Introduction à la statistique inférentielle : Echantillonnage

Avant de commencer téléchager le fichier LGN.R.

1 Introduction à la statistique inférentielle

Dans des domaines trés variés, comme l'économie, la gestion, la biologie, la médecine, ... on cherche des modèles mathématiques pour analyser, prévoir et décider.

Très souvent, les facteurs intervenant sur le caractère étudié sont plus ou moins aléatoires, et les modèles sont probabilistes. Ainsi, les caractères étudiés sont modélisés par des variables aléatoires. Dans le chapitre précédent, on a déja parlé de données, de descriptions (graphiques et numériques) et d'intuitions sur d'éventuelles modélisations.

Il faut maintenant rentrer dans le domaine plus mathématique des statistiques inférentielles.

Supposons maintenant que l'on souhaite étudier un caractère donné sur une population Ω. On modélise souvent ce caractère par une variable aléatoire X afin de rendre compte de la variabilité entre les individus. Un recensement complet nous permettrait de connaître les valeurs correspondant à tous les individus de cette population. Cela nous permettrait de proposer une modélisation très précise de la distribution de X sur la population toute entière. Cependant, la plupart du temps on n'observe qu'un échantillon de la population totale. En effet, pour des raisons de temps ou de coû t, il est en général impossible de recourir à un recensement. L'exemple du constructeur automobile qui fait des tests de collision pour évaluer le coût d'une collision frontale à 20 kilomètres heures pour le dernier modèle de la marque M illustre bien la situation : il n'accidentera pas toutes ses voitures...

Le but essentiel de la statistique (inférentielle) est d'obtenir des renseignements (ou une modélisation) d'un certain caractère noté X défini sur une population Ω à l'aide de mesures effectuées sur une sous-population $\{\omega_1, \cdots, \omega_n\}$ (appelée échantillon).

Ainsi, la statistique Inférentielle étudie les relations entre échantillons et population parente . On peut distinguer 3 types de démarche :

1. **L'échantillonnage** permet de faire le lien entre la population et l'échantillon. Si on connaissait parfaitement la population, que pourrait-on en déduire sur l'échantillon.
2. **L'estimation** permet de faire le lien entre l'échantillon et la population. Aprés avoir étudié un échantillon, que peut-on dire sur la population.
3. **Les tests** permettent des prises de décision sur la population à partir des valeurs observées sur un échantillon.

2 L'échantillonnage

2.1 Définitions.

Définition 1 *Un échantillon est une fraction d'individus de la population. Le prélèvement des éléments d'un échantillon peut-être effectué :*
avec remise : *l'individu prélevé est immédiatement remis dans la population avant de prélever le suivant.*

Un individu pouvant éventuelleemnt être prélevé plusieurs fois, les tirages sont indépendants et l'échantillon est dit non exhaustif.
sans remise : *l'échantillon est exhaustif, mais les tirages ne sont pas indépendants puisque la composition de la population parente (i.e. où le tirage est effectuée) est modifiée à chaque tirage.*

On supposera dans ce cours que l'échantillonnage est *aléatoire* et *simple* c'est-à-dire que d'une part, tous les individus de la population ont la même probabilité de faire partie de l'échantillon, et d'autre part que les choix successifs des individus composant l'échantillon sont réalisés indépendamment les uns des autres (tirage avec remise, ou sans remise dans une population trés grande).

Par conséquent, on pourra considérer que les variables $X_1, X_2, \cdots, X_n$ (représentant le caractère propre aux individus $1, 2, \cdots, n$ d'un échantillon de taille n) sont indépendantes et de même loi que X. Dans la pratique, voici quelques recommandations pour obtenir un bon échantillon : - si le cardinal de Ω est petit (taux de sondage élevé), les individus doivent être choisis au hasard (avec remise).
- si le cardinal de Ω est suffisamment grand (taux de sondage faible) on peut se contenter d'un tirage sans remise. On considère en effet que la population ne change pas beaucoup quand a retiré quelques individus.
- les individus doivent être indépendants, donc on évite les liens qui pourraient influencer les valeurs du caractère (ne pas interroger mari et femme dans un sondage d'opinion). - les mesures doivent être faites dans les mêmes conditions. Pour que les variables aient toutes la même loi. - on peut améliorer l'échantillon avec la notion de sondage stratifié qui consiste à découper la population en classes plus homogènes, et à étudier chaque classe séparément. - on verra aussi que la taille de l'échantillon influe aussi sur sa qualité. On est souvent limité par des problèmes budgétaires (institut de sondages politiques $n = 1000$) ; organismes d'état $n = 10000$; la plus grosse enquête de l'INSEE "emploi " 150000.

