous la boîte à pattes correspondante. Les extrémités des "pattes" représentent les valeurs extrêmes. La "boîte" l'intervalle interquartile et le trait séparant la boîte en 2 la médiane.
Texte ⇒ T 1.26 ⇒Boîte à pattes pour l'âge des agents de l'INRA

T 1.27 ⇒Insuffisances des boîte à pattes

Texte ⇒ Interprétation de divers boxplots, et leur insuffisance dans divers cas (distribution creuse ou bimodale). La "boîte à moustaches" peut, prise seule, se révéler insuffisante. On montre que l'information donnée par un boxplot et un histogramme (ou un système de branchage) se complètent. Il faut donc encourager les stagiaires à tracer ces 2 choses lors de l'étude de leurs données.

On introduit donc le "branchage" pour la compléter.

T 1.28 ⇒Le *stem-and-leaf* ou "branchage" ou "branche et feuilles", à comparer avec l'histogramme.

Texte ⇒ Le stem-and-leaf a l'avantage de présenter toutes les valeurs individuelles, sous la forme d'un histogramme.

Expliquer son mode de construction : en 1^{re} colonne le 1^{er} chiffre significatif, dans les autres colonnes le 2^{e} chiffre significatif. On voit donc que le minimum est 19, que le maximum 63 est représenté par 2 individus, etc.

1.4 Estimation

T 1.29 ⇒Comparaison des histogrammes d'une population et d'un échantillon

Lorsqu'il est difficile (impossible) de faire une étude exhaustive de toute une population, on se base sur des observations partielles. Par exemple, pour étudier l'âge des agents de l'INRA, on peut réaliser un sondage. Sur le transparent, les âges d'un *échantillon* d'agents. Ces agents ont été tirés au hasard (tirage au sort de leurs matricules). L'échantillon représente $40/8321 = 0.5\%$ de la population totale.

On se retrouve là dans le cadre de la statistique inférentielle : on cherche à inférer (estimer) des caractéristiques statistiques d'une variable à partir d'observations partielles. En général le résultat d'une méthode d'estimation n'est pas exact. On a des erreurs d'estimation.

Le principe de l'estimation empirique est de faire comme si la distribution observée sur l'échantillon était représentative de la distribution totale. Comme dans notre exemple, on connaît toute la population (ce qui n'est évidemment pas le cas en général), on peut comparer le résultat de l'estimation à la réalité.

Si on compare les histogrammes obtenus à partir de l'échantillon (au dessus) et de la population (en dessous), on voit des différences sensibles. L'histogramme de l'échantillon est bimodal alors que l'histogramme basé sur la population complète est unimodal. On peut poursuivre la comparaison pour les résumés semi-graphiques et numériques ...
T 1.30 ⇒Comparaison des résumés numériques d'une population et d'un échantillon

La différence entre les 2 boîtes à pattes vient essentiellement de l'étendue qui est sous-représentée dans l'échantillon, et de la médiane excentrée (par rapport à Q_1 et Q_3). Remarque : les extrémités des «pattes» ne représentent pas toujours l'étendue mais $1,5$ fois la distance interquartile, les observations au-delà étant individualisées (traits horizontaux, symboles). Cette représentation, qui est celle utilisée par défaut par des logiciels statistiques, est appropriée pour des boîtes à pattes calculées à partir d'échantillons : elle est plus stable (moins sensible aux valeurs extrêmes) et permet aussi d'attirer l'attention sur des observations qui peuvent être aberrantes.

Pour les résumés numériques, les différences ne sont pas aussi flagrantes.

En fait, plus ce qu'on cherche à estimer est fin, plus l'erreur d'estimation sera importante.
T 1.31 ⇒Définitions : Paramètres et Estimateurs.

Texte ⇒ On appelle *paramètre* la caractéristique quantitative qui permet une représentation condensée de l'information contenue dans une ou plusieurs populations.

Texte ⇒ Faire remarquer que pour les populations infinies, les paramètres résumant les données ***ne peuvent être connus***. On va alors essayer de les ***estimer*** grâce à un échantillon.

Définition : estimer consiste à rechercher une valeur numérique approchant un paramètre inconnu d'une population, ou d'une loi de distribution, à partir de données observées sur un échantillon.

Texte ⇒ L'expression mathématique permettant de mesurer à partir des données de l'échantillon, un paramètre de la population s'appelle un *estimateur* d'un paramètre.

C'est une variable aléatoire dont on espère que la valeur sera "souvent proche" du paramètre que l'on cherche à estimer.

Exemple : pour estimer l'espérance d'une variable sur une population, on prend l'espérance calculée sur l'échantillon. Cela revient à prendre la moyenne arithmétique sur l'échantillon. Si on a observé

$$S = \{1, 1, 4, 5\},$$

alors

$$\overline{X} \quad = \quad \frac{1}{4}\,(1 + 1 + 4 + 5)$$

est un estimateur de l'espérance μ.

La formule générale de la moyenne est

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_n.$$

Texte ⇒ Un estimateur est une fonction $T_n(X_1, \ldots, X_n)$ qui dépend de toutes les observations simultanément.

Vocabulaire : $\widetilde{X}$ (médiane empirique) est un estimateur de m,
$\overline{X}$ est un estimateur de l'espérance vraie (μ).

Remarque ⇒ Une statistique est une variable aléatoire qui dépend ***uniquement des observations*** réalisées.

Remarque ⇒ Un paramètre est un "résumé" d'une loi, tandis qu'une statistique est un "résumé" d'un échantillon.

Remarque ⇒ L'objectif de la statistique inférentielle est de donner une "information" sur un paramètre, μ par exemple, en utilisant l'échantillon et par exemple un estimateur de ce paramètre ($\overline{X}$).

La nature de cette "information" est à préciser ; c'est ce que nous allons voir dans la suite.

Un estimateur ***sans biais*** est un estimateur dont l'espérance est égale à la valeur du paramètre que l'on cherche à estimer.

T 1.32 ⇒Propriétés des Estimateurs

Texte ⇒ On visualise à l'aide de cibles l'importance du biais et de la variance d'un estimateur. Le meilleur estimateur est celui qui n'est pas biaisé dont la variance est minimale.

T 1.33 ⇒Estimation: définition

Texte ⇒ L'*estimation* est la valeur prise par un estimateur pour un échantillon particulier.

L'estimation d'un paramètre à partir d'un échantillon unique ne conduit généralement pas à la vraie valeur du paramètre. Cette estimation va varier d'un échantillon à l'autre.

La réalisation d'un très grand nombre d'échantillons de même taille permet de construire la *distribution (d'échantillonnage)* de l'estimateur.

L'estimation d'un paramètre peut être *ponctuelle* ou par *intervalle*. On verra au chapitre suivant ce qu'est l'estimation par intervalle (pour les stagiaires impatients : estimation avec prise en compte de la variabilité de l'estimateur).

T 1.34 ⇒Résumé sur les notions de population et d'échantillon, et de paramètres théoriques et empiriques associés.

Texte ⇒

POPULATION	ÉCHANTILLON
fixe	aléatoire
paramètres théoriques	estimateurs (versions empiriques)
	↓
"tout est connu"	Inférence $\Longleftrightarrow$ "information" sur les paramètres de la population inconnue

T 1.35 ⇒Notations

Texte ⇒ Notations (pas très homogènes, mais c'est l'usage) :

	paramètre théorique (population)	version empirique (échantillon)	
médiane :	m	$\widetilde{X}$	(ou $\widehat{m}$)
espérance :	μ	$\overline{X}$	(ou $\widehat{\mu}$)
variance :	σ^2	$\widehat{\sigma}^2$	(ou S^2)
fonction de répartition :	F	$\widehat{F}$	

Remarque ⇒ On note souvent la taille de l'échantillon en indice des estimateurs ($\widetilde{X}_n$, $\overline{X}_n$, $\widehat{\sigma}^2_n$, $\widehat{F}_n$).

Chapitre 2

Intervalles de confiance

T 2.1 ⇒Faire le lien avec le chapitre précédent, où la notion d'estimateur a été introduite : on a estimé un paramètre (médiane ou moyenne) à l'aide d'un estimateur et on veut savoir dans quelle mesure cette estimation se rapproche de la vraie valeur du paramètre. L'estimateur était une V.A., on va chercher sa loi. On pourra alors déterminer la probabilité qu'un intervalle encadrant l'estimation du paramètre contienne sa vraie valeur, c'est l'IC.

T 2.2 ⇒On va considérer 3 méthodes de calcul de l'IC. La première que nous allons voir est connue sous le nom de la méthode du signe.

2.1 Intervalle de confiance de la médiane : méthode du signe

Méthode du signe :

- on s'intéresse au signe des (x_i - médiane)
- petite précision : on postule implicitement que la distribution des observations est continue. En effet, d'une manière générale, la médiane m d'une variable aléatoire X est définie par

$$\Pr\{X < m\} \leq \frac{1}{2} \leq \Pr\{X \leq m\}.$$

Dans le cas d'une variable aléatoire continue, cela revient à dire

$$\Pr\{X < m\} = \Pr\{X \leq m\} = \Pr\{X > m\} = \Pr\{X \geq m\} = \frac{1}{2}.$$

Mais cette dernière relation n'est pas nécessairement vérifiée par la médiane d'une variable aléatoire discrète (ex. : loi uniforme sur l'ensemble $\{-1, 0, 1\}$). Or elle est utilisée dans la construction de la méthode du signe (voir plus bas). Ce n'est peut-être pas très important d'un point de vue pédagogique mais il vaut mieux le savoir.

Résumé ⇒

Comment trouver un intervalle de confiance de la médiane?
Coefficient de confiance associé à cet intervalle.

Utilisation de tables et calcul de tables.
Intervalles bilatéral et unilatéral.

2.1.1 Présentation de la méthode

À partir de l'échantillon, nous avons un estimateur $\widetilde{X}$ de la médiane m. Peut-on avoir une idée de la précision avec laquelle on approche la médiane m?

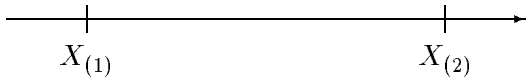
$\Rightarrow$ Objectif: encadrer m, c'est-à-dire proposer un intervalle, et savoir quelle est la probabilité que cet intervalle contienne m.

T 2.3 $\Rightarrow$Quel intervalle proposer, et comment calculer la probabilité associée?

Supposons que l'on ait un 2-échantillon X_1, X_2.

Pour encadrer m, on propose $X_{(1)} \leq m \leq X_{(2)}$ où $X_{(1)} = \min(X_1, X_2)$ et $X_{(2)} = \max(X_1, X_2)$.

Comment calculer la probabilité que l'intervalle proposé contienne m?



Calcul de $\Pr\{X_{(1)} \leq m \leq X_{(2)}\}$.

Ecrire la démonstration qui suit (points 1, 2 et 3) au tableau, elle devra y rester pour servir de guide à la démonstration suivante que les stagiaires referont seuls.

1. Calculer $\Pr\{m < X_{(1)}\}$

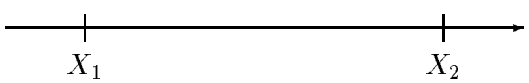
   Dire que $m < X_{(1)}$ équivaut à dire que $m < X_1$ et $m < X_2$.

   Les tirages sont *indépendants*, donc $\Pr\{m < X_{(1)}\} = \Pr\{m < X_1\} \Pr\{m < X_2\}$
   (car dans ce cas, $\Pr\{A \text{ et } B\} = \Pr\{A\} \Pr\{B\}$).

   Par définition de la médiane, $\Pr\{m < X_1\} = \Pr\{m < X_2\} = \frac{1}{2}$.

   La probabilité cherchée vaut donc $\Pr\{m < X_{(1)}\} = \frac{1}{2} \times \frac{1}{2} = \left(\frac{1}{2}\right)^2$.

2. De la même façon, on peut calculer $\Pr\{m > X_{(2)}\} = \left(\frac{1}{2}\right)^2$.

3. 

L'événement $\{$*la droite contient* $m\}$ est un événement certain, et il équivaut à :

$$\{m < X_{(1)}\} \text{ ou } \{X_{(1)} \leq m \leq X_{(2)}\} \text{ ou } \{m > X_{(2)}\}$$

et donc :

$$1 = \Pr\{m < X_{(1)}\} + \Pr\{X_{(1)} \leq m \leq X_{(2)}\} + \Pr\{m > X_{(2)}\}$$

Référence à $\Pr\{A \cup B\}$ sur fiche rappels de probabilités.

D'où :

$$\Pr\{X_{(1)} \leq m \leq X_{(2)}\} = 1 - \Pr\{m < X_{(1)}\} - \Pr\{m > X_{(2)}\}$$

$$= 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 1 - 2\left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

Conclusion : avec cette méthode, 2-échantillon et intervalle $\left[X_{(1)}\,;\ X_{(2)}\right]$, on propose un intervalle qui contient m en moyenne une fois sur deux.

Vérification empirique : normalement, à peu près un stagiaire sur deux doit avoir un intervalle ne contenant pas la vraie valeur (41,2).

Comment améliorer? (Susciter la réponse « Augmenter la taille de l'échantillon »).

T 2.4 ⇒Échantillon plus grand : par exemple, pour un 9-échantillon, on va proposer $\left[X_{(1)}\,;\ X_{(9)}\right]$, où $X_{(1)}$ est la plus petite valeur et $X_{(9)}$ est la plus grande valeur ; et pour un n-échantillon $\left[X_{(1)}\,;\ X_{(n)}\right]$.

La probabilité que cet intervalle contienne m vaut $\gamma = \Pr\{X_{(1)} \leq m \leq X_{(n)}\}$ (introduire la notation γ).

Faire calculer γ (laisser 5 minutes).

Procéder par analogie avec la démonstration ci-dessus (qui a été laissée au tableau).

Donner la piste :

$$\Pr\{m < X_{(1)}\} = \Pr\{m < X_1\}\ \Pr\{m < X_2\} \ldots \Pr\{m < X_n\}$$

d'où $\gamma = 1 - 2\left(\frac{1}{2}\right)^n = 1 - \left(\frac{1}{2}\right)^{n-1}$.

D'où la table 2.1 avec $\gamma = 1 - \left(\frac{1}{2}\right)^{n-1}$, l'erreur vaut $\left(\frac{1}{2}\right)^{n-1}$.

n	erreur $= \left(\frac{1}{2}\right)^{n-1}$	$\gamma = 1 - \left(\frac{1}{2}\right)^{n-1}$
1		
2	0,50	0,50
3	0,25	0,75
4	*	0,875
5	0,062 5	0,937 5
6	0,031 25	0,968 75
7	0,015 62	*
8	0,007 81	0,992 19
9	0,003 91	0,996 09
10	0,001 95	0,998
11	0,000 98	0,999 02
12	*	*
13	0,000 24	0,999 76
14	0,000 12	0,999 88
15	0,000 061	0,999 939

TAB. 2.1 - *Coefficient de confiance de l'intervalle* $[X_{(1)}\,;\ X_{(n)}]$

Distribuer la table 2.1 et commenter :

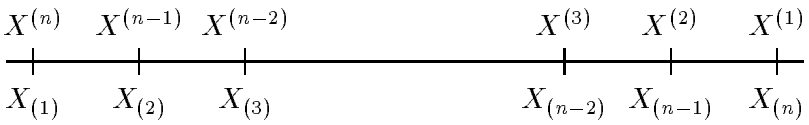
Définition de γ = probabilité que l'intervalle choisi contienne le paramètre m = **coefficient de confiance.**

On note l'erreur $\alpha = 1 - \gamma$. Faire compléter la table.

Donc pour un 9-échantillon par exemple, si on propose $\left[X_{(1)}\,;\ X_{(9)}\right]$ il y a 99,6 % de chances que cet intervalle contienne m. Précision *très élevée* : est-ce bien nécessaire?

Proposer autre chose. Modifier cet intervalle.

Introduire la notation $X_{(n)}$, $X^{(n)}$: schéma (*à faire au tableau et à laisser au tableau).*



Les $X_{(i)}$ et $X^{(i)}$ sont appelées **statistiques d'ordre**. On notera :

$$\begin{array}{lcll} X_{(n)} & = & X^{(1)} & = \text{ la plus grande valeur} \\ X^{(n)} & = & X_{(1)} & = \text{ la plus petite valeur} \\ X_{(c)} & = & X^{(n+1-c)} & = \text{ la } c\text{-ième plus petite valeur} \\ X^{(c)} & = & X_{(n+1-c)} & = \text{ la } c\text{-ième plus grande valeur} \end{array}$$

Questions pour illustrer les statistiques d'ordre :

Intervalle $\left[X_{(2)}\,;\ X^{(2)}\right]$ $\left\{\begin{array}{l}\text{plus court} \\ \\ \text{plus long}\end{array}\right.$ que $\left[X_{(1)}\,;\ X^{(1)}\right]$? *(Réponse : plus court).*

La probabilité que l'intervalle $\left[X_{(2)}\,;\ X^{(2)}\right]$ contienne m est-elle plus petite ou plus grande que $\Pr\left\{\left[X_{(1)}\,;\ X^{(1)}\right] \text{ contient } m\right\}$ *(Réponse : plus petite).*

T 2.5 ⇒Comment choisir l'intervalle, parmi tous les intervalles possibles :

$$\left[X_{(1)}\,;\ X^{(1)}\right], \left[X_{(2)}\,;\ X^{(2)}\right], \ldots, \left[X_{(c)}\,;\ X^{(c)}\right] \ ?$$

À chaque intervalle est associée une probabilité γ que cet intervalle contienne m.

On choisit γ *a priori*, ou une valeur minimale de γ, et on calcule l'intervalle tel que celui-ci ait un coefficient de confiance au moins égal à γ.

