Chapitre II. : Statistique inférentielle

Cours 3 L'estimation

Objectifs :

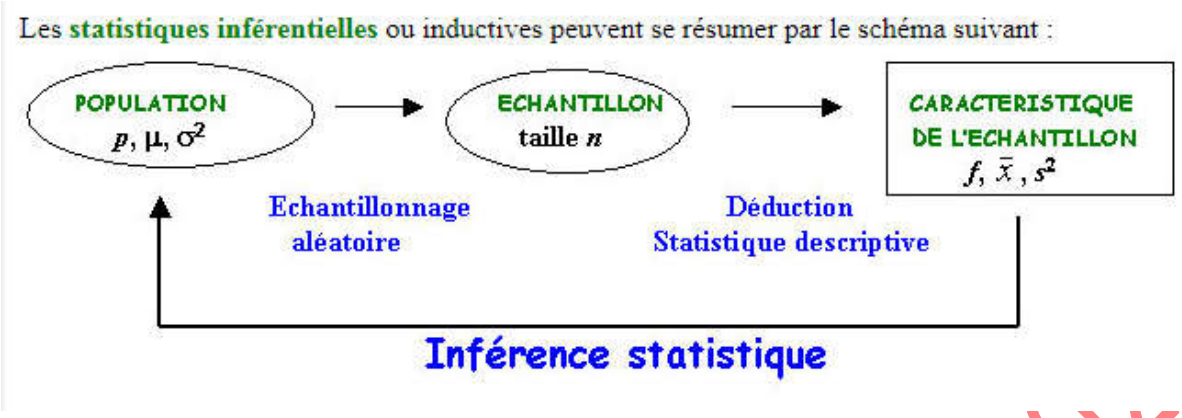
- Donner les applications de l'écart type et l'erreur standard de moyenne;
- Interpréter l'usage et calculer de l'intervalle de confiance pour une moyenne ;
- Interpréter l'usage et calculer de l'intervalle de confiance pour une proportion

Introduction

Défaut des contraintes économiques, humaines et du temps, on ne peut pas étudier une population mère toute entière. Pour contourner ce problème, on extrait un échantillon représentatif (Art d'échantillonnage) et on utilise ses statistiques (descripteurs) pour bien décrire les paramètres de la population mère (moyenne, proportion). Une question que l'on peut se poser est : peut-on généraliser les résultats observés sur l'´echantillon ´étudié à la population parente dont il est issu, et que forcément nous n'avons pas observée ? En d'autres termes, on cherche à inférer au niveau de la population parente des résultats observés, e.g. l'effet d'un facteur (ou une caractéristique particulière de la distribution des observations), sur un ´échantillon représentatif de cette population.

Formalisation

On cherche en fait a` d´écrire les paramètres de la population parente ` a l'aide d'un sous-ensemble de cette population, l'´echantillon. Ces paramètres de population sont estimés à partir des indicateurs descriptifs, qui constituent ce que l'on appelle des statistiques.



La généralisation des résultats sur la population est accompagnée de **doute** qu'on peut le quantifier par une grandeur **probabiliste. Plus** l'échantillon reflète l'image de la population (représentativité) plus les conclusions seront correctes.

On ne raisonnera alors plus en termes d'effectifs ou de fréquences observées, mais en termes probabilistes. Autrement dit, on cherchera `à évaluer la probabilité que la statistique calculée se situe dans un certain intervalle de valeurs de la variable mesurée, ou de manière équivalente au-delà d'une certaine valeur de référence. L'estimateur est donc une nouvelle variable aléatoire construite à partir des données expérimentales et dont la valeur se rapproche du paramètre que l'on cherche à connaître.

Ex.

**On suppose on voudrait connaitre la moyenne du rendement laitier chez FFPN dans la région de Biskra. (Estimation d'une moyenne).

** si on cherche de savoir la proportion de la contamination par PPR des ovins (Estimation d'une proportion).

Estimateur= vraie valeur du paramètre étudie ± erreur standard d'échantillonnage

On quantifie l'erreur d'échantillonnage par un descripteur statistique **ES « erreur standard des estimateurs, ESM ou ESP).**

- Plus la taille de l'échantillon est faible plus ES est grand (moins de précision mauvais estimateur)

- Plus les observations sont plus diversifiées plus ES est grand

Caractéristiques de la distribution des moyennes d'échantillonnage

- Distribution gaussienne ~N(μ, σ) si la variable de la population mère est normalement distribué et ou taille de l'échantillon est assez large (> 30) ; (**Normalité)**
- Moyenne de la distribution (la moyenne des moyennes des échantillons) égale μ la moyenne de la population mère (estimateur non biaisé)
- Ecart type de la distribution des moyennes est mesure de dispersion des moyennes d'échantillonnage (Erreur Standard) des échantillons = $\sigma/\sqrt{n}$

  Pour un seul échantillon, il faudra utiliser l'estimateur de la variance s au lieu de σ

  ES=$s/\sqrt{n-1}$

Analyse des données d'une série de donnée un variée

Cours : Cas un seul groupe - L'intervalle de confiance de la moyenne

Objectif :

Introduction

On peut exploiter les propriétés de la distribution d'échantillonnage de la moyenne pour indiquer comment soit bonne notre estimation. En fait, IC est l'amplitude de valeur qui doit inclure μ de la population avec 95% de certitude.

