Théorie de l'estimation

ABDELOUAHAB A
Service de Biostatistique

Extraction de *n* échantillons d'une population P

Si l'on extrait plusieurs échantillons représentatifs de taille *n* fixée, les différences observées entre les résultats obtenus sont dues à des *fluctuations d'échantillonnage*. A partir d'un échantillon, on n'a donc pas de certitudes mais des *estimations de paramètres*.

L'estimation d'un paramètre peut être faite
- par un seul nombre: estimation ponctuelle
- par 2 nombres entre lesquels le paramètre peut se trouver: estimation par intervalle

I- Caractère qualitatif :

I-1 :fluctuation d'échantillonnage d'un pourcentage

lorsqu'on évalue dans un ensemble de sujets ou « population » la proportion p de ceux qui présentent un caractère donné, on ne dispose le plus souvent dans la réalité que de groupes extrait de cette population. La proportion qu'ils nous fournissent p0 n 'est pas la proportion exacte, elle s'en écarte plus ou moins selon le hasard de l'échantillonnage

I-1- 1 le pari :

la proportion observée p0 est susceptible de prendre des valeurs échelonnées de 0à1. **q0=1-p0**

les deux bornes de l'intervalle de pari sont données au risque α par :

pour p : $p - \varepsilon\alpha \sqrt{pq/n} \leq p0 \leq p + \varepsilon\alpha \sqrt{pq/n}$

pour q : $1\text{-}pi \geq q0 \geq 1\text{-}ps$

pi borne inférieure de l'intervalle de confiance de p.
ps borne supérieure de l'intervalle de confiance de p

I-1-3-table de l'écart réduit :
donne pour tout écart réduit ε le risque α correspondant
pour ε =1.96 α =0.05
en résumé, si nous parions par exemple que la proportion p0 sur un échantillon de n=100 extrait d'une population ou p=50%
p0 sera comprise entre 40%et 60%(donc i=10) avec un risque de se tromper de 5%.

I-1-4 Condition d'utilisation de la table (validité de l'intervalle)
La table de l'écart réduit ne donne une approximation convenable que pour les grands échantillons : $n \geq 30$
$n.p$, $n.q \geq 5$

I-1-5 puissance et risque d'un pari : si p=20% , n=50

pour α = 0.001 εα =3.29 i= 3.29*0.057=0.19
pour α =0.50 εα =0.67 i= 0.67*0.057=0.04

plus en désir un risque d'erreur faible plus on est obligé de parier sur un intervalle étendu ; il y a **antagonisme** entre le risque et la puissance d'un pari

I-1-6 utilisation des nombres au lieu des pourcentages :

prenons par exemple p=20%de n=50 α =5%
l'intervalle de pari correspondant est 8%-32%
en traduisant les pourcentages en nombre :
0.08*50=4 0.32*50=16
donc on pari au risque de 5% que le nombre de boules noires sera compris entre 4 et 16.

I-2 Estimation d'un pourcentage théorique :
Faire une estimation c'est tenter de définir les paramètres d'une population, à partir des paramètres observes sur un échantillon.

I-2-1 Estimation d'un pourcentage inconnu :
Lorsqu'on observe un pourcentage p0 sur un échantillon, le problème est d'estimer le véritable pourcentage inconnu p de la population d'ou est extrait l'échantillon.
Il nous faut donc estimer un intervalle dans lequel le pourcentage inconnu p de la population a la plus grande probabilité de se trouver : cet intervalle est appelé **intervalle de confiance**

L'observation d'un pourcentage p0 sur un échantillon de n cas permet d'assigner au pourcentage inconnu p l'intervalle de confiance avec le risque α :

$$p0 - \varepsilon\alpha\sqrt{p0q0/n} \leq p \leq p0 + \varepsilon\alpha\sqrt{p0q0/n}$$

Aussi pour q :

$$1\text{-pi} \geq q \geq 1\text{-ps}$$

I-2-2 Conditions d'application (validité de l'intervalle)

Cette formule nécessite que l(effectif de l'échantillon soit suffisamment grand) $n \geq 30$

Si on appelle pi et qi les bornes inférieures et de l'intervalle de confiance de p et de q il faut que : npi , nqi $\geq$ 5.

I-2-3Table de l'intervalle de confiance

Il existe des tables donnant l'intervalle de confiance en fonction du pourcentage observé.