On utilise une table.

	min à max	$X_{(2)}$ à $X^{(2)}$	$X_{(3)}$ à $X^{(3)}$	$X_{(4)}$ à $X^{(4)}$	$X_{(5)}$ à $X^{(5)}$	$X_{(6)}$ à $X^{(6)}$	$X_{(7)}$ à $X^{(7)}$	$X_{(8)}$ à $X^{(8)}$	$X_{(9)}$ à $X^{(9)}$	$X_{(10)}$ à $X^{(10)}$
n	Coefficient de confiance									
2	0,500 0									
3	0,750 0									
4	0,875 0	0,375 0								
5	0,937 5	0,625 0								
6	0,968 8	0,781 3	0,312 5							
7	0,984 4	0,875 0	0,546 8							
8	0,992 2	0,929 7	0,711 0	0,273 4						
9	0,996 1	0,960 9	0,820 3	0,492 2						
10	0,998 0	0,978 5	0,890 6	0,656 3	0,246 1					
11	0,999 0	0,988 3	0,934 6	0,773 4	0,451 2					
12	0,999 5	0,993 7	0,961 4	0,854 0	0,612 3	0,225 6				
13	0,999 76	0,996 6	0,977 5	0,907 7	0,733 2	0,419 0				
14	0,999 88	0,998 2	0,987 1	0,942 6	0,820 4	0,576 1	0,209 5			
15	0,999 939	0,999 02	0,992 1	0,964 8	0,881 5	0,698 2	0,392 8			
16	0,999 969	0,999 48	0,995 8	0,978 7	0,923 2	0,789 9	0,545 5	0,196 4		
17	0,999 985	0,999 73	0,997 7	0,987 3	0,951 0	0,856 5	0,667 7	0,370 9		
18	0,999 992 3	0,999 86	0,998 7	0,992 5	0,969 1	0,903 7	0,762 1	0,519 3	0,185 5	
19	0,999 996 2	0,999 924	0,999 27	0,995 6	0,980 8	0,936 4	0,832 9	0,640 7	0,352 4	
20	0,999 998 1	0,999 960	0,999 60	0,997 4	0,988 2	0,968 6	0,884 7	0,738 5	0,496 5	0,176 2
c	1	2	3	4	5	6	7	8	9	10

TAB. 2.2 - *intervalle de confiance de la médiane avec les statistiques d'ordre* $\left[X_{(c)}\,;\ X^{(c)}\right]$

Distribuer et commenter la table 2.2.

On verra un peu plus loin comment établir cette table.

Utilisation de la table :

Avec $n = 9$, si on veut $\gamma \geq 0,95$, quel intervalle choisir? Quelle sera la valeur de γ?

Reporter les réponses sur le transparent T 2.4 : $\left[X_{(2)}\,;\ X^{(2)}\right]$, et $\gamma = 0,960\,9$.

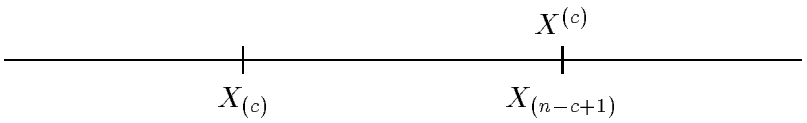
Si on choisit $\gamma \geq 0,99$ on obtient l'intervalle $\left[X_{(1)}\,;\ X^{(1)}\right]$ avec $\gamma = 0,996\,09$ (voir table 2.1).

Donc plus γ est élevé, plus l'intervalle est long. Si on prend $\left[X_{(1)}\,;\ X^{(1)}\right]$, on a un intervalle très grand, avec un niveau de confiance associé très élevé 0,99. Mais est-il bien nécessaire d'avoir un niveau de confiance aussi élevé?

Calcul de la table

T 2.6 ⇒En préliminaire, signaler qu'on peut laisser de côté les cas où l'une des observations coïncide avec la médiane m. Pour la suite de la démonstration, on s'appuie sur le schéma du transparent qu'on complète au fur et à mesure. Indiquer que Y est le nombre d'observations à gauche de m. Maintenant dire que $X_{(c)} \leq m$ signifie que $X_{(c)}$ est à gauche de m et par suite $X_{(c-1)}$, ..., $X_{(1)}$ aussi (les placer au fur et à mesure). Donc on a au moins c observations à gauche de m c'est-à-dire $Y \geq c$. De l'autre côté, dire que $X^{(c)} \geq m$ signifie que $X^{(c)}$ est à droite de m et par suite $X^{(c-1)}$, ...,$X^{(1)}$ aussi (les placer au fur et à mesure). Donc on a au moins c observations à droite de m c'est-à-dire au plus $n - c$ observations à gauche de m c'est-à-dire $Y \leq n - c$, d'où $c \leq Y \leq n - c$.

Schéma :



T 2.7 ⇒Loi de la variable aléatoire Y :

Montrer à l'aide du transparent en suivant le parallèle avec la loi binomiale que $Y \sim \mathcal{B}(n, \frac{1}{2})$. Donc on sait calculer $\Pr\{c \leq Y \leq n - c\}$: table de la loi binomiale, complètement connue.

T 2.8 ⇒Notation.

L'erreur, c'est-à-dire la probabilité que $\left[X_{(c)}\,;\ X^{(c)}\right]$ ne contienne pas m, est notée $\alpha = 1 - \gamma$ (remontrer α dans la table 2.1).

La loi binomiale est symétrique, donc :

$$\Pr\{m < X_{(c)}\} = \Pr\{m > X^{(c)}\} = \alpha' = \frac{\alpha}{2}$$

et on note l'erreur $\alpha = 2\alpha'$. Remarque pour les formateurs : α' est la notation pour le risque associé aux intervalles de confiance unilatéraux tandis que α est utilisé pour les intervalles bilatéraux.

T 2.9 ⇒La variable Y dépend à la fois des observations et du paramètre d'intérêt (médiane). C'est en cela qu'elle nous apporte une information sur le paramètre. On remarquera que Y n'est pas calculable à partir d'un échantillon : avec un n-échantillon, on ne peut pas calculer le nombre de valeurs plus petites que m, puisque m est inconnu. *Y n'est donc pas une statistique.*

Par contre, la loi de Y est connue, c'est une loi binomiale, on connaît sa distribution, les valeurs des probabilités ont été calculées et rassemblées dans une table, cette loi est totalement indépendante de la forme de la distribution de la population. C'est ainsi qu'on peut construire un IC à partir de Y. On dit que Y est une variable pivot.

T 2.10 ⇒Bilan. La méthode de construction d'IC qu'on va décrire est appelée méthode du signe : la variable pivot Y sur laquelle elle est basée ne dépend que du signe des $X_i - \widetilde{X}$.

Comment a-t-on défini un intervalle de confiance de la médiane m ? On a supprimé le même nombre de valeurs observées, $c-1$, aux deux extrémités de l'échantillon réordonné.

Que faut-il connaître pour choisir ce nombre de valeurs, $c-1$, à supprimer ?
(Attendre les réponses puis présenter le transparent suivant).

T 2.11 ⇒Tableau de γ en fonction de c et n

1- par exemple pour un 9-échantillon, (ligne horizontale sur transparent) si on augmente γ, que devient c ? *(Il diminue).*

Que devient l'intervalle ... ? *(Il devient plus large).*

Donc ce nombre c dépend de γ.

2- pour γ fixé, $\gamma = 0,96$, (valeurs à entourer sur le transparent) si on augmente n, que devient c ... ? *(Il augmente).*

Donc la valeur de c dépend de γ fixé *a priori* et de n, la taille de l'échantillon.

Remarquons que cet intervalle de confiance dépend également de la dispersion des données, car il est construit en éliminant « à chaque extrémité » un certain nombre de valeurs.

2.1.2 Cas des grands échantillons

T 2.12 ⇒La table 2.2 n'est pas calculée pour n grand, car quand n est grand, $\mathcal{B}(n,p)$ devient une gaussienne $\mathcal{N}(np, npq)$.

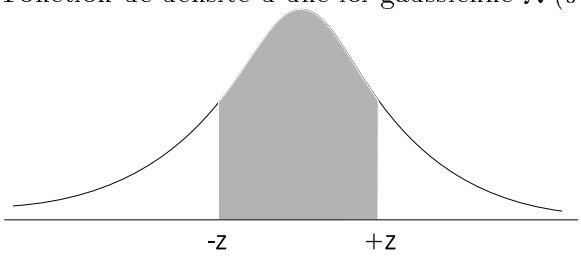
Donc $\mathcal{B}(n, \frac{1}{2}) \approx \mathcal{N}(\frac{n}{2}, \frac{n}{4})$.

Alors c devient :

$c = \frac{n+1}{2} - z\frac{\sqrt{n}}{2}$ avec z écart gaussien.

Rappel de la définition de l'écart gaussien :

Fonction de densité d'une loi gaussienne $\mathcal{N}(0,1)$:



On sait calculer la surface $\Pr\{-z \leq X \leq z\}$ sous la courbe comprise entre $-z$ et $+z$, donc connaissant $\gamma = \Pr\{X_{(c)} \leq m \leq X^{(c)}\}$ fixé *a priori*, on lit dans la table la valeur de z correspondant, d'où :

$$c = \frac{n+1}{2} - z\frac{\sqrt{n}}{2}$$

Démonstration (pour formateurs) :
Quand n est grand, la loi $\mathcal{B}\left(n, \frac{1}{2}\right)$ est proche de la loi gaussienne $\mathcal{N}\left(\frac{n}{2}, \frac{n}{4}\right)$. Soit $\widetilde{Z}$ une variable qui suit une $\mathcal{N}\left(\frac{n}{2}, \frac{n}{4}\right)$. L'approximation avec correction de continuité est basée sur les formules de substitution suivantes :

$$\begin{aligned}
\Pr\{Y=0\} &\leftarrow \Pr\left\{\widetilde{Z} \leq 1/2\right\} \\
\Pr\{Y=1\} &\leftarrow \Pr\left\{1/2 \leq \widetilde{Z} \leq 3/2\right\} \\
&\vdots \\
\Pr\{Y=i\} &\leftarrow \Pr\left\{i-1/2 \leq \widetilde{Z} \leq i+1/2\right\} \\
\Pr\{Y=n-1\} &\leftarrow \Pr\left\{n-3/2 \leq \widetilde{Z} \leq n-1/2\right\} \\
\Pr\{Y=n\} &\leftarrow \Pr\left\{n-1/2 \leq \widetilde{Z}\right\}.
\end{aligned}$$

On cherche c tel que

$$\Pr\{Y < c\} \leq \frac{\alpha}{2}.$$

Or

$$\Pr\{Y < c\} = \Pr\{Y=0\} + \Pr\{Y=1\} + \ldots + \Pr\{Y=c-1\}.$$

Soit en utilisant les approximations

$$\begin{aligned}
\Pr\{Y < c\} &\simeq \Pr\left\{\widetilde{Z} \leq 1/2\right\} + \Pr\left\{1/2 \leq \widetilde{Z} \leq 3/2\right\} + \ldots + \Pr\left\{c-3/2 \leq \widetilde{Z} \leq c-1/2\right\} \\
&\simeq \Pr\left\{\widetilde{Z} \leq c-1/2\right\}.
\end{aligned}$$

On est donc ramené à la recherche d'un c tel que

$$\Pr\left\{\widetilde{Z} \leq c-1/2\right\} \leq \frac{\alpha}{2}.$$

Si on centre et réduit, on obtient l'inégalité

$$\Pr\left\{Z \leq \frac{2}{\sqrt{n}}\left(c-\frac{1}{2}\right) - \sqrt{n}\right\} \leq \frac{\alpha}{2},$$

où Z est une gaussienne centrée réduite. Par définition de l'écart gaussien, on trouve donc

$$\frac{2}{\sqrt{n}}\left(c-\frac{1}{2}\right)-\sqrt{n}=-z,$$

c'est-à-dire

$$c=\frac{n+1}{2}-z\frac{\sqrt{n}}{2}.$$

2.1.3 Intervalle de confiance unilatéral

Jusqu'ici nous avons construit un intervalle de confiance bilatéral, c'est-à-dire que l'on recherchait une information à la fois sur les bornes inférieure et supérieure de la médiane. Nous allons voir maintenant, sur un même exemple de données, que suivant la question qu'on se pose, on peut vouloir calculer soit un intervalle de confiance bilatéral soit un intervalle de confiance unilatéral. Dans ce dernier cas, on ne recherche une information que sur une seule borne (inférieure ou supérieure) de la médiane.

T 2.13 ⇒De l'IC bilatéral à l'IC unilatéral, IC bilatéral.

Texte ⇒ On vient de créer une nouvelle race de vaches laitière et on veut estimer (par intervalle) sa production laitière médiane. Pour évaluer le rendement de la nouvelle race, on dispose des productions individuelles d'un échantillon de 50 bêtes choisies au hasard.

On va calculer un intervalle du type $[X_{(c)}, X^{(c)}]$. On veut un niveau de confiance a peu près égal à 95%.

On utilise la table 2.3 pour remplir les blancs sur le transparent. Pour $n = 50$, on trouve un γ immédiatement supérieur à $0,95$ égal à $0,967$. Le α correspondant est égal à $0,033$ et le c correspondant est 18. On a donc

$$\Pr\{[X_{(18)}, X^{(18)}] \ni m\} = 0,967.$$

On détermine la 18ième plus petite observation et la 18ième plus grande observation et on trouve comme IC

$$[X_{(18)}, X^{(18)}] = [4101, 4727].$$

T 2.14 ⇒De l'IC bilatéral à l'IC unilatéral, IC unilatéral

Texte ⇒ Il s'agit maintenant de voir si on peut recommander la nouvelle race à un agriculteur qui souhaite augmenter la production de son troupeau, actuellement composé de vaches de provenance locale. La production laitière médiane de la race locale est connue et égale à $R = 4190$ litre/an. On utilise les mêmes données que pour le problème précédent. On ne s'intéresse qu'à la borne inférieure de l'IC. En effet, on souhaite seulement savoir si la nouvelle race produit plus que la race locale (qu'elle produise autant ou moins, on ne la recommandera pas).

On va donc chercher un c' tel que

$$\Pr\{X_{(c')} < m\} = \gamma' \simeq 0,95.$$

n	c	γ	α	$\alpha/$
5	1	.938	.062	.031
6	1	.969	.031	.016
	2	.781	.219	.109
7	1	.984	.016	.008
	2	.875	.125	.062
8	1	.992	.008	.004
	2	.930	.070	.035
	3	.711	.289	.145
9	1	.996	.004	.002
	2	.961	.039	.020
	3	.820	.180	.090
10	1	.998	.002	.001
	2	.979	.021	.011
	3	.891	.109	.055
11	1	.999	.001	.000
	2	.988	.012	.006
	3	.935	.065	.033
	4	.773	.227	.113
12	2	.994	.006	.003
	3	.961	.039	.019
	4	.854	.146	.073
13	2	.997	.003	.002
	3	,978	.022	.011
	4	.908	.092	.046
	5	.733	.267	.133
14	2	.998	.002	.001
	3	.987	.013	.006
	4	.943	.057	.029
	5	.820	.180	.090
15	3	.993	.007	.004
	4	.965	.035	.018
	5	.882	.118	.059
16	3	.996	.004	.002
	4	.979	.021	.011
	5	.923	.077	.038
	6	.790	.210	.105
17	3	.998	.002	.001
	4	.987	.013	.006
	5	.951	.049	.025
	6	.857	.143	.072
18	4	.992	.008	.004
	5	.969	.031	.015
	6	.904	.096	.048
	7	.762	.238	.119
19	4	.996	.004	.002
	5	.981	.019	.010
	6	.936	.064	.032
	7	.833	.167	.084
20	4	.997	.003	.001
	5	.988	.012	.006
	6	.959	.041	.021
	7	.885	.115	.058
21	5	.993	.007	.004
	6	.973	.027	.013
	7	.922	.078	.039
	8	.811	.189	.095
22	5	.996	.004	.002
	6	.983	.017	.008
	7	.948	.052	.026
	8	.866	.134	.067
23	5	.997	.003	.001
	6	.989	.011	.005
	7	.965	.035	.017
	8	.907	.093	.047
	9	.790	.210	.105
24	6	.993	.007	.003
	7	.977	.023	.011
	8	.936	.064	.032
	9	.848	.152	.076
25	6	.996	.004	.002
	7	.985	.015	.007
	8	.957	.043	.022
	9	.892	.108	.054
26	7	.991	.009	.005
	8	.971	.029	.014
	9	.924	.076	.038
	10	.831	.169	.084
27	7	.994	.006	.003
	8	.981	.019	.010
	9	.948	.052	.026
	10	.878	.122	.061
28	7	.996	.004	.002
	8	.987	.013	.006
	9	.964	.036	.018
	10	.913	.087	.044
	11	.815	.185	.092
29	8	.992	.008	.004
	9	.976	.024	.012
	10	.939	.061	.031
	11	.864	.136	.068
30	8	.995	.005	.003
	9	.984	.016	.008
	10	.957	.043	.021
	11	.901	.099	.049
	12	.800	.200	.100
31	8	.997	.003	.002
	9	.989	.011	.005
	10	.971	.029	.015
	11	.929	.071	.035
	12	.850	.150	.075
32	9	.993	.007	.004
	10	.980	.020	.010
	11	.950	.050	.025
	12	.890	.110	.055
33	9	.995	.005	.002
	10	.986	.014	.007
	11	.965	.035	.018
	12	.920	.080	.040
	13	.837	.163	.081
34	10	.991	.009	.005
	11	.976	.024	.012
	12	.942	.058	.029
	13	.879	.121	.061
35	10	.994	.006	.003
	11	.983	.017	.008
	12	.959	.041	.020
	13	.910	.090	.045
	14	.825	.175	.088
36	10	.996	.004	.002
	11	.989	.011	.006
	12	.971	.029	.014
	13	.935	.065	.033
	14	.868	.132	.066
37	11	.992	.008	.004
	12	.980	.020	.010
	13	.953	.047	.024
	14	.901	.099	.049
	15	.812	.188	.094
38	11	.995	.005	.003
	12	.986	.014	.007
	13	.966	.034	.017
	14	.927	.073	.036
	15	.857	.143	.072
39	12	.991	.009	.005
	13	.976	.024	.012
	14	.947	.053	.027
	15	.892	.108	.054
40	12	.994	.006	.003
	13	.983	.017	.008
	14	.962	.038	.019
	15	.919	.081	.040
	16	.846	.154	.077
41	12	.996	.004	.002
	13	.988	.012	.006
	14	.972	.028	.014
	15	.940	.060	.030
	16	.883	.117	.059
42	13	.992	.008	.004
	14	.980	.020	.010
	15	.956	.044	.022
	16	.912	.088	.044
	17	.836	.164	.082
43	13	.995	.005	.003
	14	.986	.014	.007
	15	.968	.032	.016
	16	.934	.066	.033
	17	.874	.126	.063
44	14	.990	.010	.005
	15	.977	.023	.011
	16	.951	.049	.024
	17	.904	.096	.048
	18	.826	.174	.087
45	14	.993	.007	.003
	15	.984	.016	.008
	16	.964	.036	.018
	17	.928	.072	.036
	18	.865	.135	.068
46	14	.995	.005	.002
	15	.989	.011	.006
	16	.974	.026	.013
	17	.946	.054	.027
	18	.896	.104	.052
47	15	.992	.008	.004
	16	.981	.019	.009
	17	.960	.040	.020
	18	.921	.079	.039
	19	.856	.144	.072
48	15	.994	.006	.003
	16	.987	.013	.007
	17	.971	.029	.015
	18	.941	.059	.030
	19	.889	.111	.056
49	16	.991	.009	.005
	17	.979	.021	.011
	18	.956	.044	.022
	19	.915	.085	.043
	20	.848	.152	.076
50	16	.993	.007	.003
	17	.985	.015	.008
	18	.967	.033	.016
	19	.935	.065	.032
	20	.881	.119	.059

TAB. 2.3 - *Test du signe et intervalle de confiance de la médiane*

Ce qu'on peut réécrire

$$\Pr\{[X_{(c')}, +\infty[\ni m\} = \gamma' \simeq 0,95.$$

Pour trouver γ', on utilise toujours la table 2.3. Cette table donne $\alpha' = 1 - \gamma'$ plutôt que γ'. On cherche donc un α' immédiatement inférieur à $1 - 0,95 = 0,05$. Et on trouve $\alpha' = 0,032$. C'est-à-dire $\gamma' = 1 - 0,032 = 0,968$. Reporter cette valeur sur le transparent ainsi que le $c' = 19$ correspondant (colonne c dans la table). On cherche la 19ième plus petite observation et on trouve $X_{(19)} = 4155$ qu'on reporte sur le transparent.