Si IC est large : m est un mauvais estimateur de μ

IC est étroit : m est un bon estimateur de μ.

Théoriquement, si on répète l'expérience n fois (souvent ce n'est pas le cas) avec la même taille de l'échantillon et on calcule pour chaque échantillon IC. 95% de ces IC peuvent inclure μ la vraie valeur de la moyenne de la population. L'amplitude de IC dépend :

- Seuil de certitude fixé préalablement (souvent 95% en sciences biologiques 1- α)
- Talle de l'échantillon n (+ n grand + de précision + l'étroitesse de IC)
- La variabilité du phénomène étudié (+ de variabilité 6 de précision de l'estimateur + la largeur de IC)

Calcul de IC pour une moyenne μ

Si *σ est connu et n assez large*

IC 95% = m ± 1,96 ESM

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = \left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

NB : 2,58 pour 99% au lieu de 1,96 pour 95%

Si *σ est inconnu et n assez large*

On remplace ***σ par*** son estimateur **s**

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Et on utilise la distribution de Studient « t (0.05, n-1) » au lieu de la distribution normale IC 95% soit:

$$\bar{x} \pm t_{0.05}\frac{s}{\sqrt{n}} = \left(\bar{x} - t_{0.05}\frac{s}{\sqrt{n}}, \bar{x} + t_{0.05}\frac{s}{\sqrt{n}}\right)$$

I. IC pour une proportion π

On suppose que nous sommes intéressés à une proportion **π** des individus qui possèdent un attribut dans une population. Ex. on veut savoir la proportion des ovins atteints par la PRR à Biskra. On sélectionne d'une manière aléatoire un échantillon représentatif d'une taille n de la population ovine dans la région d'étude. On observe la proportion

de la contamination par cette maladie PRR dans l'échantillon soit **p=r/ n est l'estimateur de π**

Caractéristique de la distribution d'échantillonnage d'une proportion

Sont les mêmes pour la distribution de moyenne :

- Réellement la distribution est binomiale β(λ)mais pour n assez large est approximativement normale N (μ, σ)
- La proportion moyenne de la distribution est **π (p un estimateur non biaisé)**
- L' écart type de la distribution des proportion est $\sqrt{\pi(1-\pi)}/n$ **est SEP. SEP** est une mesure de précision de **p** comme une estimateur de **π. A partir d'un échantillon,** on peut l'estimer par :

SEP= $\frac{\sqrt{p(1-p)}}{n} = SEP\% = \sqrt{P\%\ (100-P\%)}/n$

Détermination de IC pour une proportion

IC 95% de π de la population soit :

P± 1,96 SEP= P± 1,96 = $\frac{\sqrt{P(1-P)}}{n}$

Application

Sur un échantillon aléatoire de 115 brebis, le diagnostic sanguin a donnée 36 brebis séropositives de la PRR dans la région de Biskra.

P= 36/115= 0.31, SEP= $\frac{\sqrt{0.31(1-0.31)}}{115}$ =0.043, IC 95%= 0.31±1 .96. 0.043.

On est 95% certain que la proportion de la contamination avec la PRR dans la population ovine à Biskra soit entre 23% et 40%.

Chapitre III. Test d'hypothèse

- **Objectifs : Connaitre l'approche générale d'un test d'hypothèse**

Introduction

Dans la statistique inférentielle, si l'estimation permet la description ; le test d'hypothèse est processus probabiliste de décision sur toute la population à partir d'un (plusieurs) échantillon(s).

Jargon statistique

H0 /H1 : H0 : pas de différence : il n'y a pas d'impact « d'effet » du traitement dans la population .

Ex. on s'intéresse au phénomène biologique : les bœufs sont en danger au pâturage du début de printemps au hypo magnésie.

Un niveau sanguin bas de mg dans l'élevage à l'herbe contre celui à l'étable suggère un risque de la maladie.

H0 : la moyenne de Mg plasmatique dans les 2 groupes est identique.