2.2 Distribution d'échantillonnage.

Nous avons vu au cours des TP précédents différentes manières de décrire des caractéristiques importantes de jeux de données au travers des notions de moyenne, variance, médiane, quartiles, ... Lorsque l'on procède par sondage, ces caractéristiques ne sont plus déterministes. Ce sont des variables aléatoires qui prendront des valeurs aléatoires en fonction de l'échantillon considéré.

Une variable calculée sur un échantillon (moyenne, somme, variance, ...) est elle-même une variable aléatoire. En effet, si on prélève plusieurs échantillons de même taille n dans une même population, la variable calculée pour chacun des échantillons prendra une valeur différente d'un échantillon à l'autre en raison du caractère aléatoire des prélèvements.

La distribution de probabilité de cette statistique est appelée distribution d'échantillonnage.

La démarche d'échantillonnage consiste à déduire (entre autres) les distributions des variables aléatoires

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

á partir de ce que l'on connait ou que l'on suppose sur la population entière. On suppose à partir de maintenant que l'échantillonnage est aléatoire et simple donc les variables $X_1, X_2, , X_n$ (mesurées sur les individus $1, 2, \cdots, n$ d'un échantillon de taille n) sont indépendantes et de même loi que X d'espérance m et de variance σ^2.

2.3 Loi des grands nombres et Théorême de la limite centrale

Présentons tout d'abord deux outils majeurs de la statistique inférentielle. La loi des grands nombres assure que lorsque la taille de l'échantillon tend vers l'infini, la moyenne arithmétique sur cet échantillon tend vers l'espérance de la distribution de X. Le Théorème de la limite centrale permet d'aller un peu plus

loin et de montrer que lorsque la aille d'échantillon est grand, la loi de $\overline{X}$ est proche d'une loi normale.

Théorème 1 (Loi des grands nombres.) *Soient $X_1, ..., X_N$ des variables indépendantes, de même loi et intégrables ($E[|X|]$ existe et est fini). Alors*

$$\overline{X} = \frac{X_1 + ... + X_N}{N}$$

converge (en probabilité et presque surement) vers $E[X]$.

EXERCICE :

- Taper une (ou plusieurs) commandes suivantes afin de visualiser la convergence de $\overline{X}$ pour différentes lois.
  x< −rbinom(1000,1,0.5)
  z< −0 ;y< −0
  for(i in 1 :1000) {y[i+1]< −y[i]+x[i+1] ;z[i]< −y[i]/i} ;plot(z,type="l")

  x< −runif(1000,-1,1)
  z< −0 ;y< −0
  for(i in 1 :1000) {y[i+1]< −y[i]+x[i+1] ;z[i]< −y[i]/i} ;plot(z,type="l")

  x< −rcauchy(1000,0,1)
  z< −0 ;y< −0
  for(i in 1 :1000) {y[i+1]< −y[i]+x[i+1] ;z[i]< −y[i]/i} ;plot(z,type="l")

  x< −rexp(1000,2)
  z< −0 ;y< −0
  for(i in 1 :1000) {y[i+1]< −y[i]+x[i+1] ;z[i]< −y[i]/i} ;plot(z,type="l")

- On peut penser que l'on a eu de la chance. On voudrait plutôt voir comment se comportent les moyennes calculéées sur différents échantillons. Utiliser la fonction $L.G.N$ pour visualiser la concentration des moyennes correspondant aux différents échantillons autour de la valeur moyenne.
  $L.G.N("binom", 1, 0.5, mean = 0.5)$
  $L.G.N("unif", -1, 1, mean = 0)$
  $L.G.N("pois", 3, mean = 1/3)$
  $L.G.N("exp", 2, mean = 0.5)$.

Théorème 2 (Théorème de la limite centrale.) *Soient $X_1, ..., X_N$ des variables indépendantes, de même loi et de carré intégrable ($E[X]$ et $V(X)$ existent et sont finis) avec $V(X) > 0$. Posons $\overline{X} = \frac{X_1+...+X_N}{N}$. On montre que*

$$Z := \sqrt{N}\frac{\overline{X} - E[X]}{\sqrt{V(X)}}$$

est asymptotiquement (lorsque N tend vers l 'infini) de loi $N(0, 1)$. On peut donc approcher la loi de $\overline{X}$ par la loi $N(E[X], \sqrt{\frac{V(X)}{N}})$ pour N assez grand ($N > 30$ suffit en général).