L'IC calculé contient la médiane R de la production médiane de la race locale. L'étude menée n'a pas mis en évidence un accroissement de la production laitière.

On compare maintenant les deux IC obtenus. Faire remarquer que γ a presque la même valeur que γ', mais que les bornes inférieures des intervalles ne sont pas identiques.

Donc pour une même valeur du coefficient de confiance, les bornes inférieures des intervalles unilatéral et bilatéral sont différentes.

Conclusion : à niveau de confiance égal, avec l'intervalle unilatéral, on n'a pas d'information sur la borne supérieure (on n'en a pas besoin), mais on a pu augmenter la borne inférieure en augmentant la valeur de c. Autrement dit, le test unilatéral mettra plus facilement en évidence l'hypothèse unilatérale H_1 que le test bilatéral.

2.2 Intervalle de confiance de la médiane et de la moyenne (distribution symétrique)

T 2.15 ⇒⇒ Jusqu'à présent, nous avons travaillé sur des distributions des X_i quelconques et la méthode du signe nous a permis de construire des IC de la médiane. Nous allons maintenant progressivement introduire des postulats sur la distribution des X_i, qui vont nous permettre de réduire la taille de l'IC de la médiane (ou de la moyenne). D'autres méthodes seront mises en oeuvre pour construire ces IC. Ces postulats sur la distribution des X_i doivent être faits *a priori*. La méthode de Wilcoxon que nous allons voir maintenant (flèche sur le transparent) est basée sur le postulat que la distribution est symétrique.

2.2.1 Introduction : recherche d'un intervalle de confiance symétrique

T 2.16 ⇒Propriété : si la distribution de la variable X étudiée est symétrique, la médiane et la moyenne sont confondues.
Dans le cas d'une distribution symétrique, il apparaît naturel d'encadrer ce paramètre par un intervalle symétrique autour de son estimation. Dans un premier temps, on va rechercher un intervalle de confiance autour de la médiane, avec symétrie des *probabilités* de part et d'autre de la médiane. Puis, on envisagera d'encadrer la moyenne empirique de l'échantillon en supposant que les X_i sont des variables gaussiennes : on retiendra alors la même *longueur* de part et d'autre de la moyenne.

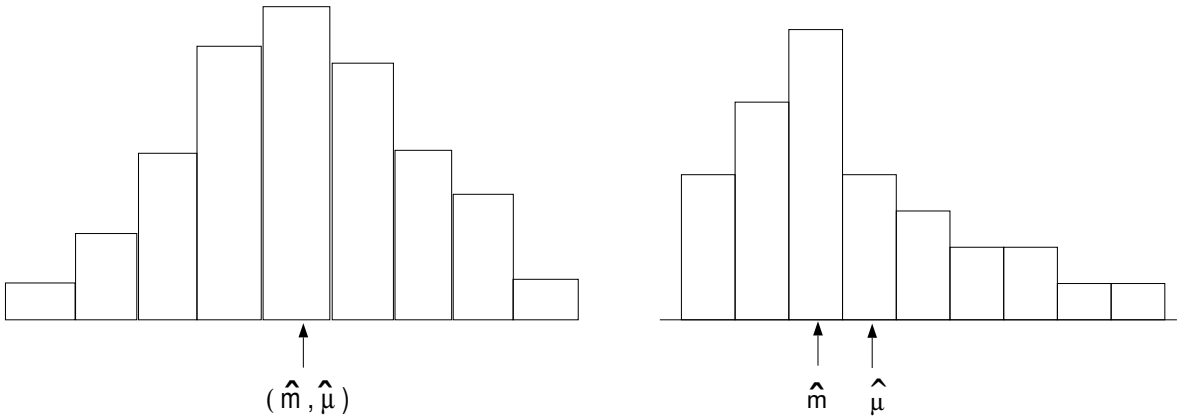


FIG. 2.1 - *Représentation des individus puis des variables.*

2.2.2 Intervalle de confiance de la médiane : méthode de Wilcoxon

Résumé ⇒

L'introduction de l'hypothèse de symétrie de la distribution des X_i permet de « *raccourcir* » l'intervalle de confiance de la médiane. Mais le modèle statistique utilisé est différent du précédent.

T 2.17 ⇒La longueur de l'intervalle de confiance pour la médiane obtenu par la méthode du signe dépend de la dispersion des observations. Montrer sur le transparent l'exemple de 2 échantillons de taille 7. L'intervalle de de confiance de l'échantillon le plus dispersé est plus grand.

L'idée à la base de la méthode de Wilcoxon est de construire l'intervalle de confiance non pas à partir des observations initiales X_i mais de leurs moyennes 2 à 2 Z_{ij}. Montrer sur le 2nd graphique du transparent les Z_{ij}. Les segments verticaux en pointillé indiquent la correspondance entre les (X_i, X_i) et les Z_{ii}, les segments obliques en pointillé indiquent pour trois couples (X_i, X_j) les correspondances avec les Z_{ij}. La réduction de la dispersion est indiquée par la réduction de l'intervalle interquartile.

NB : *De plus si la distribution des X_i est symétrique, les Z_{ij} ont même médiane que les X_i.* C'est-à-dire que ce postulat de symétrie assure que les observations initiales et leurs moyennes 2 à 2 sont bien concentrées autour de la même valeur.

T 2.18 ⇒Construction de l'IC pour la médiane par la méthode de Wilcoxon.

Cette méthode est basée sur un postulat de symétrie de la distribution des observations. Elle utilise les moyennes 2 à 2 des observations qui ont la même médiane que les observations initiales mais sont moins dispersées.

L'intervalle de confiance (bilatéral) est obtenu en éliminant un même nombre c de Z_{ij} extrêmes. Pour calculer le coefficient de confiance γ d'un IC donné ou déterminer l'IC pour un coefficient de confiance γ donné, on peut utiliser la même variable pivot que pour la méthode du signe, c'est-à-dire le nombre de Z_{ij} plus petits que la médiane. Mais *attention*,

les Z_{ij} ne sont pas indépendants et Y_Z ne suit pas une loi binômiale. Il faut donc utiliser une table spécifique (table 2.4) pour calculer le coefficient de confiance ou déterminer l'IC.

T 2.19 ⇒Commenter la table et retrouver ensemble la valeur de $c = 5$ pour $\gamma = 0,89$. Représenter l'intervalle de confiance obtenu sur le graphique du transparent. Comparer à la table précédente utilisée pour la méthode du signe où $c = 2$ pour $\gamma = 0,87$ (représenter aussi cet intervalle sur le transparent).

On insistera sur le fait que la réduction de l'intervalle est associé à une hypothèse de symétrie de la distribution de départ. Le modèle statistique utilisé, dans ce cas, est différent de celui utilisé pour la méthode du signe.

Exercice :

Sous l'hypothèse d'une distribution symétrique de la variable «âge» dans la population de l'INRA, faire calculer l'intervalle de confiance de la médiane de chacun des 9-échantillons travaillés par les stagiaires, avec un coefficient de confiance de 0,961.

Pour $n = 9$ et $\gamma = 0,961$, on obtient $c = 6$, alors que l'on avait $c = 2$ dans le cas précédent.

L'exemple sera traité sur un 9-échantillon préparé par le formateur (les Z_{ij} étant calculées auparavant), en montrant qu'il n'est pas utile de calculer toutes les valeurs des Z_{ij} pour conclure : seules les valeurs extrêmes sont intéressantes !

2.2.3 Intervalle de confiance de la moyenne : méthode de Student

T 2.20 ⇒L'introduction d'un postulat encore plus fort que le précédent (normalité des X_i, qui implique aussi la symétrie de la distribution) conduit à restreindre encore l'intervalle de confiance de la moyenne. Pour un échantillon de taille donnée et un niveau de confiance fixé, plus une distribution sera bien caractérisée *a priori*, meilleure sera la précision que l'on aura sur sa médiane (ou sa moyenne).

T 2.21 ⇒Nous allons chercher à construire un IC autour de la moyenne, en travaillant non plus sur la distribution des X_i mais sur la distribution des $\overline{X}$ (qui ont la même moyenne). Montrer visuellement qu'un IC construit sur $\overline{X}$ sera plus petit qu'un IC construit sur les X_i (cas d'une distribution gaussienne).

T 2.22 ⇒On va déterminer une marge autour de la moyenne empirique $\overline{X}$ des variables X_i, telle que :

$$\overline{X} - \text{ marge } \leq \mu \leq \overline{X} + \text{ marge}$$

La marge précédente sera définie en fonction de : *(Laisser dire les stagiaires !)...*

- la taille n de l'échantillon,

n	c	γ	α	α'	n	c	γ	α	α'	n	c	γ	α	α'	n	c	γ	α	α'
3	1	.750	.250	.125	11	6	.990	.010	.005	16	20	.991	.009	.005	21	43	.991	.009	.005
4	1	.875	.125	.062		7	.986	.014	.007		21	.989	.011	.006		44	.990	.010	.005
5	1	.938	.062	.031		11	.958	.042	.021		30	.956	.044	.022		59	.954	.046	.023
	2	.875	.125	.062		12	.946	.054	.027		31	.949	.051	.025		60	.950	.050	.025
6	1	.969	.031	.016		14	.917	.083	.042		36	.907	.093	.047		68	.904	.096	.048
	2	.937	.063	.031		15	.898	.102	.051		37	.895	.105	.052		69	.897	.103	.052
	3	.906	.094	.047	12	8	.991	.009	.005	17	24	.991	.009	.005	22	49	.991	.009	.005
	4	.844	.156	.078		9	.988	.012	.006		25	.989	.011	.006		50	.990	.010	.005
7	1	.984	.016	.008		14	.958	.042	.021		35	.955	.045	.022		66	.954	.046	.023
	3	.953	.047	.023		15	.948	.052	.026		36	.949	.051	.025		67	.950	.050	.025
	4	.922	.078	.039		18	.908	.092	.046		42	.902	.098	.049		76	.902	.098	.049
	5	.891	.109	.055		19	.890	.110	.055		43	.891	.109	.054		77	.895	.105	.053
8	1	.992	.008	.004	13	10	.992	.008	.004	18	28	.991	.009	.005	23	55	.991	.009	.005
	2	.984	.016	.008		11	.990	.010	.005		29	.990	.010	.005		56	.990	.010	.005
	4	.961	.039	.020		18	.952	.048	.024		41	.952	.048	.024		74	.952	.048	.024
	5	.945	.055	.027		19	.943	.057	.029		42	.946	.054	.027		75	.948	.052	.026
	6	.922	.078	.039		22	.906	.094	.047		48	.901	.099	.049		84	.902	.098	.049
	7	.891	.109	.055		23	.890	.110	.055		49	.892	.108	.054		85	.895	.105	.052
9	2	.992	.008	.004	14	13	.991	.009	.004	19	33	.991	.009	.005	24	62	.990	.010	.005
	3	.988	.012	.006		14	.989	.011	.005		34	.989	.011	.005		63	.989	.011	.005
	6	.961	.039	.020		22	.951	.049	.025		47	.951	.049	.025		82	.951	.049	.025
	7	.945	.055	.027		23	.942	.058	.029		48	.945	.055	.027		83	.947	.053	.026
	9	.902	.098	.049		26	.909	.091	.045		54	.904	.096	.048		92	.905	.095	.048
	10	.871	.129	.065		72	.896	.104	.052		55	.896	.104	.052		93	.899	.101	.051
10	4	.990	.010	.005	15	16	.992	.008	.004	20	38	.991	.009	.005	25	69	.990	.010	.005
	5	.986	.014	.007		17	.990	.010	.005		39	.989	.011	.005		70	.989	.011	.005
	9	.951	.049	.024		26	.952	.048	.024		53	.952	.048	.024		90	.952	.048	.024
	10	.936	.064	.032		27	.945	.055	.028		54	.947	.053	.027		91	.948	.052	.026
	11	.916	.084	.042		31	.905	.095	.047		61	.903	.097	.049		101	.904	.096	.048
	12	.895	.105	.053		32	.893	.107	.054		62	.895	.105	.053		102	.899	.101	.051

$\alpha = 2\alpha' = 1-\gamma$ = niveau de signification bilatéral, $\alpha' = \alpha/2 = (1-\gamma)/2$ = niveau de signification unilatéral,
γ = coefficient de confiance. Si $n > 25$, $c \approx n(n+1)/4 + 1/2 - z\sqrt{n(n+1)(2n+1)/24}$, ($z$: écart gaussien).

TAB. 2.4 - *test de Wilcoxon signes et rangs et intervalle de confiance de la moyenne*

- la dispersion $\widehat{\sigma}^2$ des données,
- et du coefficient de confiance γ choisi.

Rappel

Soient n variables X_i.

La moyenne $(\overline{X})$ et la variance empiriques $(\widehat{\sigma}^2)$ de l'échantillon des X_i sont :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Si les variables X_i sont indépendantes, les variables $(X_i - \overline{X})$ ne le sont pas : l'une d'entre elles est complètement déterminée par la connaissance des autres ! Il ne reste donc plus que $(n-1)$ variables indépendantes dans l'expression de l'estimateur $\widehat{\sigma}^2$; ceci conduit à diviser la somme des carrés des écarts à la moyenne par $(n-1)$ et non par n.

T 2.23 $\Rightarrow$

- Si les $X_1, X_2, \ldots, X_n$ sont des gaussiennes $\mathcal{N}(\mu, \sigma^2)$, alors :

  $\overline{X}$ est une gaussienne $\mathcal{N}(\mu, \sigma^2/n)$

  $(\overline{X} - \mu)/(\sigma/\sqrt{n})$ est une gaussienne standard $\mathcal{N}(0,1)$ centrée et réduite.

- Si z est l'écart gaussien associé au risque $\alpha = 2\alpha'$, alors :

$$\Pr\{-z \leq (\overline{X} - \mu)/(\sigma/\sqrt{n}) \leq +z\} = 1 - \alpha = 1 - 2\alpha'$$

ce qui peut encore s'écrire :

$$\Pr\{-z(\sigma/\sqrt{n}) \leq (\overline{X} - \mu) \leq z(\sigma/\sqrt{n})\} = \Pr\{\overline{X} - z(\sigma/\sqrt{n}) \leq \mu \leq \overline{X} + z(\sigma/\sqrt{n})\}$$

En réalité, σ^2 n'est pas connu *a priori* et on l'estime à partir des données expérimentales :

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

T 2.24 $\Rightarrow$

– Cas des petits échantillons

La variable $T = (X - \mu)/(\hat{\sigma}/\sqrt{n})$ suit une distribution standard, nommée « distribution de Student à $(n-1)$ degrés de liberté » ; cette distribution est un peu plus dispersée que la distribution $\mathcal{N}(0,1)$. Elle sera notée t_{n-1}. On remplace l'écart gaussien Z par l'écart de Student t :

$$Pr\ \{\ \overline{X} - t\ \frac{\hat{\sigma}}{\sqrt{n}}\ \leq\ \mu\ \leq\ \overline{X}\ +\ t\ \frac{\hat{\sigma}}{\sqrt{n}}\}\ = 1 - \alpha$$

Dans les cas des petits échantillons, on doit postuler que les X_i suivant une loi gaussienne.