H1 : les deux moyennes sont différentes.

il est possible d'émettre l'un des deux jugements :

** H0 est vraie, donc H1 est fausse

** H1 est vraie, donc H0 est fausse. On peut symboliser dans le tableau :

	Etat réalisé	
Jugement	H0 réalisé	H1 réalisé
H0	Jugement correct	Jugement faux
H1	Jugement faux	Jugement correct

Parmi les deux hypothèses H0 et H1, il en existe en général une dont le rejet à tort a des conséquences plus fâcheuses que pour l'autre. Il est donc normal de ne pas traiter H0 et H1 de façon symétrique. Ainsi, on peut commettre deux types d'erreur :

** l'erreur de 1ère espèce qui est la probabilité de rejeter H0 alors que H0 est vraie ;

** l'erreur de 2e espèce qui est la probabilité d'accepter H0 alors que H1 est vraie.

Pour relier maintenant le jugement porté à l'observation de la variable X, on opère ainsi :

— on dit que H0 est vraie si la valeur observée de X, soit x, se trouve dans un certain domaine ω, appelé **région d'acceptation** de l'hypothèse H0 ;

— on dit que H1 est vraie si la valeur observée appartient à $\overline{\omega}$, appelé **région critique** ou région de rejet.

Pour choisir le domaine ω, on impose en général deux conditions :

— que la probabilité de commettre l'erreur de première espèce soit égale à un seuil déterminé α choisi a priori aussi faible qu'on le veut (souvent 5% en biologie) ;

— que la probabilité β de commettre l'erreur de deuxième espèce soit minimale. Il importe de noter en effet que la première condition ne suffit pas, sauf cas très particulier, à définir ω de façon unique.

Il est possible maintenant de compléter le tableau précédent en indiquant les règles de jugement et les probabilités pour qu'il soit correct ou faux :

	état réalisé	
décision	H0 est réalisée	H1 est réalisée

H0($X \in \omega$)	jugement correct $(1-\alpha)$	jugement faux (β)
H1($X /\in \omega$)	jugement faux (α)	jugement correct $(1-\beta)$

Un tel mode de raisonnement est appelé test d'hypothèses. Le complément à l'unité de β, soit $(1-\beta)$ est appelé **puissance** du test : un test est d'autant plus puissant, pour un risque de première espèce fixé, que le risque de deuxième espèce est plus **petit**. L'hypothèse H0 sur laquelle sera mené le test est appelée **l'hypothèse nulle**. Autrement dit, Si, par exemple, on compare H0 à Hi , la méthode exposée plus haut permet de trouver une région ω1 telle que le risque de première espèce soit égal à α et que le risque de deuxième espèce β1 soit minimum.

Souvent H1 désigne qu'il y a une différence mais elle veut déterminer dans quelle direction soit-elle (μ1><μ or π1<>π). Donc il s'agit de comparer deux hypothèses de la forme : H0 : $\theta = \theta_0$ et H1 : $\theta > \theta_0$, on est conduit à ce qu'on appelle un test unilatéral à droite, où le risque de première espèce est bloqué à droite

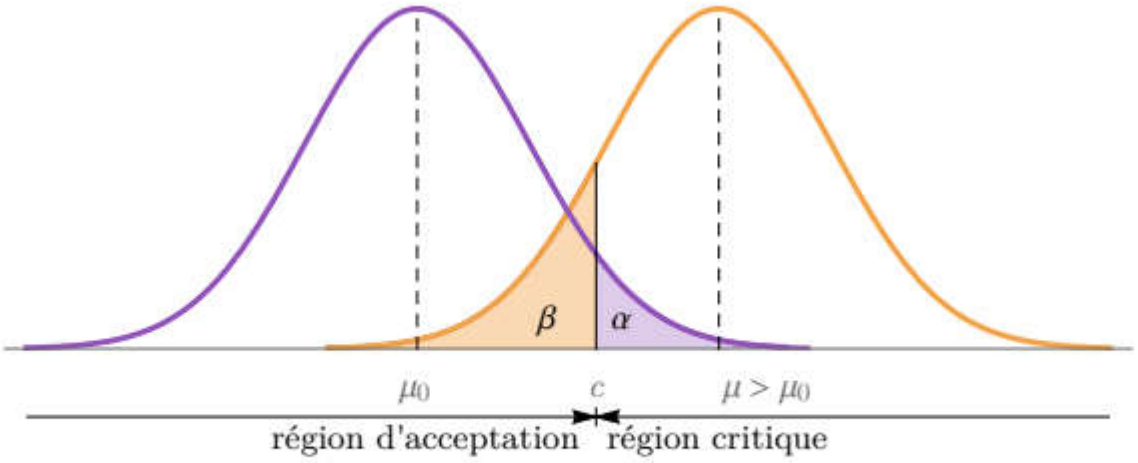


FIGURE 1 – Principe d'un test unilatéral à droite

Le test d'hypothèses de la forme H0 : θ = θ0 et H1 : θ < θ0, conduit à un test unilatéral à gauche. Enfin, dans le cas d'hypothèses de la forme H0 : θ = θ0 et H1 : θ 6= θ0, il apparaît logique de répartir le risque α aux deux extrémités de la distribution. Le test est alors dit symétrique ou bilatéral

