Estimation et Tests

FX Jollois

TC - 2ème année - 2021/2022

Section 1

Estimation

Introduction

Comment puis-je connaître un indicateur sur la population française ?

- Impossible à réaliser (trop coûteux, trop compliqué à mettre en oeuvre, ...)
- Sélection d'un sous-ensemble de la population, appelé **échantillon**

Comment sélectionner correctement un échantillon ?

- Notion de représentativité
- Méthodes de sondage pour répondre à ce problème

Que fait-on ?

Quand on cherche à analyser un phénomène (biologique, économique, météorologique. . .), on a 2 possibilités

Loi de probabilité connue a priori

On vérifie a posteriori que les observations faites à partir d'un échantillon sont en accord avec elle. On effectue alors un test d'ajustement entre la distribution théorique et la distribution observée

Loi de probabilité inconnue

Mais elle est suggérée par la description de l'échantillon (nature de la variable, forme de la distribution des fréquences, valeurs des paramètres descriptifs). Dans ce cas, il est nécessaire d'estimer les paramètres de la loi de probabilité à partir des paramètres établis sur l'échantillon.

Inférence statistique

Inférence

Opération qui consiste à admettre une proposition en raison de son lien avec une proposition préalable tenue pour vraie.
::: ### Inférence statistique Ensemble de techniques permettant d'induire les caractéristiques d'un groupe général (la population) à partir de celles d'un groupe particulier (l'échantillon), en fournissant une mesure de la certitude de la prédiction (via la probabilité d'erreur) :::

2 problèmes différents

Estimation

Déterminer les **valeurs inconnues** des paramètres de la population à partir des données de l'échantillon. Il est alors nécessaire de déterminer la précision de ces estimations en établissant un *intervalle de confiance* autour des valeurs prédites.

Tests d'hypothèses

A partir d'une hypothèse posée, déterminer les conséquences de cette hypothèse sur la population et/ou l'échantillon, et comparer ces conséquences aux observations faites sur l'échantillon. On conclut **en acceptant ou en rejetant l'hypothèse de travail** à partir de règles de décisions objectives.

Distribution d'échantillonage

Dans un problème d'estimation, il est nécessaire d'étudier la **loi de probabilité** suivie par l'estimateur

Trois concepts importants :

- Paramètres de la **population** (comme la proportion p, la moyenne μ, ou la variance σ^2)
- Paramètres de l'**échantillon** (comme la fréquence f, la moyenne $\bar{x}$, ou la variance s^2)
- Variables aléatoires des paramètres (comme $\bar{X}$, ...)

Estimation

- Problème statistique : estimation d'un paramètre inconnu de la population via un échantillon
- Résumer l'échantillon à une statistique
- Plusieurs catégories de paramètres :
  - Paramètres de position
    - Paramètres de dispersion
    - Paramètres de liaison
- Deux types d'estimation :
  - Estimation ponctuelle
    - Estimation par intervalle

Estimation ponctuelle

- Estimation d'un résultat sur la population
- Unique valeur mesurée dans l'échantillon

Définition

Soit θ un paramètre inconnu intervenant dans la loi de probabilité (connue analytiquement) de la variable aléatoire X. Soient $x_1, x_2, \ldots, x_n$ les n valeurs prises par la v.a. X dans un échantillon de taille n. On appelle **estimateur** de θ, noté T_n la fonction qui fait correspondre aux valeurs x_i de l'échantillon la valeur du paramètre θ. On note la valeur numérique de cette estimation par

$$\hat{\theta} = T_n(x_1, x_2, \ldots, x_n)$$

Exemple d'estimation ponctuelle

- Estimation de la taille moyenne de la population française
- Echantillon : les étudiants de ce cours
- Variable aléatoire suivant une loi normale
- Proposer une estimation de la taille moyenne μ, via l'échantillon x_i ?
  - la moyenne
  - la médiane
  - le mode
  - la taille de l'individu 3
  - ...

Quel estimateur ?

- Meilleur estimateur de la taille moyenne ?
- Définition mathématique impossible de *meilleur*
- Comparer les estimateurs avec certains critères :
  - **Biais** : l'estimation ne doit pas être systématiquement décalée par rapport à la vraie valeur,
  - **Précision** : la variation d'un échantillon à l'autre de l'estimation doit être faible,
  - **Convergence** : lorsque la taille de l'échantillon augmente, l'estimateur converge vers le paramètre inconnu θ,
  - **Complexité** : le calcul de l'estimation ne doit pas nécessiter trop de calculs,
  - **Robustesse** : les perturbations doivent avoir un impact très limité sur l'estimation.

Variable quantitative

Moyenne

Soit X une variable aléatoire d'espérance μ inconnue, la moyenne $\hat{\mu}$ (ou $\bar{x}$) de l'échantillon est un estimateur correct de μ ($E(\hat{\mu}) = \mu$: sans biais et $V(\hat{\mu}) = \frac{\sigma^2}{n} \to 0$: convergent).