– Cas des grands échantillons

Dans le cas des grands échantillons, on utilise directement l'écart gaussien Z, pour encadrer la moyenne :

$$Pr\ \{\ \overline{X} - Z\ (\frac{\hat{\sigma}}{\sqrt{n}})\ \leq\ \mu\ \leq\ \overline{X} + Z\ (\frac{\hat{\sigma}}{\sqrt{n}})\ \}\ = 1 - \alpha$$

Dans le cas des grands échantillons, si la distribution des X_i est symétrique, la loi de $\overline{X}$ devient très vite proche de la loi gaussienne même si X_i ne suit pas une loi gaussienne. Si n est grand, il n'est donc pas nécessaire de postuler que la distribution des X_i est gaussienne, pour pouvoir encadrer la moyenne par cette méthode.

Exercice :

Sous les hypothèses d'une distribution (symétrique et) gaussienne de la variable « âge » dans la population de l'INRA, faire calculer l'intervalle de confiance de la médiane (ou de la moyenne, puisque ce sont les mêmes) sur un 9-échantillon préparé à l'avance par le formateur, avec un coefficient de confiance de 0,961.

Pour $\gamma = 0,961$ et $n = 9$, la table de Student donne $t_{n-1} = 2,5$, pour un intervalle bilatéral. On remarquera que l'intervalle obtenu est encore plus court que dans le cas précédent où seule l'hypothèse de symétrie avait été introduite. Le modèle utilisé est différent des précédents puisqu'il suppose, outre la symétrie de la distribution, la normalité des X_i. On insistera sur le fait que, si le modèle gaussien est effectivement correct, alors l'intervalle de confiance calculé est le meilleur.

Remarque ⇒ On pourra ici remontrer la distribution de la variable « âge d'un agent » dans l'ensemble de la population INRA et discuter les points suivants :

– Le modèle choisi était-il bon ? (les hypothèses de normalité et de symétrie étaient-elles justifiées ?) Il semble *a priori* que oui ; attention, quand même, aux queues de la

distribution ! Dans le cas où ces hypothèses n'auraient pas été respectées, le coefficient de confiance choisi au départ n'est qu'illusion, car son calcul repose sur des bases fausses !

- Si la variance de la population avait été parfaitement connue au départ (et non estimée à partir des échantillons de chacun), comment aurait été calculé l'intervalle de confiance autour de la moyenne? *(à partir d'une distribution normale et non de Student !)*

2.3 Synthèse sur les intervalles de confiance

T 2.25 ⇒Dans un premier temps, on est conduit à identifier :

- une population (ajouter sur le transparent «une nouvelle race de vaches»),
- une variable d'étude (ajouter sur le transparent «le rendement laitier»),
- un paramètre de la loi de cette variable (ajouter sur le transparent «l'espérance ou la médiane»).

T 2.26 ⇒Dans un second temps, on est amené à calculer sur un n-échantillon donné, pour un coefficient de confiance γ choisi :

- soit directement un intervalle de confiance (distribution quelconque, ou petits échantillons et distribution postulée symétrique) ; dans ce cas, on a effectivement défini l'intervalle avec un nombre de valeurs identiques de part et d'autre de la médiane ($[(n+1)/2]$-ième valeur centrale si n impair) ;
- soit un estimateur du paramètre identifié et une marge autour de cet estimateur (grands échantillons et distribution postulée symétrique, ou distribution postulée gaussienne) : dans ce cas (postulat de normalité de $\overline{X}$), on a défini un intervalle en considérant une marge identique $z\widehat{\sigma}/\sqrt{n}$ de part et d'autre de la moyenne empirique $\overline{X}$.

Dans tous les cas, on aboutit à un intervalle de confiance aléatoire autour du paramètre identifié.

Ceci correspond au fait qu'en tirant un grand nombre d'échantillons de la population, une proportion γ contiendra la valeur vraie du paramètre et une proportion $(1-\gamma)$ ne contiendra pas cette valeur.

Chapitre 3

Les tests d'hypothèses

T 3.1 ⇒Les test d'hypothèses

Texte ⇒

- La question qui se pose
- 2 façons d'y répondre

La duexième démarche donne les mêmes tests que la première, c'est simplement une autre façon de voir.

T 3.2 ⇒Plan proposé

Texte ⇒ Dans la partie «Démarche et vocabulaire», nous verrons ce que sont les erreurs de 1ère et seconde espèce, les risques correspondants, la puissance d'un test.

3.1 Introduction des principales notions sur l'exemple du test du signe

3.1.1 Du test du signe à l'erreur de décision

T 3.3 ⇒Test du signe

Texte ⇒ Un agent INRA effectue une mission au centre de recherches d'Antibes. Pendant le déjeuner, il s'étonne de l' âge apparemment élevé des agents d'Antibes.

Pour préciser cette impression, il veut comparer l'âge des agents d'Antibes à celui de l'ensemble des agents de l'INRA, il effectue un tirage aléatoire dans la liste du personnel, et note les âges de 13 personnes :

48 41 33 60 58 27 48 40 43 32 57 51 45

Soit, si on réordonne :

27 32 33 40 43 44 45 48 48 51 57 58 60

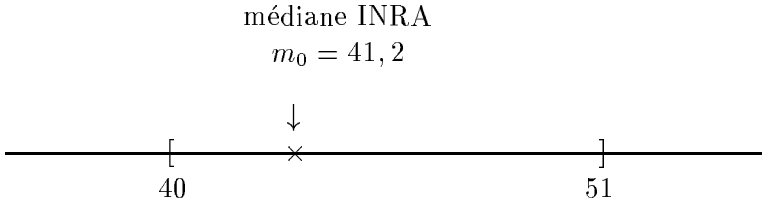
T 3.4 ⇒I.C. de la médiane m_A

Texte ⇒ Il calcule un I.C. de la médiane.

Il choisit un coefficient de confiance $\gamma = 0,90$, et obtient $c = 4$ (table 2.3).

Il en déduit un I.C. de la médiane des âges d'Antibes [40 ; 51].

La médiane des âges des agents INRA vaut $m_0 = 41,2$.



$41,2 \in$ I.C. [40 ; 51]

L'intervalle de confiance autour de m_A contient m_0. Il n'a donc pas pu démontrer que les agents d'Antibes sont plus âgés que dans le reste de l'INRA.

Il vaut mieux dire que l'I.C. contient m_0, plutôt que m_0 appartient à l'I.C. : la deuxième formulation suggère une variation de m_0, alors que m_0 est connu et fixé, et que c'est l'I.C. autour de m_A qui est aléatoire.

Qu'avons-nous fait ?

On sait que la médiane de l'âge des agents INRA, m_0, vaut 41,2.

- *Question* : « la médiane des âges d'Antibes vaut-elle m_0 ? », c'est-à-dire « $m_A = m_0$? »

  m_0 : valeur de référence (fixée, connue).

  Population : les agents d'Antibes.

  Variable aléatoire (V. A.) : âge.

  On a tiré un 13-échantillon (les tirages sont indépendants et tous les agents d'Antibes ont la même probabilité d'être tirés au sort).

- On a *construit un I.C.* autour de m_A tel que $\Pr\left\{\left[X_{(4)}\,;\ X^{(4)}\right] \ni m_A\right\} = 0,908$,

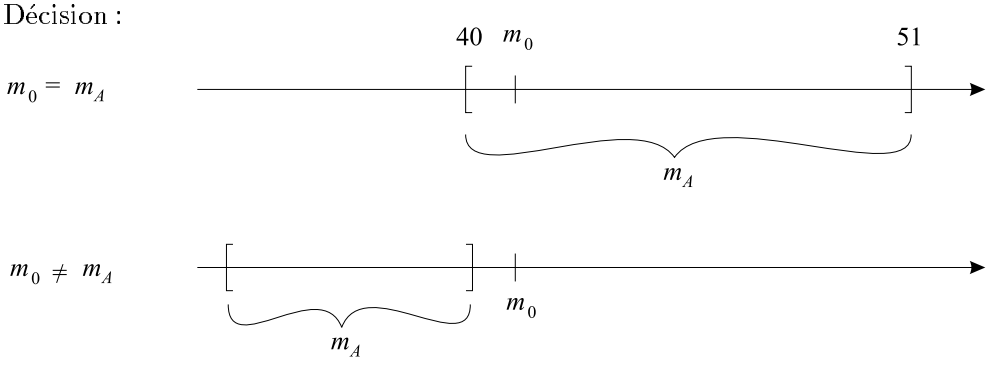
  c'est-à-dire que la probabilité pour que l'I.C. $\left[X_{(4)}\,;\ X^{(4)}\right]$ contienne m_A (inconnu) vaut 0,908. Il y a 91 % de chances que l'I.C. $\left[X_{(4)}\,;\ X^{(4)}\right]$ contienne m_A.

– *Règle de décision* :

si l'I.C. $\left[X_{(4)}\,;\ X^{(4)}\right]$ (autour de m_A) contient m_0 ..... je décide $m_A = m_0$.

si l'I.C. $\left[X_{(4)}\,;\ X^{(4)}\right]$ ne contient pas m_0 .....je décide $m_A \neq m_0$.

Décision :



Remarque ⇒ Bien aligner m_0 qui est connu, fixé. C'est l'I.C. qui varie.

En option, on peut demander aux stagiaires quelle est la population étudiée. Il est fort possible que certains répondent «les agents INRA-France» alors que c'est «les agents INRA-Antibes» qu'il faut répondre.

T 3.5 ⇒

Texte ⇒ La décision est-elle correcte? Les risques d'erreur.

Question : la règle de décision est donnée. Quelles sont les erreurs possibles ?

Réponse ⇒

Décision		*Réalité*		
		$m_A = m_O$		
$m_A = m_O$	↗			
	↘	$m_A \neq m_O$	$\longrightarrow$	erreur de décision
		$m_A \neq m_O$		
$m_A \neq m_O$	↗			
	↘	$m_A = m_O$	$\longrightarrow$	erreur de décision

3.1.2 Du test bilatéral au test unilatéral

T 3.6 ⇒Test bilatéral, test unilatéral

Texte ⇒ **Rappel**

H_0 : « $m_A = m_0$ »
H_1 : « $m_A \neq m_0$ » $\Longleftrightarrow \underbrace{<<\ m_A < m_0 \text{ ou } m_A > m_0\ >>}_{\textit{alternative bilatérale}}$

On a fait un test *bilatéral* sans le dire.

T 3.7 ⇒Test unilatéral

Texte ⇒ **Reprenons le problème**:

On a un a priori : l'âge médian des agents d'Antibes est plus élevé que l'âge médian national. Il semble que $m_A > m_0$.

On peut raisonnablement n'envisager que les deux possibilités : ou $m_A = m_0$, ou $m_A > m_0$. On exclut d'office $m_A < m_0$.

On pose :

H_0 : « $m_A = m_0$ »
H_1 : « $m_A > m_0$ » $\Longrightarrow$ *alternative unilatérale*.

On fait un *test unilatéral*.

Remarque ⇒ Ce qui nous a conduit à *choisir un test unilatéral*, c'est un *raisonnement* que l'on a fait, *avant* même *de recueillir les données et de les étudier*. C'est important. Ce raisonnement, dans notre cas, pourrait être: Antibes, c'est dans le Sud, on arrive a y être muté que tardivement l'âge est plus élevé qu'ailleurs. S'ils ne sont pas plus âgés, alors ils ont le même âge que dans n'importe quel autre centre INRA, mais on exclut le fait qu'ils puissent être plus jeunes.

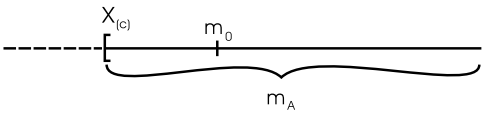
H_0 : « $m_A = m_0$ »
H_1 : « $m_A > m_0$ »

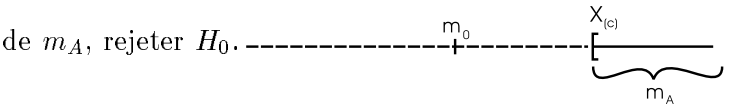
On veut savoir si m_0 est dans l'I.C., ou en dessous.

L'I.C. recherché est donc de la forme $\left[X_{(c)}\,;\ +\infty\right[$.
C'est un *I.C. unilatéral*.

La *règle de décision*, c'est :

Si m_0 est dans l'I.C. de m_A, accepter H_0.



Si m_0 est plus petit que l'I.C. de m_A, rejeter H_0. 

3.1.3 Test du signe, méthode directe

T 3.8 ⇒Test du signe, méthode directe

Texte ⇒ Étudions la loi (distribution) de la statistique Y_0 : nombre de valeurs X_i plus petites que m_0 (c'est une statistique, puisque c'est calculable à partir d'un échantillon : on connaît m_0, on peut donc compter le nombre de valeurs de l'échantillon plus petites que m_0).

Si l'hypothèse $m_A = m_0$ est vraie, m_0 est égal à la médiane de la population d'Antibes, et le nombre de valeurs X_i plus petites que m_0 est compris entre 0 et 13, et voisin de 13/2 (np). La loi de Y_0 est... (Laisser dire) *: une binomiale $\mathcal{B}(13, 1/2)$.*

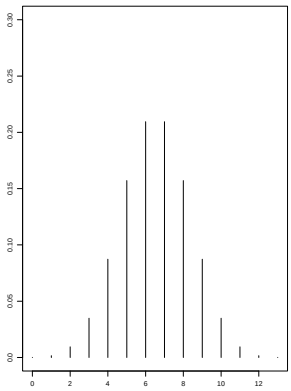


FIG. 3.1 *: loi de Y_0 si $m_A = m_0$ ($\mathcal{B}(13, 1/2)$).*

Si $m_A > m_0$, le nombre de valeurs X_i inférieures à m_0 a tendance à être plus petit, et Y_0 a tendance à être plus petit que 13/2.

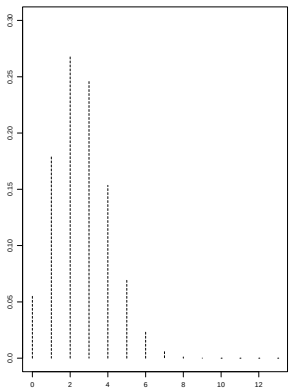


FIG. 3.2 : loi de Y_0 si : $m_A > m_0$.

Remarque ⇒ La loi de Y_0 si $m_A > m_0$ est encore une binomiale, de paramètre $p < 1/2$.

3.1.4 Démarche et vocabulaire

T 3.9 ⇒Démarche et vocabulaire

Texte ⇒ On récapitule en essayant de faire comprendre le principe général de la règle de décision.

T 3.10 ⇒

Texte ⇒ Tableau résumant la situation *(à construire au tableau ou sur transparent).*

H_0 : « $m_A = m_0$ »
H_1 : « $m_A > m_0$ »

TABLEAU 3.1 : les quatre issues possibles d'un test statistique

		la vraie valeur de m_A, inconnue, est $= m_0$ (H_0)	$m_0(H_1)$
décision	$m_A = m_0$ acceptation de H_0	pas d'erreur	erreur de 2^{e} espèce
	$m_A > m_0$ rejet de H_0	erreur de 1re espèce	pas d'erreur

Vocabulaire

(Présenter le vocabulaire par oral, en commentant le tableau, puis écrire au tableau les différents termes,avec une définition la plus succincte possible) .

- Deux « états possibles de la nature » :

  - *l'hypothèse nulle* H_0 : c'est le cas où il n'y a pas d'effet, pas de différence, rien de nouveau... (ici, $m_A = m_0$) ;

  - *l'hypothèse alternative* H_1 : c'est le cas « intéressant », dont on aimerait apporter la « preuve statistique », c'est-à-dire en contrôlant le risque d'erreur lorsqu'on le choisit comme conclusion (ici, $m_A > m_0$).

- Un *test statistique*, c'est une *règle de décision.*

- Lorsque l'on fait un test, on ne sait pas si c'est H_0 ou H_1 qui est vraie :

  - si c'est H_0 qui est vraie (s'il n'y a pas de différence ...), et si l'on conclut H_1 (il y a une différence), on fait une erreur, appelée *erreur de 1re espèce*,

  $\Longleftrightarrow$ si on conclut (à tort) qu'il y a une différence (H_1), on fait une erreur de 1re espèce.

  La probabilité, sous H_0, de l'erreur de 1re espèce est encore appelée *niveau de signification*, ou *niveau de significativité* (*level*, *significance level*).

- si c'est H_1 qui est vraie (il y a une différence), et que l'on conclut H_0, on fait une erreur appelée *erreur de 2^e espèce*,

$\Longleftrightarrow$ si on accepte à tort H_0 (pas de différence), on fait une erreur de 2$^{\text{e}}$ espèce.

Exercice 2

Un suspect est accusé de meurtre. On ignore s'il est coupable ou innocent. Le jury vote, après audition du procès, et décide coupable ou innocent : il prend une décision, et choisit entre une hypothèse et son alternative.

Question :

Quelle est l'hypothèse H_0 ?
Quelle est l'alternative H_1 ?
Précisez les erreurs possibles du jury. (Utilisez le vocabulaire statistique).

T 3.11 ⇒Exercice

Texte ⇒ Réponse
H_0 : « le suspect est innocent ».
H_1 : « le suspect est coupable ».

		réalité	
		innocent H_0	coupable H_1
décision	relaxé (H_0)	OK	erreur de 2$^{\text{e}}$ espèce
	condamné (H_1)	erreur de 1$^{\text{re}}$ espèce	OK

Le jury fait une erreur :
soit en décidant « coupable », alors que le suspect est innocent,
soit en décidant « innocent », alors que le suspect est coupable.

L'erreur de 1$^{\text{re}}$ espèce consiste à décider coupable à tort, c.-à-d. à rejeter H_0 à tort. L'erreur de 2$^{\text{e}}$ espèce consiste à décider non coupable à tort, c'est-à-dire à accepter H_0 à tort.

L'erreur de 1$^{\text{re}}$ espèce consiste à condamner un innocent, l'erreur de 2$^{\text{e}}$ espèce à relaxer un coupable.

Bilan

On a déterminé une règle de décision, pour faire un test statistique. On a vu qu'il existe deux types d'erreur de décision.