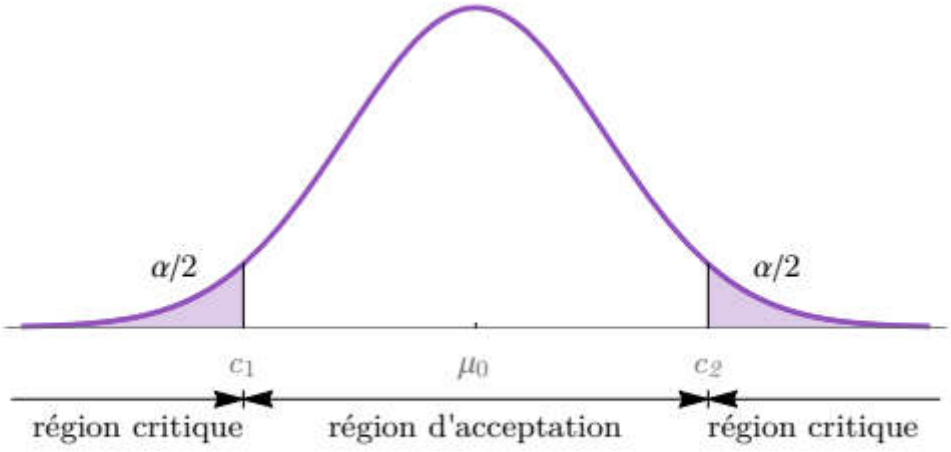


FIGURE 2 – Principe d'un test symétrique ou bilatéral

Démarche générale d'un test

En résumé, la démarche générale d'un test consiste à suivre les étapes suivantes :

1. Choisir et formuler les hypothèses H0 et H1.

2. Déterminer la variable de décision, celle avec laquelle le test sera effectivement mené et qui sera évaluée sur un échantillon.

3. Choisir le risque α et calculer la région d'acceptation ou, préférentiellement, la région critique qui en dépend.

4. Calculer éventuellement la puissance du test 1 – β

5. Calculer la valeur expérimentale de la variable de décision à partir d'un échantillon (test approporié)

6. La comparer à la région critique pour conclure en rejetant ou non l'hypothèse H0.

7. Le cas échéant, calculez l'intervalle de confiance pour l'effet d'intérêt exprimé en fonction de la spécification du paramètre pour H0 vraie.

On gardera à l'esprit, compte tenu de la dissymétrie des deux hypothèses testées, que la conclusion du test est plus forte quand on rejette l'hypothèse H0, le non-rejet ne valant pas vérité. Ainsi, le plus souvent, on choisira les hypothèses de telles sorte que H1 soit l'hypothèse pour laquelle on aimerait conclure.

Test statistique et P-value

A partir de données, on calcule la valeur du test statistique (une expression algébrique particulière à l'hypothèse que nous testons). A chaque valeur du test est attaché une probabilité « p-value » qui décrit la chance d'obtenir l'effet observé sous H0 vraie.

La valeur P nous permet de déterminer si nous disposons de suffisamment de preuves pour rejeter l'hypothèse nulle en faveur de l'hypothèse alternative.

• Si la valeur P est très petite, il est donc peu probable que nous ayons pu obtenir les résultats observés si l'hypothèse nulle était vraie. Nous rejetons donc H0.

• Si la valeur P est très grande, alors il y a de grandes chances que nous ayons pu obtenir les résultats observés si l'hypothèse nulle était vraie et nous ne rejetons pas H0

Prendre une décision en utilisant P-value

**Si les résultats observés ne correspondent pas à ce à quoi nous nous attendions si l'hypothèse nulle était vraie, nous concluons que nous disposons de suffisamment de preuves pour rejeter l'hypothèse nulle. Nous disons que le résultat du test est statistiquement significatif.

** Si, toutefois, les résultats observés correspondent à ce à quoi nous nous attendions si l'hypothèse nulle était vraie, nous ne la rejetons pas. Nous disons que le résultat du test est non significatif.

NB. L'asterisk dans les output de certains logiciels statistiques peuvent distinguent entre :

*** : désigne très hautement significatif ($p < 0.001$)

** : désigne hautement significatif ($0.001 < p < 0.01$)

* : désigne hautement significatif (0.01<p< 0.05)

NS : désigne non significatif (p>= 0.05)