EXERCICE

Utiliser la fonction T.C.L pour visualiser comment la loi de la statistique Z se rapproche de la loi $\mathcal{N}(0, 1)$ lorsque la taille de l'échantillon grandit. On compare les distributions empiriques des valeurs de Z obtenues pour différents échantillons et la loi théorique.
T.C.L("norm",1,2,mean=1,sd=2)
T.C.L("binom",1,0.5,mean=0.5,sd=0.5)

T.C.L("unif",-1,1,mean=0,sd=sqrt(1/3))
T.C.L("pois",3,mean=3,sd=sqrt(3))
T.C.L("exp",2,mean=0.5,sd=0.5)

2.3.1 Distribution d'échantillonnage de la moyenne.

On définit la moyenne de la manière suivante :

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

On montre facilement les propriétés suivantes

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = m,$$

$$Var(\overline{X}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} * n * Var(X) = \frac{Var(X)}{n}.$$

Donc $\overline{X}$ vaut m en moyenne (i.e. son espérance vaut m) et sa variance est inversement proportionnelle à n. Comme nous l'avons vu il dans un des TPs précédents, la variance qui s'écrit

$$V(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

représente l'écart carré moyen à l'espérance. Ainsi, plus l'échantillon est grand, plus $\overline{X}$ approche la moyenne m.

Plus précisément,
Si la variables X suit une loi $N(m, \sigma)$, la variable $\overline{X}$ suit la loi $N(m, \frac{\sigma}{\sqrt{n}})$.
Si la variable X suit une loi quelconque, on peut quand même appliquer le théorème de la limite centrale et quand n est "grand", $\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}}$ suit approximativement la loi $N(0, 1)$.

2.3.2 Distribution d'échantillonnage de la variance.

Les calculs sont plus compliqués, mais on retiendra que :

$$E(S^2) = \frac{n-1}{n}\sigma^2$$

Si la variable X suit une loi $N(m, \sigma^2)$, alors

$$\frac{nS^2}{\sigma^2} \text{ suit la loi du khi-deux de paramètre } n-1.$$

On utilisera plus souvent l'estimateur corrigé de la variance $S_{n-1} = \frac{n}{n-1}S^2$ pour lequel on a

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1).$$

Si la variable ne suit pas une loi normale, la distribution de la variance est en général inconnue.

EXERCICE :
A. Echantillon de loi normale

– Générer 10000 échantillons de taille 100 de loi $\mathcal{N}(160, 5)$ avec la commande :

$$x < -matrix(0, 10000, 100); for(p\ in 1 : 10000)\ x[p,] < -rnorm(100, 160, 5)$$

– Représenter la distribution de ces données par un histogramme et tracer sur le même graphique la densité de la loi $\mathcal{N}(150, 5)$.
– Calculer les moyennes sur les différents échantillons avec la commande

$$X < -apply(x, 1, mean)$$

– Représenter sur un même graphique l'histogramme des moyennes et la densité de la loi $\mathcal{N}(160, 5/10)$
– Calculer les variances sur les différents échantillons avec la commande

$$V < -99 * apply(x, 1, var)/25$$

– Représenter sur un même graphique l'histogramme des variances et la densité d'une $\chi^2(99)$.

B. Faire de même pour des échantillons de loi $\mathcal{E}(5)$ et $\mathcal{U}(-1, 1)$. On remarquera que la loi de $\frac{nS^2}{\sigma^2}$ ne suit plus une loi $\chi^2(99)$.

2.3.3 Distribution d'échantillonnage de T.

Dans de nombreux cas on ne connait pas la valeur de σ et on peut légitimement vouloir l'estimer en utilisant S^2. On est alors amené à considérer la variable T

$$T = \frac{\overline{X} - m}{\frac{S}{\sqrt{n}}}$$

On peut montrer que si X suit une loi $N(m, \sigma^2)$, la variable T suit une loi de student de paramètre $n - 1$. On peut montrer également, que si X suit une loi quelconque et que n est grand, T suit approximativement une loi $N(0, 1)$.

EXERCICE :

A. Echantillon de loi normale

– Générer 10000 échantillons de taille 10 de loi $\mathcal{N}(160, 5)$ avec la commande :

$$x < -matrix(0, 10000, 10); for(p\ in 1 : 10000)\ x[p,] < -rnorm(10, 160, 5)$$

– Calculer les moyennes sur les différents échantillons avec la commande

$$X < -apply(x, 1, mean)$$

– Calculer les variances sur les différents échantillons avec la commande

$$V < -apply(x, 1, var)$$

– Caculer les valeurs de T pour les différents échantillons avec la commande

$$T < -sqrt(10) * (X - 160)/sqrt(V)$$

– Représenter sur un même graphique l'histogramme des moyennes et la densité des lois de Student à 9 degrés de liberté et de la loi $\mathcal{N}(0, 1)$

B. Echantillon de taille 30 de loi $\mathcal{E}(5)$. Mêmes questions.

EXERCICE SUPPLEMENTAIRE : Voir sur le fichier Exercice 2.