T 3.12 ⇒Comment préciser la valeur d'un test?

Texte ⇒ **Vocabulaire et notations**
Pr{faire une erreur de 1re espèce} : *risque de 1re espèce*, noté α.

Pr{faire une erreur de 2e espèce} : *risque de 2e espèce*, noté β.

Puissance : 1 - β

À partir des quatre issues possibles d'un test statistique, on construit le tableau suivant :

TABLEAU 3.2

		état de la nature	
		$H_0 \ : \ m_A = m_0$	$H_1 \ : \ m_A \neq m_0$
décision	$H_0 \ : \ m_A = m_0$	γ	$\Pr_{m_A}\left\{\left[X_{(4)} \ ; \ X^{(4)}\right] \ni m_0\right\}$ risque de 2e espèce = ? β ?
	$H_1 \ : \ m_A \neq m_0$ $[C; C'] \not\ni m_0$	α risque de 1re espèce	$1 - \beta$?

Attention

Après avoir fait un test, ne pas dire : « je décide H_1 avec une probabilité d'erreur égale à 9 % ». Lorsque le test est fait, la décision est prise, elle est correcte ou incorrecte ; elle n'est pas « correcte avec une probabilité égale à tant ».

Avant de faire le test, on peut dire que le risque d'erreur, dans le cas où on déciderait H_1, vaut 9 %.

Commentaire

Lorsque l'on fait un test, on se fixe α (risque de 1re espèce), et donc γ ($= 1 - \alpha$). Par conséquent, lorsque l'état de la nature est H_0 (ce que l'on ne sait pas), on connaît parfaitement le risque d'erreur α.

Cependant, lorsque l'on fait une expérience et un test, on ne connaît pas l'état de la nature, on ne sait pas dans quelle colonne on se trouve, on sait seulement dans quelle ligne on se trouve (suivant la décision prise).

Si l'on décide H_1, le risque que l'on court, c'est que l'état de la nature soit H_0. Ce risque est connu, c'est α. Soit on répond juste, soit on fait partie des $\alpha = 9,2\,\%$ de malchanceux qui se

trompent. Le *risque* α étant parfaitement *connu*, *contrôlé*, la situation est *confortable*.

Si l'on décide H_0, le risque que l'on court, c'est que l'état de la nature soit H_1. Ce risque, c'est le *risque de* 2^e *espèce*, β. Il est *inconnu*. La probabilité de se tromper (probabilité d'accepter l'hypothèse nulle alors qu'elle est fausse) n'est *pas contrôlée*.

Il faut être prudent...

Si je décide H_1, je connais la probabilité de me tromper.

Si je décide H_0, je ne connais pas la probabilité de me tromper. Je dirai donc que *je n'ai pas pu prouver que la différence observée est significative* (je n'ai pas réussi à rejeter H_0) ; ça ne veut pas dire que cette différence n'existe pas (peut-être que si j'avais tiré un échantillon de taille double ou triple, ou plus, j'aurais pu prouver que la différence est significative).

Remarque $\Rightarrow$ Attention à l'emploi du terme « significatif » : on observe une différence, et on se demande si cette différence observée sur les moyennes ou médianes empiriques reflète une réelle différence des espérances ou médianes théoriques. La différence observée est-elle significative? Si je rejette H_0, j'en déduirai que la différence observée est significative (d'une réelle différence au niveau des populations). Par contre, éviter de dire que « le test est significatif » (ça n'a pas tellement de sens).

Puissance d'un test

On appelle $1-\beta$ la *puissance* du test. De manières équivalentes, c'est :

$\Longleftrightarrow$ la probabilité de décider H_1, lorsque l'état de la nature est H_1.

$\Longleftrightarrow$ la probabilité de prouver qu'une différence observée est significative (lorsqu'elle l'est effectivement).

$\Longleftrightarrow$ la capacité d'un test à prouver qu'une différence observée est significative.

Remarque $\Rightarrow$ Pour certains tests, on sait quand même obtenir de l'information sur β, ou sur $1-\beta$: par exemple, un test unilatéral est plus puissant qu'un test bilatéral (dans le seul cas où ils sont comparables, c.-à-d. pour tester contre une alternative unilatérale) ; un test de Wilcoxon (qui postule la symétrie de la distribution) est plus puissant qu'un test du signe (qui ne suppose rien sur la distribution) lorsque les deux sont applicables (et donc si la distribution est réellement symétrique) ; pour une alternative du type « la médiane vaut m_1 » ou « l'espérance vaut μ_1 » où m_1, μ_1 sont connus, on peut (souvent) calculer la puissance.

T 3.13 $\Rightarrow$Puissance d'un test

Texte $\Rightarrow$ Effet de $m_A - m_0$ sur la puissance.

La distribution de Y_0 sous H_0 est connue. C'est une loi binomiale de paramètres n (taille de l'échantillon) et $p = 1/2$ $\mathcal{B}(n, 1/2)$.

L'hypothèse H_1 n'est pas connue avec précision. Regardons cependant deux cas possibles avec une différence $m_A - m_0$ plus ou moins grande (positive avec test unilatéral).

Sous H_1 Y_0 suit $\mathcal{B}(n,p)$ avec $p < 1/2$ (montrer sur la figure à quoi correspondent $\alpha, \beta, 1-\beta$).

Plus la différence $m_A - m_0$ est grande, plus la probabilité p d'observer une valeur dans l'échantillon inférieure à m_0 est faible, et plus la distribution du Y_0 est décaléee vers la gauche. Cela fait augmenter la surface correspondant à $1-\beta$ (montrer sur la figure).

T 3.14 ⇒Effet de la taille de l'échantillon sur la puissance.

Texte ⇒ En augmentant la taille de l'échantillon, les binomiales de paramètre p constant ont une distribution plus distincte de la $\mathcal{B}(n, 1/2)$ et donc la puissance du test augmente.

T 3.15 ⇒

T 3.16 ⇒

3.2 Les différents tests

T 3.17 ⇒Les différents tests

Texte ⇒
Remarque ⇒ plus il y a de postulats (plus l'information initiale est riche) plus le test est puissant si les postulats sont vérifiés.

3.2.1 Test du signe (pour rappel)

T 3.18 ⇒Exercice

Texte ⇒ Résultats :

- $Y_0 = 4$
- $c = 4$
- $Y_0 \geq c \rightarrow$ l'hypothèse nulle ne peut être rejetée.

En utilisant l'approche IC, on trouve comme IC unilatéral $[40, +\infty[$ $(c = 4)$ et on a $[40, +\infty[\ni 41,2$.

Avec $n = 13$, la valeur la plus proche de $\alpha' = 0,05$ et qui lui est inférieure ou égale est $0,046$ (ce qui correspond à un $\gamma' = 0,954$).

3.2.2 Test de Wilcoxon

T 3.19 ⇒Test de Wilcoxon (méthode directe).

Texte ⇒ **Rappel**

Le test de Wilcoxon à partir de l'I.C. autour de la médiane :

H_0 : « $m_A = m_0$ »
H_1 : « $m_A \neq m_0$ »

n fixé, α fixé, on lit c dans la table 2.4, page 31, ⇒ on obtient $Z_{(c)}$ et $Z^{(c)}$, lus dans le tableau des moyennes deux à deux.

Si $\left[Z_{(c)}\,;\ Z^{(c)}\right] \ni m_0$, accepter H_0

Si $\left[Z_{(c)}\,;\ Z^{(c)}\right] \not\ni m_0$, rejeter H_0

Ce test s'exprime de facon équivalente, à l'aide de la statistique :

$$Y_Z^0 = \text{ nombre de moyennes } Z_{ij} = \frac{X_i + X_j}{2} \text{ inférieures à } m_0$$

Règle de décision :

Si $[c\,;\ n - c + 1] \ni Y_Z^0$, accepter H_0

Si $[c\,;\ n - c + 1] \not\ni Y_Z^0$, rejeter H_0

T 3.20 ⇒Test de Wilcoxon direct, somme des rangs

Texte ⇒ Autre façon de calculer Y_Z^0 :

- calculer les différences $X_i - m_0$;
- affecter un rang aux valeurs absolues des différences $X_i - m_0$;
- calculer la somme des rangs des différences négatives.

⇒ On obtient Y_Z^0 = nombre de moyennes Z_{ij} inférieures à m_0.

T 3.21 ⇒Exercice

Texte ⇒ On reprend les données «âge des agents d'Antibes». On veut tester si l'âge médian (moyen) à Antibes est supérieur à l'âge médian (moyen) national qui est égal à $41,2$. On

va donc faire un test unilatéral. Sur le transparent, les âges des agents échantillonnés sont classés par ordre croissant. Les stagaires doivent remplir les colonnes vides.

X_i	$X_i - m_0$	rang de $\lvert X_i - m_0 \rvert$	signe de $X_i - m_0$
27	-14.2	10	-
32	-9.2	8	-
33	-8.2	7	-
40	-1.2	2	-
41	-0.2	1	-
43	1.8	3	+
45	3.8	4	+
48	6.8	5.5	+
48	6.8	5.5	+
51	9.8	9	+
57	15.8	11	+
58	16.8	12	+
60	18.8	13	+

Comme on a 2 fois $6,8$, on attribue à ces valeurs un rang moyen. Comme statistique de test, on peut utiliser soit la somme des rangs pour les différences négatives, soit la somme des rangs pour les différences positives. Demander aux stagiaires la forme de la région de rejet pour les 2 statistiques ($[0, c]$ pour les différences négatives, $[91 - c, 91]$ pour les différences positives, 91 est la somme des rangs pour $n = 13$). On prend les différences négatives. On lit la valeur de c dans la table 2.4, page 31. Pour $\alpha' = 0,05$, on trouve $c = 22$. On calcule la statistique de test :

$$10 + 8 + 7 + 2 + 1 = 28.$$

Comme $28 > 22$, on ne rejette pas l'hypothèse «l'âge médian (moyen) à Antibes est égal à l'âge médian (moyen) national».

3.2.3 Test de Student

Rappel

On a vu deux tests :

- *test du signe* (I.C. de la médiane)
- *test de Wilcoxon* (I.C. de l'espérance ou de la médiane : méthode de Wilcoxon)

On peut toujours utiliser le *test du signe* (aucune condition d'utilisation n'est nécessaire), sinon la condition implicite : on dispose d'un *échantillon*, c'est-à-dire que l'on a *indépendance entre les* X_i (indépendance assurée par un tirage aléatoire des individus).

Pour le test de *Wilcoxon*, on doit aussi (comme pour n'importe quel test) disposer d'un *échantillon*. De plus, la *distribution théorique* (sur la population) de la V.A. doit être *symétrique*.

Remarque ⇒ Dans le cas du test de *Wilcoxon*, la *distribution* théorique est supposée *symétrique*. Donc, espérance et médiane coïncident. L'I.C. donné par la méthode de Wilcoxon est aussi bien un *I.C. autour de l'espérance* qu'un *I.C. autour de la médiane*.

Par contre, dans le cas du *test du signe*, la *distribution* théorique est *quelconque* (symétrique ou dissymétrique). Donc, espérance et médiane diffèrent. L'I.C. donné par le test du signe est bien un *I.C. autour de la médiane*.

T 3.22 ⇒Test de Student, présentation.

Texte ⇒ *Supposons* que la *distribution théorique* est *normale* (ou pas trop différente), et donc symétrique.

On sait : espérance âge INRA : $\mu_0 = 42$ (valeur de référence).
On note : espérance âge Antibes : μ_A (inconnue).

H_0 : « $\mu_A = \mu_0$ »
H_1 : « $\mu_A > \mu_0$ »

La méthode de Student utilise la statistique :

$$T_0 = \frac{\overline{X} - \mu_0}{\widehat{\sigma}/\sqrt{n}} = \frac{\overline{X} - 42}{\widehat{\sigma}/\sqrt{n}}$$

(T_0 est calculable à partir de l'échantillon, c'est donc une statistique).

Sous H_0, c'est-à-dire si $\mu_A = \mu_0$, $T_0 \sim t_{n-1}$ (T_0 suit une loi de Student à $n-1$ degrés de liberté). Cette loi est connue et tabulée.

Sous H_1, c'est-à-dire si $\mu_A > \mu_0$, la loi de la statistique T_0 n'est pas centrée, c'est une autre loi, décalée vers la droite. Cette loi n'est pas connue car elle dépend de la valeur de μ_A (inconnue).

Remarque sur le nombre de degrés de liberté ⇒ La loi de Student à d degrés de liberté peut être définie de manière synthétique par

$$t(d) = \sqrt{d}\frac{\mathcal{N}(0,1)}{\sqrt{\chi^2(d)}},$$

où $\chi^2(d)$ désigne une loi du χ^2 à d degrés de liberté. La gaussienne et le χ^2 doivent être indépendants. Un χ^2 à d degrés de liberté est une somme de carrés de d gaussiennes indépendantes centrées réduites. On peut écrire T_0 comme

$$T_0 = \sqrt{n-1}\frac{D}{\sqrt{N}},$$

où

$$D = \frac{\bar{X} - \mu_0}{\sqrt{n}\sigma}$$

et

$$N = \sum_i (\frac{X_i - \bar{X}}{\sigma})^2.$$

Les variables D et N sont bien indépendantes. La variable D suit une loi normale centrée réduite. Si on remplaçait le $\bar{X}$ par μ_0 dans l'expression de N, N serait un χ^2 à n degrés de liberté. Mais comme l'espérance de X est estimée par $\bar{X}$, on perd un degré de liberté et N est un χ^2 à $n-1$ degrés de liberté.

Règle de décision :

Si $T_0 > t_c$, (la statistique est supérieure à la valeur critique), rejeter H_0.
Si $T_0 \leq t_c$, accepter H_0.

Pour déterminer la valeur critique t_c, choisir un risque unilatéral α', chercher la valeur critique dans une table de Student.

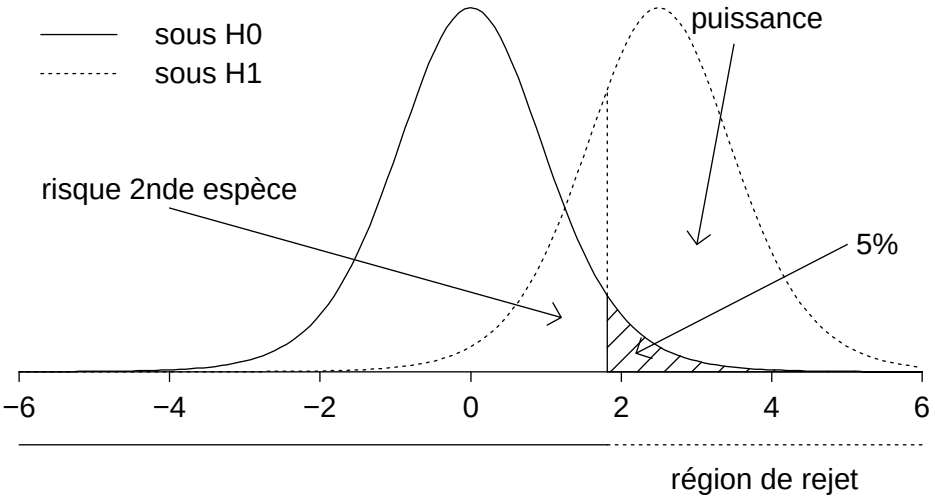


FIG. 3.4 : test de Student unilatéral de $\mu_A = \mu_0$ contre $\mu_A > \mu_0$.

Indiquer une région de rejet sur le transparent. Demander aux stagiaires d'indiquer sur le graphique à quoi correspondent le niveau du test, le risque de 2ième espèce et la puissance.

T 3.23 ⇒Test de Student unilatéral

Texte ⇒

- Reprenons le test de Student précédent.

  C'était un test unilatéral :

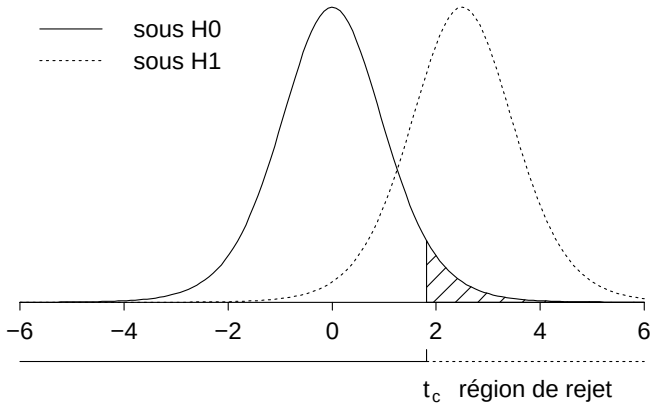
  H_0 : « $\mu_A = \mu_0$ »

  H_1 : « $\mu_A > \mu_0$ »

  On utilise la statistique $T_0 = \frac{\overline{X} - \mu_0}{\hat{\sigma}/\sqrt{n}}$

Si H_1 est vraie, $\mu_A > \mu_0$, T_0 a tendance à être positif, la loi de T_0 sous H_1 est donc décalée vers la droite.

On rejette H_0 si T_0 est « trop grand ».



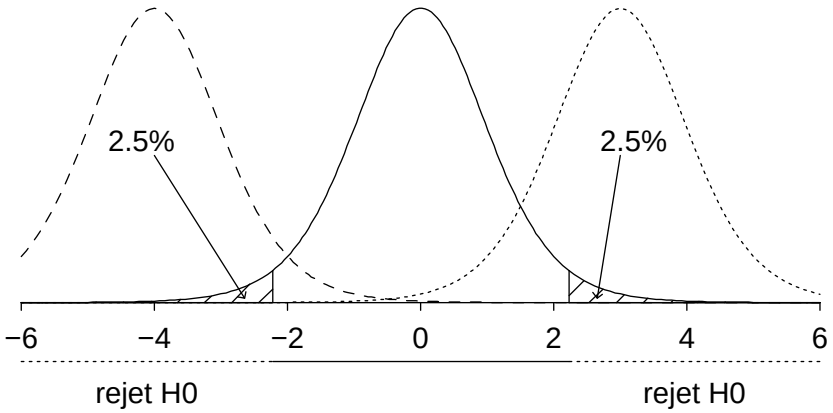
Décision	Réalité	
	H_0	H_1
H_0	$\gamma = 1 - \alpha$	β
H_1	α	$1 - \beta$

T 3.24 ⇒Test de Student bilatéral

Texte ⇒ Considérons maintenant une alternative bilatérale :
H_0 : « $\mu_A = \mu_0$ »
H_1 : « $\mu_A \neq \mu_0$ »

Sous l'alternative H_1 (c.-à-d. si H_1 est vraie), la loi de T_0 peut être décalée soit vers la droite, soit vers la gauche.

Il faut donc choisir une règle de décision qui rejette quand T_0 est « trop grand » ($\mu_A \gg \mu_0$) **et** quand T_0 est « trop petit » ($\mu_A \ll \mu_0$).



Remarque ⇒ On voit sur ce graphique que le risque d'erreur de 2^e espèce (probabilité

d'accepter H_0 à tort), dépend bien de la vraie valeur de m_A ; plus m_A est éloigné de m_0, moins ce risque est élevé.

Par exemple, ce risque est plus faible dans le cas de la courbe de gauche, centrée sur $\mu_A = -4$ et donc plus éloignée de $\mu_0 (= 0$ ici) que la courbe de droite, centrée sur $\mu_A = +2,7$.

Exercice 5

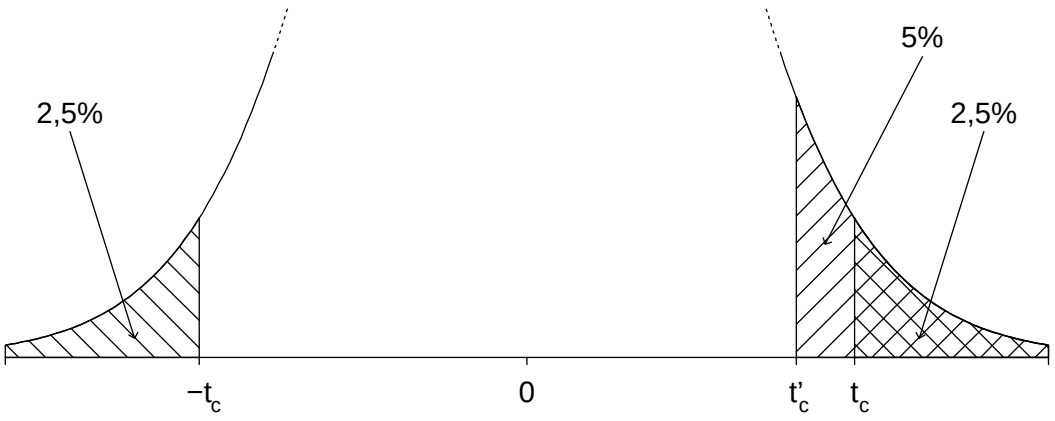
On fait un test bilatéral de niveau $\alpha = 0,05 \Longrightarrow$ valeur critique t_c.

On fait un test unilatéral de niveau $\alpha' = 0,05 \Longrightarrow$ valeur critique t'_c.

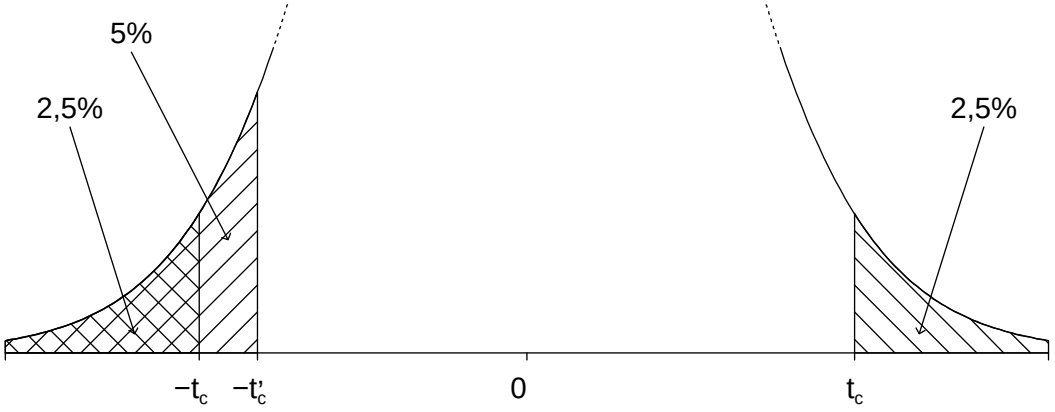
A-t-on $t_c = t'_c$, $t_c > t'_c$ ou $t_c < t'_c$?

Porter t_c et t'_c sur un graphique.

Réponse $\Rightarrow$



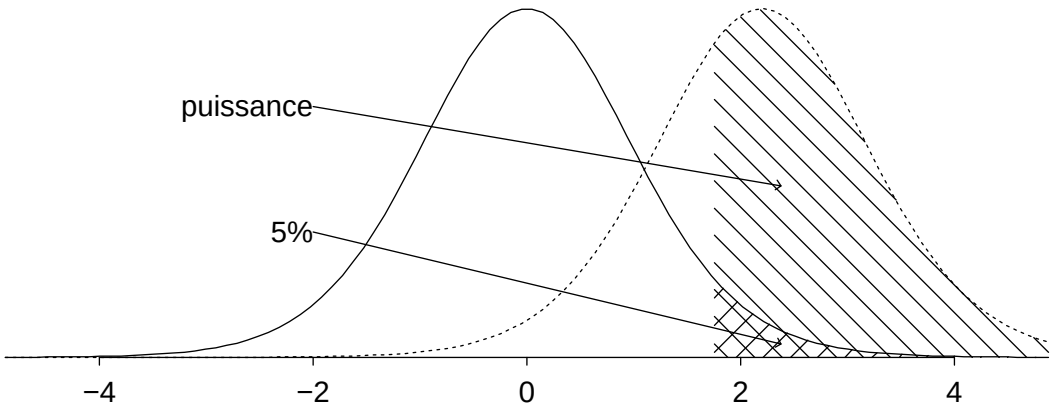
Test bilatéral comparé au test unilatéral $H_1 : \mu_A > \mu_0$



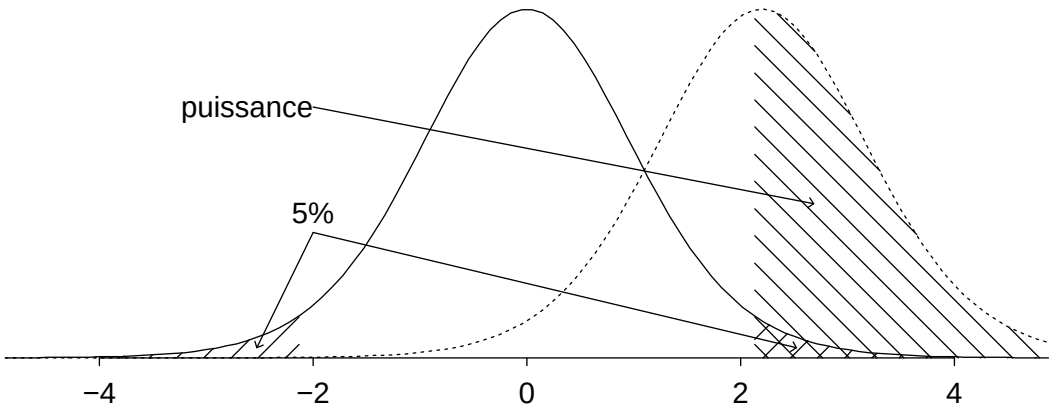
Test bilatéral comparé au test unilatéral $H_1 : \mu_A < \mu_0$

Si la valeur réelle de μ_A est 46, que dire de la puissance des tests unilatéral de niveau 0,05, et bilatéral de niveau 0,05 ? Faire une figure représentant les deux tests.

Réponse ⇒



Puissance du test unilatéral $H_1 : \mu_A > \mu_0$



Puissance du test bilatéral $H_1 : \mu_A \neq \mu_0$

Le test unilatéral est plus puissant. C'est normal, on a fait une hypothèse en plus (μ_A ne peut pas être inférieur à μ_0).

3.3 Niveau descriptif ou P-variable

Résumé ⇒

On introduit ici une nouvelle statistique, qui permet de donner une vue globale des niveaux de confiance obtenus pour un échantillon donné.

On reprendra l'exemple précédent où l'on souhaitait tester si l'espérance d'âge des agents d'Antibes était supérieure ou égale à celle de l'ensemble des agents de l'INRA.

T 3.25 ⇒Niveau descriptif ou P-variable, principe du test

Texte ⇒

- Soit un échantillon de 13 individus tirés de la population des agents d'Antibes, dont on veut étudier l'espérance, m_A.

$$H_0 : m_A = m_0, \quad \text{contre} \quad H_1 : m_A > m_0$$

Remarque ⇒ Le formateur reprendra cet exemple et fera préciser oralement l'ensemble des paramètres à définir pour faire un test (évaluation collective de « Bilan des tests »). Il en viendra alors à un test de Student, sous l'hypothèse d'une distribution gaussienne des âges des agents d'Antibes.

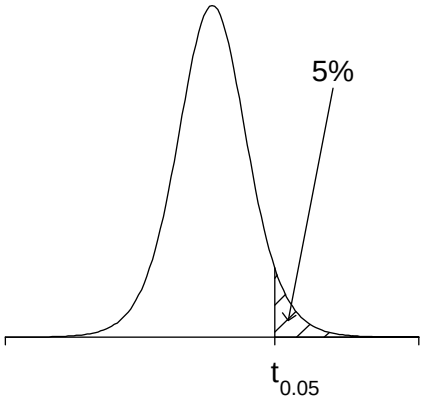
- Après avoir choisi l'erreur α' associée au test unilatéral, on lit, dans la table relative à la loi de Student et pour un nombre de degrés de liberté de 12, la valeur $t_{\alpha'}$ associée.

T 3.26 ⇒Niveau descriptif ou P-variable, règle de décision

Texte ⇒ Extrait de la table de Student à 12 degrés de liberté.

Probabilité de l'extrémité de la loi		Valeur critique
0,5		0,00
0,25		0,69
0,10		1,36
0,05		1,78
⋮		⋮
α'	$\longrightarrow$	$t_{\alpha'}$
⋮		⋮
P'_{T_0}	$\longleftarrow$	T_0
⋮		⋮

/



– La règle est : si $T_0 > t_{\alpha'}$, H_0 est rejetée,

ce qui est strictement équivalent à : si $P'_{T_0} < \alpha'$, H_0 est rejetée.

T 3.27 ⇒Niveau descriptif ou P-variable, définition de la P-variable

Texte ⇒ P'_{T_0} est la valeur lue dans la table, pour la valeur critique T_0 : $P'_{T_0} = 1 - F_0(T_0)$.

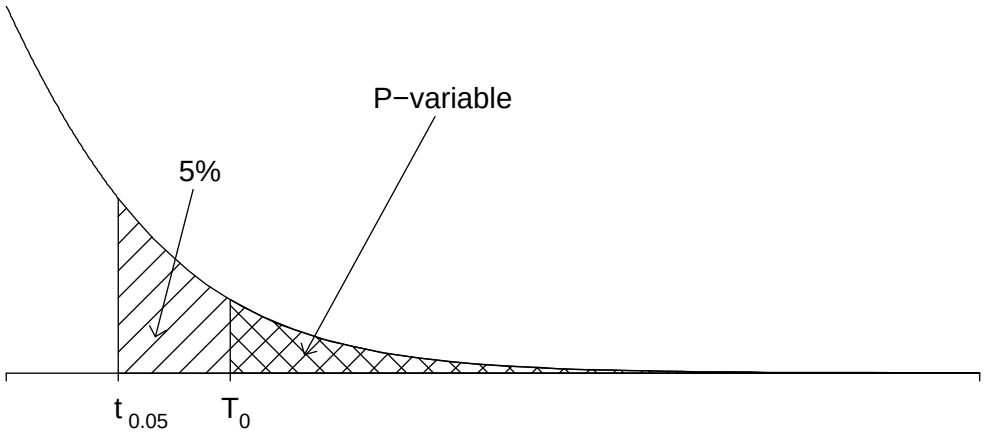


FIG. 3.5 : P-variable associée à la statistique T_0, pour un test unilatéral.

– On a alors défini une nouvelle statistique P'_{T_0}, appelée « niveau descriptif » ou « P-variable » (*P-value* en anglais) ou statistique pivotée, qui est un nombre compris entre

0 et 1 (comme une probabilité), et pour laquelle le calcul de la valeur critique en fonction du niveau de test est très facile ! *(puisque identique...)*

Règle : pour un niveau d'erreur choisi, α',

si $P'_{T_0} < \alpha'$, rejet de H_0,

si $P'_{T_0} > \alpha'$, H_0 non rejetée.

- Dans les cas où H_0 a été rejetée (c'est-à-dire quand le risque d'erreur est bien contrôlé) cette statistique permet de décrire très rapidement la « zone de précision » dans laquelle on se trouve : elle a donc bien un rôle descriptif.

Remarque ⇒ C'est une façon très commode de présenter le résultat d'un test, mais cela ne doit en aucun cas inciter à « choisir le niveau du test en fonction des résultats de l'échantillon »...

Cas d'un test bilatéral

Remarque ⇒ Cette dernière partie est tout à fait facultative. Elle pourra éventuellement être traitée en tant qu'exercice, pour construire la nouvelle statistique pivotée.

T 3.28 ⇒Niveau descriptif ou P-variable, cas bilatéral

Texte ⇒

- On suppose que l'on veut tester si l'âge moyen des agents de Montpellier est égal à celui de l'ensemble des agents de l'INRA, contre l'alternative bilatérale : « l'âge moyen des agents de Montpellier est différent (supérieur ou inférieur) de celui de l'ensemble des agents de l'INRA ».

- On tire un 13-échantillon de la population de Montpellier dont on veut étudier l'âge moyen, m_M.

$$H_0 \ : \ m_M = m_0 \qquad H_1 \ : \ m_M \neq m_0$$

- La statistique pivotée est alors différente de la précédente, puisqu'elle prend en compte les deux extrémités de la distribution (fig. 3.6) :

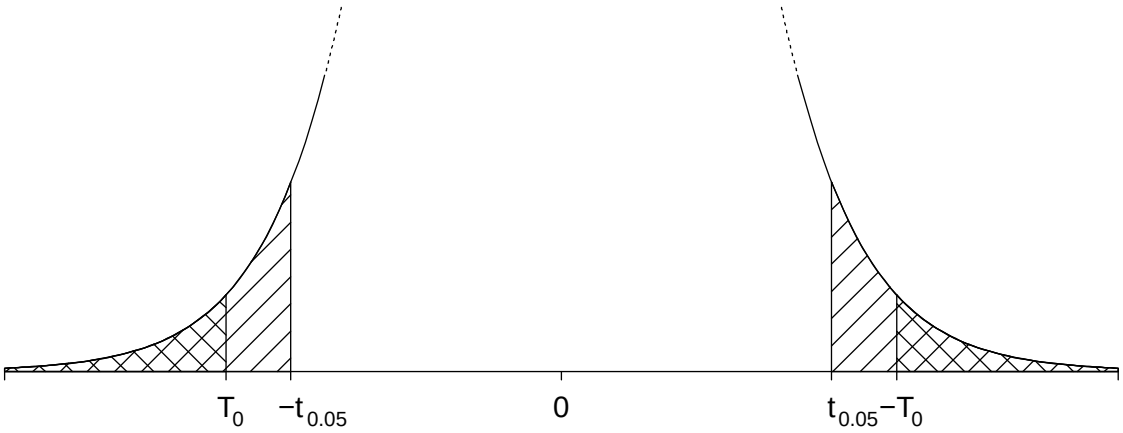


FIG. 3.6 : P-variable associée à la statistique T_0 (ou $|T_0|$) pour le test bilatéral (P_{T_0} est la surface de deux zones hachurées).

Si on appelle F_0 la fonction de répartition de la statistique, on a :

$$\text{si} \quad T_0 < 0, \quad P_{T_0} = F_0(T_0) + [1 - F_0(-T_0)]$$

$$\text{si} \quad T_0 > 0, \quad P_{T_0} = [1 - F_0(T_0)] + F_0(-T_0)$$

$$\text{ou encore :} \quad \forall \ T_0, \quad P_{T_0} = F_0(-|T_0|) + [1 - F_0(|T_0|)]$$

3.4 Bilan sur les tests

T 3.29 ⇒Bilan sur les tests

Texte ⇒ Les étapes à suivre.

- On identifiera :

  - la population,

  - la variable d'étude,

  - le paramètre de la loi à étudier.

- On précisera :

  - la distribution postulée de la variable,

  - les hypothèses H_0 *et* H_1.

T 3.30 ⇒Bilan sur les tests, suite et fin

Texte ⇒

- On choisira une statistique S

  - de loi connue sous H_0,

  - dont la forme de la loi est connue au moins approximativement sous H_1.

- On choisira un niveau de confiance γ.

- On pourra alors définir le domaine de rejet de H_0, et calculer la valeur critique s_α pour un échantillon de taille n $(\alpha = 1 - \gamma)$.

- Après avoir tiré un n-échantillon de mesures, on calculera la réponse du test sur les mesures de cet échantillon :

  si $S > s_\alpha$ on rejette H_0, et

  si $S < s_\alpha$ on ne rejette pas H_0.

Dans les exemples de tests de Student, on a pris $S = T_0$ ou $S = -T_0$ pour les tests unilatéraux, et $S = |T_0|$ pour le test bilatéral.

Chapitre 4

Exemples d'applications des tests

4.1 Introduction

Ce chapitre, «exemples d'application de tests», est constitué d'un ensemble de problèmes. Chaque problème est résolu indépendamment des autres. Cette partie du module est un chapitre à la carte, où le formateur en concertation avec les stagiaires pourra faire un choix de présentation de méthodes à l'aide de la résolution d'exercices. L'objectif essentiel de ce chapitre est de familiariser les stagiaires avec les fiches aide-mémoire.

T 4.1 $\Rightarrow$Arbre de NESI[1] (comparaison des paramètre de position et des distributions)

Il s'agit de comparer 2 échantillons, on ne traitera pas le cas où les 2 échantillons sont appariés qui se rapporte aux chapitres précédents (dans ce cas on fabrique un unique échantillon sur la différence des données appariées que l'on compare à la valeur de référence 0). On parle d'échantillons appariés lorsqu'on compare sur un même individu (ou sur deux individus très proches, par ex. jumeaux) deux traitements différents.

Dans le cas de 2 échantillons indépendants, on les comparera soit en fonction d'un paramètre de position (moyenne, médiane, quantile, ...), soit en fonction de leurs distributions.

Les fiches aide-mémoire seront consultées par les stagiaires pour choisir le ou les tests appropriés.

4.2 Problème 1 : Comparaison des hauteurs des arbres de deux types de forêts.

Résumé :

$$\text{non appariés} \rightarrow \text{forme identique} \begin{cases} \text{dispersions identiques} \rightarrow \text{gaussienne} \rightarrow \text{Welch} \\ \text{dispersions différentes} \rightarrow \text{gaussienne} \rightarrow \text{Student} \end{cases}$$

La hauteur dominante d'un peuplement est la hauteur moyenne des 100 plus gros arbres à l'hectare. Cette hauteur dominante à un âge donné est un bon indicateur de la productivité

1. NESI est un module de SPLUS d'aide au choix d'un test statistique développé à l'unité de biométrie de Jouy-en-Josas

du peuplement forestier. Les 2 types de forêts peuvent être différenciés sur le lieu, le type d'arbre, ... A l'aide de l'exploration graphique (histogrammes, boxplot et qqnorm), on admet que les populations de hauteurs sont normales et leurs variances sont égales. Les échantillons sont de taille $m = 13$ et $n = 14$.
Choix du test de Student : Comparaison des moyennes des deux types de forêts

Postulats et Limites du Test

- indépendance : indispensable.
- inégalité des variances : pas trop important si les échantillons sont de même taille.
- normalité : pas trop important si on a de grands échantillons (les risques de première et deuxième espèce sont peu modifiés), notamment si les distributions sont symétriques et même si elles sont très différentes de la normalité.

Dans le cas où les échantillons sont d'effectifs très inégaux, il devient indispensable de s'assurer de l'égalité des variances. Pour l'exemple des forêts, le test de Fisher permet d'accepter l'hypothèse nulle d'égalité des variances.

C'est la façon dont on a réalisé l'échantillon qui permet de conclure à l'indépendance des X_i entre eux et à l'indépendance des Y_j entre eux. Cependant, on peut, si on a des doutes, effectuer une représentation graphique de X_{i+1} en fonction de X_i, ou faire un test de run. Le test de run est fondé sur l'observation de la longueur des suites d'observations identiques (en-dessous de la médiane ou de la moyenne par exemple). Ces méthodes sont adaptées dans le cas de données temporelles par exemple, pour mettre en évidence des linéarités. Elles ne sont pas intéressantes pour cet exemple.

On peut vérifier l'indépendance deX_i et Y_j en faisant un graphique de X_i en fonction de Y_j associé à un test de corrélation. Si X_i et Y_j suivent des distributions normales on utilisera le coefficient de corrélation de Pearson, sinon les coefficients de corrélation non paramétriques (coefficient de corrélation de rangs de Spearman r_s et coefficient de corrélation de rangs de Kendall τ). On pourra se référer aux fiches aide-mémoire pour plus d'informations.

On peut tester la normalité graphiquement à l'aide de la fonction S `qqnorm`. La représentation graphique est sensiblement une droite (droite de Henry : quantiles observés / quantiles d'une loi normale) si la distribution est normale. Si on compare deux distributions, sur le même principe on utilisera la fonction S `qqplot` ou la procédure SAS `univariate` avec l'option `normal`.

Méthode de calcul

- rejet de H_0 si $|T| > t_{m+n-2}(5\%)$
- sous l'hypothèse alternative $H_1 = \{\mu_X \neq \mu_Y\}$, T suit une loi inconnue.
- avec les données : hauteurs d'arbres

$$\text{ddl} = 13 + 14 - 2 = 25$$

$$t_{1-\alpha/2} = 2,06 \text{ (test bilatéral - à lire dans la table)}$$

Calcul de la statistique T :

$$\bar{X} = \frac{1}{m}\sum x_i \simeq 25,97$$

$$\sum (x_i - \bar{X})^2 \simeq 22,15$$

$$\bar{Y} = \frac{1}{n}\sum y_i \simeq 25,38$$

$$\sum (y_i - \bar{Y})^2 \simeq 32,58$$

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sum (x_i-\bar{X})^2 + \sum (y_i-\bar{Y})^2}{m+n-2}(\frac{1}{m}+\frac{1}{n})}}$$

$$T \simeq 0,9542$$

Conclusion :

$$T \simeq 0,9542 < t_{1-\alpha/2} = 2,06$$

On ne rejette pas H_0, les moyennes entre `FORET1` et `FORET2` ne sont pas significativement différentes.

Commentaires sur la sortie Splus

Test de Fisher (ou Berhens-Fisher-Snedecor ou Fisher-Snedecor)

- `F` correspond à la statistique de Fisher, quotient des variances estimées, c'est-à-dire quotient de $\hat{\sigma}^2(X) = (\sum (x_i - \bar{X})^2)/m - 1$ et de $\hat{\sigma}^2(Y) = (\sum (y_i - \bar{Y})^2)/n - 1$.
- `num df` correspond au degré de liberté du numérateur `FORET1` et vaut $m - 1$.
- `denom df` correspond au degré de liberté du dénominateur `FORET2` et vaut $n - 1$.
- `p-value` correspond au risque d'erreur réel lorsque l'on rejette l'hypothèse nulle. Si le risque est faible ($< 5\%$), on rejette l'hypothèse nulle.
- `alternative hypothesis` : l'hypothèse alternative est définie par «le quotient des variances est différent de 1».

Test de Student

- `t` correspond à la statistique de Student (notée T dans ce cours).
- `df` correspond au nombre de degré de liberté, df $= m + n - 2$.
- `95 percent confidence interval` correspond à l'intervalle de confiance à 95 % de la différence des moyennes :

$$I.C. = [I_m; I_M]$$

$$I_m = (\bar{X} - \bar{Y}) - t_{1-\alpha/2} \times \sqrt{\frac{\sum (x_i - \bar{X})^2 + \sum (y_i - \bar{Y})^2}{m+n-2}(\frac{1}{m}+\frac{1}{n})}$$

$$I_M = (\bar{X} - \bar{Y}) + t_{1-\alpha/2} \times \sqrt{\frac{\sum (x_i - \bar{X})^2 + \sum (y_i - \bar{Y})^2}{m+n-2}(\frac{1}{m} + \frac{1}{n})}$$

$$I_m \leq \mu_X - \mu_Y \leq I_M$$

$$I.C. = [-0.676; 1.843]$$

- `mean of X` correspond à la moyenne empirique de X, $\bar{X} = \frac{1}{m}\sum x_i$

Résumé des commandes Splus utilisées

Graphiques

- `par` : instruction graphique permettant d'obtenir plusieurs graphiques sur une même page.
- `hist` : construction d'un histogramme.
- `boxplot` : construction d'un boxplot.
- `qqnorm` : vérification graphique de la normalité d'une distribution.

Tests

- `var.test` : test de Fisher d'égalité des variances.
- `t.test` : test de student d'égalité de 2 moyennes dans le cas de variances égales.

Commentaires sur la sortie SAS

- le `Std Dev` correspond à $\widehat{\sigma}_X$;
- le `Std Error` correspond à $\widehat{\sigma}_{\overline{X}}$;
- la valeur `Prob > |T|` correspond à la P-variable $= P$;
- le programme fait un test d'homogénéité des variances, or :

  - on a supposé $\sigma_X = \sigma_Y = \sigma$ dans nos hypothèses de départ, et

  - ce test n'est «pas terrible» de toute façon,

  et tout marche à peu près bien si les effectifs sont égaux mais les variances pas bien homogènes ;
- donc on n'a pas à regarder la ligne de résultats pour `variance unequal`, qui correspond à un calcul de T modifié, avec un nombre de degrés de liberté calculé différemment ;
- on n'a donc à prendre que les résultats de la ligne `equal` ; on obtient donc T_0 avec les degré de liberté et la P-variable associés.

Remarque $\Rightarrow$ Dans le cas de variances «inégales», SAS calcule un T_0 modifié :

$$T'_O = \frac{\overline{X} - \overline{Y}}{\sqrt{(W_1 + W_2)}}$$

où $W_1 = \widehat{\sigma}_X^2/m$ et $W_2 = \widehat{\sigma}_Y^2/n$

Le nombre de degrés de liberté est alors calculé avec la formule :

$$ddl = (W_1 + W_2)^2 \Big/ \left(\frac{W_1^2}{m-1} + \frac{W_2^2}{n-1} \right)$$

Si *a priori* les variances sont «peu différentes», on peut prendre ce test corrigé!
Sinon, envisager plutôt un autre test : la comparaison de distributions par exemple.

Résumé des commandes SAS utilisées

Remarque : Le programme est écrit sur deux colonnes sur le transparent, mais en réalité la deuxième colonne est à la suite de la première.

- `proc ttest` : procédure réalisant le test de Student d'égalité des moyennes (les variances des 2 échantillons sont considérées identiques) et le test de Welch d'égalité des moyennes (les variances des 2 échantillons sont considérées différentes). Il fournit aussi le test de Fisher d'égalité des variances.

4.3 Problème 2 : Comparaison du taux horaire de diminution de sucres

Exploration graphique

Le test de Student n'est pas applicable sur cet exemple puisque les distributions des deux échantillons indépendants ne sont pas normales. On observe quelques valeurs exceptionnelles, mais les distributions des deux échantillons indépendants semblent être identiques à un décalage près de la position de la médiane. On peut alors utiliser le test de Mann-Withney pour tester la différence des médianes des réponses des deux échantillons.

Hypothèses et postulats

A priori on a aucune raison de supposer une réponse supérieure à l'autre selon la dose de médicament, on effectue un test bilatéral.

$$H_0 = \{\text{égalité des médianes } m_X = m_Y\}$$

$$H_1 = \{m_X \neq m_Y\}$$

Postulats :

- les deux échantillons sont indépendants,

- les X_i ont tous une même distribution,
- les Y_j ont tous une même distribution,
- ces distributions sont continues.

Calcul de la statistique U :
On note n et m la taille respective des échantillons X (`doseA`) et Y (`doseB`).

- Ordonner toutes les valeurs de X_i et Y_j ensemble.
- Attribuer un rang (par ordre croissant) à chacune.

$$S_X = \text{somme des rangs des } X_i$$
$$S_Y = \text{somme des rangs des } Y_j$$

Remarque : Si ex-æquo prendre le rang moyen.

On calcule ensuite

$$U_X = S_X - n(n+1)/2$$
$$U_Y = S_Y - m(m+1)/2.$$

On observera que $n(n+1)/2$ est la somme des rangs des X_i si tous les X_i sont inférieurs aux Y_j. De même, $m(m+1)/2$ est la somme des rangs des Y_j si tous les Y_j sont inférieurs aux X_i. Une petite valeur de U_X ou de U_X indique que les deux distriubutions sont décalées. La statistique de test est

$$U = \inf(U_X, U_Y).$$

$\rightarrow$ Remarque : $U_X + U_Y = mn$ (nombre total de couples). Donc, en pratique, on peut se contenter d'effectuer un seul calcul soit U_X soit U_Y pour obtenir la statistique U.

Pour effectuer le test, on compare la statistique U calculée à la valeur critique lue dans la table.

Exemple :

	Rang Dose A	Rang Dose B
-16.20	1	
-15.87	2	
-12.81	3	
-11.13	4	
-10.87		5
-10.19	6	
-10.10	7	
-9.67	8	
-6.27		9
-3.76		10
-2.32		11
0.21	12	
1.59		13
1.96	14	
2.66		15
3.02		16
7.23		17
15.01		18

$$S_X = 57 \text{ et } S_Y = 114$$

$$U_X = 57 - 9 \times 10/2 = 12$$

$$U_Y = 114 - 9 \times 10/2 = 69$$

Statistique : $U = 12$ (on vérifie que $9 \times 9 = 81$ et $81 - 12 = 69$: $U_X + U_Y = mn$).
Valeurs critiques du test bilatéral :

$$U_c = 17 \text{ (5 \%) et } U_c = 11 \text{ (1 \%)}$$

Rejet de H_0 au profit de H_1 si $U < U_c$
Conclusion :
On rejette H_0 au profit de H_1 à 5 %. On conclut à une différence de réponse selon les variables `doseA` ou `doseB`.

Commentaires sur la sortie Splus

Le test de somme des rangs de Wilcoxon calculant la statistique W est équivalent au test de Mann-Withney.

- `W` correspond à la statistique de Wilcoxon, W = la plus petite somme des rangs des 2 échantillons.
- `n` correspond à la taille de l'échantillon `doseA`.
- `m` correspond à la taille de l'échantillon `doseB`.
- `p-value` correspond au risque d'erreur réel lorsque l'on rejette l'hypothèse nulle. Si le risque est faible (< 5 %), on rejette l'hypothèse nulle.
- `alternative hypothesis` : l'hypothèse alternative est définie par «la différence des moyennes est 0».

Résumé des commandes Splus utilisées

Graphiques

- `par` : instruction graphique permettant d'obtenir plusieurs graphiques sur une même page.
- `hist` : construction d'un histogramme.
- `boxplot` : construction d'un boxplot.
- `qqnorm` : vérification graphique de la normalité d'une distribution.

Tests

- `wilcox.test` : test de Wilcoxon.

Commentaires sur la sortie SAS

- `N` correspond à la taille de l'échantillon.
- `Sum of Scores` correspond à la somme des rangs de l'échantillon.
- `Expected Under H0` correspond à la somme des rangs espérés sous l'hypothèse nulle H_0, ici (57+114)/2=85.5.
- `Std Dev Under H0` correspond à l'écart type estimé de la somme des rangs sous l'hypothèse nulle H_0.
- `mean score` correspond à la moyenne de la somme des rangs, c'est-à-dire 57/9 et 114/9.
- `S` correspond à la statistique de Wilcoxon, S = la plus petite somme des rangs des 2 échantillons.
- `Z` correspond à la statistique Z qui est une approximation normale avec correction de la continuité de la statistique S, $Z = (S - \text{Expected})/\text{Std Dev}$.
- `Prob > |Z|` correspond à p-value.

Résumé des commandes SAS utilisées

Remarque : Le programme est écrit sur deux colonnes sur le transparent, mais en réalité la deuxième colonne est à la suite de la première.

- `proc npar1way` : cette procédure fournit un ensemble de résultats de comparaison de deux échantillons, et en particulier le test de Wilcoxon.

4.4 Problème 3 : Changement d'isolation

La firme Hotpot veut tester l'efficacité d'une nouvelle isolation mois coûteuse pour ses fours. Plus les temps de refroidissement sont élevés, meilleure est l'isolation.

Le statisticien est confronté à deux optiques de comparaison. L'une consiste en une comparaison de paramètre de position (médiane, moyenne, ...), et l'autre en une comparaison globale des distributions.

4.4.1 Comparaison de paramètre de position

Pour les deux échantillons, à un four près, on est assez proche de la normalité. On peut envisager des tests basés ou non sur le postulat de normalité.

Sans postulat de la normalité

On envisage le test de Mann-Whitney-Wilcoxon puisque les X_i ont tous même distribution continue, les Y_j ont tous même distribution continue et la distribution des X_i ne diffère de celle des Y_j que par un décalage de la médiane.

$$H_0 = \{\text{égalité des médianes : } m_X = m_Y\}$$

$$H_1 = \{m_X \neq m_Y\}$$

Calcul de la statistique U :
On note n et m la taille respective de l'échantillon X (`isolstd`) et Y (`isolnvl`), et $N = n+m$ le nombre total d'observations.

$$S_X = \text{somme des rangs des } X_i$$

$$S_X = 3 + 6 + 9 + 10 + 12 + 15 + 16 + 17 = 88$$

$$S_Y = \text{somme des rangs des } Y_j$$

$$S_Y = 1 + 2 + 4 + 5 + 7 + 8 + 11 + 13 + 14 = 65$$

$$U_X = S_X - n(n+1)/2 = 52$$

$$U_Y = S_Y - m(m+1)/2 = 20$$

$$U = \inf(U_X, U_Y) = 20$$

Au seuil de 5 %, la valeur critique du test bilatéral est $U_c = 15$. Comme $U \geq U_c$, on ne peut pas rejeter l'hypothèse nulle au seuil de 5 %. Ainsi, la firme a raison, on ne peut pas conclure à une différence de perte de chaleur.
Test Unilatéral :

$$H'_0 = \{m_X \leq m_Y\}$$

$$H'_1 = \{m_X > m_Y\}$$

Au seuil de 5 %, la valeur critique du test unilatéral est $U_c = 18$. Comme $U \geq U_c$, on ne peut pas rejeter l'hypothèse nulle au seuil de 5 %. Ainsi, la firme a raison, on ne peut pas conclure à une différence de perte de chaleur.

Commentaires sur la sortie Splus Le test réalisé est unilatéral.

Le test de somme des rangs de Wilcoxon calculant la statistique W est équivalent au test de Mann-Withney.

- `W` correspond à la statistique de Wilcoxon,W = la plus petite somme des rangs des 2 échantillons.
- `n` correspond à la taille de l'échantillon `isolstd`.
- `m` correspond à la taille de l'échantillon `isolnvl`.
- `p-value` correspond au risque d'erreur réel lorsque l'on rejette l'hypothèse nulle. Si le risque est faible (< 5 %), on rejette l'hypothèse nulle.
- `alternative hypothesis` : l'hypothèse alternative du test unilatéral est définie par «la différence des moyennes est supérieure à 0».

Résumé des commandes Splus utilisées

- `wilcox.test` : test de Wilcoxon.

Commentaires sur la sortie SAS Le test réalisé est bilatéral.

- `N` correspond à la taille de l'échantillon.
- `Sum of Scores` correspond à la somme des rangs de l'échantillon.
- `Expected Under H0` correspond à la somme des rangs espérés sous l'hypothèse nulle H_0.
- `Std Dev Under H0` correspond à l'écart type estimé de la somme des rangs sous l'hypothèse nulle H_0.
- `mean score` correspond à la moyenne de la somme des rangs.
- `S` correspond à la statistique de Wilcoxon, S = la plus petite somme des rangs des 2 échantillons.
- `Z` correspond à la statistique Z qui est une approximation normale avec correction de la continuité de la statistique S, $Z = (S - \text{Expected})/\text{Std Dev}$.
- `Prob > |Z|` correspond à p-value.

Résumé des commandes SAS utilisées

- `proc npar1way`: cette procédure fournit un ensemble de résultats de comparaison de deux échantillons, et en particulier le test bilatéral de Wilcoxon.

Avec postulat de la normalité

Néanmoins on peut supposer la normalité et l'égalité des variances satisfaites. On effectue un test de Fisher. On ne rejette pas l'hypothèse nulle et on accepte que les variances soient égales.

On considère que :

$$X_i \sim \mathcal{N}(\mu_X, \sigma^2) \text{ et } Y_i \sim \mathcal{N}(\mu_Y, \sigma^2)$$

$$H_0 = \{\text{égalité des moyennes} : \mu_X = \mu_Y\}$$

$$H_1 = \{\mu_X \neq \mu_Y\}$$

$$\bar{X} = \frac{1}{8}\sum_i x_i \simeq 15.26 \qquad \sum_i (x_i - \bar{X})^2 \simeq 7.98$$

$$\bar{Y} = \frac{1}{9}\sum_i y_i \simeq 14.43 \qquad \sum_i (y_i - \bar{Y})^2 \simeq 6$$

$$\hat{\sigma}^2 = \left[\sum_i (x_i - \bar{X})^2 + \sum_i (y_i - \bar{Y})^2\right] / (m + n - 2) \simeq 0.93$$

$$t = \frac{\bar{X} - \bar{Y}}{\hat{\sigma}\sqrt{1/m + 1/n}} \simeq 1.77$$

ddl $= m + n - 2 = 15$

La valeur critique au seuil de 5 % du test bilatéral de student à 15 degré de liberté est : $t_{1-\alpha/2} = 2.093$. Comme $|t| < t_{1-\alpha/2}$, on ne peut pas rejeter l'hypothèse nulle.
Test unilatéral :

$$H_0' = \{\mu_X \leq \mu_Y\}$$
$$H_1' = \{\mu_X > \mu_Y\}$$

La valeur critique au seuil de 5 % du test unilatéral de student à 15 degré de liberté est : $t_{1-\alpha} = 1.753$. Comme $t > t_{1-\alpha}$, on rejette l'hypothèse nulle. On conclut à une différence de perte de chaleur.

Commentaires sur la sortie Splus Test de Fisher :

- `F` correspond à la statistique de Fisher, quotient des variances estimées, c'est-à-dire quotient de $\hat{\sigma}^2(X) = (\sum (x_i - \bar{X})^2)/n - 1$ et de $\hat{\sigma}^2(Y) = (\sum (y_i - \bar{Y})^2)/m - 1$.
- `num df` correspond au degré de liberté du numérateur `isolstd` et vaut $n - 1$.
- `denom df` correspond au degré de liberté du dénominateur `isolnvl` et vaut $m - 1$.
- `p-value` correspond au risque d'erreur réel lorsque l'on rejette l'hypothèse nulle. Si le risque est faible (< 5 %), on rejette l'hypothèse nulle.
- `alternative hypothesis` : l'hypothèse alternative est définie par «le quotient des variances est différent de 1».

Test de Student :

- `t` correspond à la statistique de Student.
- `df` correspond au nombre de degré de liberté, df $= m + n - 2$.
- `95 percent confidence interval` correspond à l'intervalle de confiance à 95 % de la différence des moyennes et est donné par la relation :

Pour un test bilatéral :

$$I.C. = [I_m; I_M]$$
$$I_m = (\bar{X} - \bar{Y}) - t_{1-\alpha/2} \times \sqrt{\frac{\sum (x_i - \bar{X})^2 + \sum (y_i - \bar{Y})^2}{m+n-2}(\frac{1}{m} + \frac{1}{n})}$$
$$I_M = (\bar{X} - \bar{Y}) + t_{1-\alpha/2} \times \sqrt{\frac{\sum (x_i - \bar{X})^2 + \sum (y_i - \bar{Y})^2}{m+n-2}(\frac{1}{m} + \frac{1}{n})}$$
$$I_m \leq \mu_X - \mu_Y \leq I_M$$
$$I.C. = [-017; 1.83]$$

Pour un test unilatéral ($H_0' = \{\mu_X \leq \mu_Y\}$ contre $H_1' = \{\mu_X > \mu_Y\}$) :

$$I_m = (\bar{X} - \bar{Y}) - t_{1-\alpha} \times \sqrt{\frac{\sum (x_i - \bar{X})^2 + \sum (y_i - \bar{Y})^2}{m+n-2}(\frac{1}{m} + \frac{1}{n})}$$
$$I.C. = [I_m; \infty]$$

- Avec Splus, on a fait un test unilatéral, 0 n'appartient pas à l'IC $[0.007, +\infty[$. Donc on rejette l'hypothèse nulle au profit de H_1'.
- `mean of X` correspond à la moyenne empirique de X, $\bar{X} = \frac{1}{n}\sum x_i$

Résumé des commandes Splus utilisées

- `var.test` : test de Fisher d'égalité des variances.
- `t.test` : test de student d'égalité de 2 moyennes dans le cas de variances égales.

Commentaires sur les sorties SAS

- le `Std Dev` correspond à $\widehat{\sigma}_X$;
- le `Std Error` correspond à $\widehat{\sigma}_{\overline{X}}$;
- la valeur `Prob > |T|` correspond à la P-variable $= P$; le test réalisé est bilatéral ; on ne rejette pas l'hypothèse nulle au niveau 5%;
- le programme fait un test d'homogénéité des variances, or :

  - on a supposé $\sigma_X = \sigma_Y = \sigma$ dans nos hypothèses de départ, et

  - ce test n'est «pas terrible» de toute façon,

  et tout marche à peu près bien si les effectifs sont égaux mais les variances pas bien homogènes ;
- donc on n'a pas à regarder la ligne de résultats pour `variance unequal`, qui correspond à un calcul de T modifié, avec un nombre de degrés de liberté calculé différemment ;
- on n'a donc à prendre que les résultats de la ligne `equal` ; on obtient donc T_0 avec les degré de liberté et la P-variable associés.

Remarque $\Rightarrow$ Dans le cas de variances «inégales», SAS calcule un T_0 modifié :

$$T_O' = \frac{\overline{X} - \overline{Y}}{\sqrt{(W_1 + W_2)}}$$

où $W_1 = \widehat{\sigma}_X^2/m$ et $W_2 = \widehat{\sigma}_Y^2/n$

Le nombre de degrés de liberté est alors calculé avec la formule :

$$ddl = (W_1 + W_2)^2 \Big/ \left(\frac{W_1^2}{m-1} + \frac{W_2^2}{n-1} \right)$$

Si *a priori* les variances sont «peu différentes», on peut prendre ce test corrigé!
Sinon, envisager plutôt un autre test : la comparaison de distributions par exemple.

Résumé des commandes SAS utilisées Remarque : Le programme est écrit sur deux colonnes sur le transparent, mais en réalité la deuxième colonne est à la suite de la première.

- `proc ttest` : procédure réalisant le test de Student d'égalité des moyennes (les variances des 2 échantillons sont considérées identiques) et le test de Welch d'égalité des moyennes (les variances des 2 échantillons sont considérées différentes). Il fournit aussi le test de Fisher d'égalité des variances.

4.4.2 Comparaison des distributions

En général, comparer globalement des distributions est moins puissant que d'effectuer des comparaisons sur des paramètres de position. Le test de Cramer-Von-Mises est une extension du test de Kolmogorov-Smirnov et est plus puissant.

x_i	y_i	$F_X(z)$	$F_Y(z)$	$F_X(z) - F_Y(z)$	$(F_X(z) - F_Y(z))^2$
	12.9	0	1/9	-1/9	1/81
	13.7	0	2/9	-2/9	4/81
13.9		1/8	2/9	-7/72	$49/(72)^2$
	14	1/8	3 /9	-15/72	$225/(72)^2$
	14.1	1/8	4/9	-23/72	$529/(72)^2$
14.2		2/8	4/9	-14/72	$196/(72)^2$
	14.4	2/8	5/9	-22/72	$484/(72)^2$
	14.7	**2/8**	**6/9**	**-30/72**	$900/(72)^2$
14.8		3/8	6/9	-21/72	$441/(72)^2$
14.9		4/8	6/9	-12/72	$144/(72)^2$
	15.1	4/8	7/9	-20/72	$400/(72)^2$
15.3		5/8	7/9	-11/72	$121/(72)^2$
	15.4	5/8	8/9	-19/72	$361/(72)^2$
	15.6	5/8	9/9	-27/72	$729/(72)^2$
15.7		6/8	9/9	-18/72	$324/(72)^2$
16.1		7/8	9/9	-9/72	$81/(72)^2$
17.2		8/8	9/9	0	0

Test de Kolmogorov-Smirnov

La plus forte déviation est observée pour une température de 14.7, on a ainsi :

$$D = \sup_z |F_X(z) - F_Y(z)|$$

$$D = \frac{30}{72} \simeq 0.416$$

$D \simeq 0.416$ est inférieur à la valeur critique 0.639 (lue dans la table) au seuil de 5 % du test bilatéral (table A8 du Sprent). On ne rejette pas l'hypothèse nulle, c'est à dire, on ne ne peut pas conclure à une différence significative de la nouvelle isolation par rapport à l'ancienne. Avec le test unilatéral, la valeur critique aurait été de $0,556$, on ne rejette pas l'hypothèse nulle, non plus.

Test de Cramer-Von-Mises

$$D^2 = \sum_z [F_X(z) - F_Y(z)]^2$$

$$T = \frac{mnD^2}{(m+n)^2}$$

On trouve $D^2 \simeq 1.023$ et $T \simeq \frac{8\times 9\times 1.023}{17^2} \simeq 0.255$. On compare T à $0,461$ (resp. $0,743$) pour un test bialtéral de niveau 5% (resp. 1%). On ne rejette pas l'hypothèse nulle.
Remarque : D^2 est une notation et par conséquent $D^2 \neq (0.416)^2$.

Commentaires sur la sortie SAS

Les tests suivants sont bilatéraux.

Test de Kolmogorov-Smirnov

- `N` correspond à la taille de l'échantillon.
- `EDF at maximum` correspond à la valeur de la fonction de distribution empirique lorsque l'on est au maximum de la déviation (ou écart) entre les deux distributions empiriques F_X et F_Y, et vaut $2/8$ pour `isolstd` et $6/9$ pour `isolnvl`.
- `deviation from Mean at maximum` correspond à l'écart entre la distribution empirique de l'échantillon (F_X ou F_Y) et la distribution empirique moyenne F calculée en considérant un seul échantillon lorsque l'on est au maximum de la déviation, et vaut $\sqrt{n}(F_X(z) - F(z))$ pour `isolstd` et $\sqrt{m}(F_Y(z) - F(z))$ pour `isolnvl`.

  $$F(z) = (1/(m+n))(\text{nombre de } x_i \text{ et de } y_i \leq z).$$

  Dans l'exemple : $F(z) = (1/(8+9)) \times (2+6) = 0,47$.
- `Maximum Deviation occured at Observation` correspond au numéro de l'observation lorsque l'on est au maximum de la déviation, c'est l'observation 11 (numéro dans la liste donnée au programme SAS).
- `Value of TEMPS at maximum` correspond à la valeur prise par l'observation lorsque l'on est au maximum de la déviation, ici 14.7.
- `KS` correspond à la statistique de Kolmogorov-Smirnov, mesure de l'écart maximal de la distribution empirique de chaque échantillon à la distribution empirique F.
- `KSa` correspond à la statistique de Kolmogorov-Smirnov asymptotique, et vaut $\sqrt{m+n}KS$.
- `D` correspond à l'écart maximal entre les deux distributions empiriques F_X et F_Y, $D = \sup_z |F_X(z) - F_Y(z)|$.
- `Prob > KSa` correspond à la p-value.

Test de Cramer-Von-Mises

- `Summed Deviation from Mean` correspond à la somme des écarts au carré de chaque distribution empirique F_X et F_Y à la distribution empirique F.
- `CM` correspond à la statistique de Cramer.
- `CMa` correspond à la statistique de Cramer asymptotique, $\texttt{CMa} = (m+n)\texttt{CM}$.

Conclusion

Seul le test unilatéral de Student a permis de mettre en évidence une différence de perte de chaleur. Si la normalité des échantillons est satisfaite, le test de Student est plus puissant que les autres tests proposés.

4.5 Problème 4 : Examen

Comparer l'homogénéité des candidats dans les différents jury par le **test du χ^2 d'homogénéité de deux populations**

Calcul de la statistique

- $P_{X,i}$ représente la probabilité de la catégorie ou classe i dans la population X.
- Les classes ou catégories peuvent être des variables sans ordre entre elles (variable discrète nominale) ou ordinale. Par exemple, une variable représentant les jury est une variable nominale et une variable prenant les critères bon, moyen, mauvais est une variable ordinale.
- Sous H_0, les probabilités $P_{X,i}$ et $P_{Y,i}$ sont égales mais inconnues. On les estime par le rapport N_i/N (effectif de la classe i divisé par l'effectif total).
  - effectif espéré dans la classe i de la population X est : $e_{X,i} = N_X \frac{N_i}{N}$
  - effectif espéré dans la classe i de la population Y est : $e_{Y,i} = N_Y \frac{N_i}{N}$
- Statistique :

$$\frac{(N_{X,i} - e_{X,i})^2}{e_{X,i}} \text{ représente } \frac{(\text{effectif observé} - \text{effectif espéré})^2}{\text{effectif espéré}}$$

Exemple

	JURY 1	JURY 2	JURY 3	TOTAUX
Nombre de reçus	50	47	56	153
Nombre de refusés	5	14	8	27
TOTAUX	55	61	64	180

Effectifs espérés :

	JURY 1	JURY 2	JURY 3
Nombre de reçus	46.75	51.85	54.4
Nombre de refusés	8.25	9.15	9.6

$$\chi^2 = \frac{(50-46.75)^2}{46.75} + \frac{(47-51.85)^2}{51.85} + \frac{(56-54.4)^2}{54.4} + \frac{(5-8.25)^2}{8.25} + \frac{(14-9.15)^2}{9.15} + \frac{(8-9.6)^2}{9.6}$$

$$\chi^2 = 4.84$$

$$\chi^2_{0.95,2} = 5.99$$

On ne peut pas rejeter H_0 au seuil de 5 %.

Commentaires sur la sortie Splus

- `X-squared` correspond à la statistique du χ^2.
- `df` correspond au nombre de degré de liberté : $((3-1) \times (2-1)$.

Résumé des commandes Splus utilisées

- `chisq.test` : réalise le test du χ^2 sur des données sous forme de matrice.

Commentaires sur la sortie SAS

On s'intéresse seulement à la ligne `Chi-Square` du tableau `STATISTICS FOR TABLE OF A BY B`.

Résumé des commandes SAS utilisées

- `proc freq` : calcule plusieurs statistiques et notamment la statistique du χ^2.

4.6 Problème 5 : Activité manuelle

Afin de comparer les deux distributions, on effectue un test de Kolmogorov-Smirnov et un test de Cramer-Von-Mises (celui ci est un peu plus puissant que Kolmogorov et demande peu de calculs supplémentaires).

On examine la sortie SAS de Kolmogorov-Smirnov.

La forte déviation est observée au temps 227. La statistique D, écart maximal entre les deux distributions vaut 0.607. D est inférieur à la valeur critique 0.714 lue dans la table au seuil 5 % du test bilatéral. On ne rejette pas l'hypothèse nulle, on ne peut pas conclure à une différence entre les deux populations. La valeur de la p-value nous conduit à la même conclusion.

On examine la sortie SAS de Cramer-Von-Mises.

On compare la valeur de la statistique `CMa` à la valeur 0.461 au seuil de 5 %. On trouve $0.485 \geq 0.461$, par conséquent on rejette l'hypothèse nulle au seuil 5 %.

4.7 Problème 6 : Variété de haricots

Solution

1. Classes phénotypiques attendues si les lois de Mendel sont vraies:

$$(3R + r)(3B + b) = 9BR + 3Br + 3bR + br$$

donc 9/16 individus à graines noires, 3/16 individus à graines rouges et 4/16 individus à graines blanches.

2. Comparaison des effectifs théoriques et observés : calcul : $42.75 = (9/16) \times 76$

phénotypes	noires	rouges	blanches	total
eff. observés	42	19	15	76
eff. théoriques	42,75	14.25	19	76

3. Calcul du khi2

$$\chi^2 = \frac{(42 - 42.75)^2}{42.75} + \frac{(19 - 14.25)^2}{14.25} + \frac{(15 - 19)^2}{19}$$

$$\chi^2 \simeq 2.44$$

4. Lecture du khi2 théorique dans table

$$\chi^2_{5\%,2} = 5.99$$

5. Décision : On ne peut pas rejeter l'hypothèse nulle.