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Variance

$\hat{\sigma}^2$ n'est pas un bon estimateur de σ^2 car $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$. Par contre, $\hat{\sigma}^{*2}$ est un estimateur sans biais de σ^2, et convergent. Mais $\hat{\sigma}^*$ n'est pas un estimateur sans biais de σ.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

$$\hat{\sigma}^{*2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Variable quantitative (suite)

Médiane

Valeur pour laquelle 50 % des individus ont une valeur plus grande et 50 % plus petite. Intéressant car insensible aux données aberrantes, contrairement à la moyenne.

$$\hat{m} : p(X < \hat{m}) = 0.5$$

En triant les données x_i par ordre croissant, on obtient la médiane avec

$$\text{Si } n \text{ pair} \qquad \hat{m} = \frac{x_{n/2} + x_{n/2+1}}{2}$$

$$\text{Si } n \text{ impair} \qquad \hat{m} = \frac{x_{(n+1)}}{2}$$

Variable qualitative

Mode

Mesure prise le plus fréquemment.

$$x_{mode} : p(X = x_{mode}) = \max_x p(X = x)$$

Proportion

Soit $\hat{p}$ l'estimation d'une proportion inconnu p et k le nombre d'individus présentant la caractéristique étudiée, la proportion p approxime la vraie valeur de p :

$$\hat{p} = \frac{k}{n}$$

Variable qualitative

Ecart-type d'une proportion

Soit $F_n = \frac{k}{n}$, c'est une v.a. construite par la somme de n v.a. suivant une loi de Bernouilli et de même paramètre p. C'est donc (d'après le TCL) une v.a. dont la loi de probabilité tend vers une loi normale de moyenne p. Son écart-type est estimé par

$$\hat{\sigma_p} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Cette estimation n'est valable que pour les cas où $n > 30$.

Estimation par intervalles

- Intervalle souvent plus intéressant et plus correct que l'affirmation $\hat{\theta} = c$
- Estimation par intervalle de confiance (souvent symétrique)

Définition

Soit X une v.a., θ le paramètre inconnu et $\hat{\theta}$ son estimation sur X, on cherche ainsi $c1$ et $c2$ tel que

$$p(c1 < X < c2 | \theta = \hat{\theta}) = 1 - \alpha$$

Choix de α dépendant du problème posé

- Etude de marché prospective : α élevé (intervalles restreints)
- Etude sur une maladie ou dans une centrale nucléaire : α très faible (intervalles grands)
- Pratique : prendre un risque α égal à 5 %.

Cas de la loi Normale centrée-réduite

Table de valeurs connues pour $P(X < u)$ (fonction de répartition) Table complète

$\rightarrow$ Au croisement, on lit $P(X < 0,31) = 0.6217$

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0.5000	**0.5040**	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0,1	0.5398	**0.5438**	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0,2	0.5793	**0.5832**	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0,3	**0.6179**	**0.6217**	**0.6255**	**0.6293**	**0.6331**	**0.6368**	**0.6406**	**0.6443**	**0.6480**	**0.6517**
0,4	0.6554	**0.6591**	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0,5	0.6915	**0.6950**	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0,6	0.7257	**0.7291**	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0,7	0.7580	**0.7611**	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0,8	0.7881	**0.7910**	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0,9	0.8159	**0.8186**	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	**0.8438**	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1,1	0.8643	**0.8665**	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1,2	0.8849	**0.8869**	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1,3	0.9032	**0.9049**	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1,4	0.9192	**0.9207**	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1,5	0.9332	**0.9345**	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1,6	0.9452	**0.9463**	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1,7	0.9554	**0.9564**	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1,8	0.9641	**0.9649**	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1,9	0.9713	**0.9719**	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

Quelques valeurs à connaître

- $P(X < 1.64) = 0.9494974 =\sim 0.95$
  - Si X suit une loi Normale centrée réduite, il y a 95% de chance que sa valeur soit inférieur à 1.64
- $P(X < 1.96) = 0.9750$
  - Si X suit une loi Normale centrée réduite, il y a 97.5% de chance que sa valeur soit inférieur à 1.96

Et pour les valeurs négatives ? On se base sur la symétrie de la loi Normale.

- $P(X < -1.64) = P(X > 1.64) = 1 - P(X < 1.64) =\sim 0.95$
- $P(X < -1.96) = P(X > 1.96) = 1 - P(X < 1.96) =\sim 0.975$

Conclusion

Pour une variable X suivant une loi Normale centrée-réduite, on a donc 5% de chances que sa valeur soit comprise dans l'intervalle $[-1.96; 1.96]$.

Intervalle de confiance d'une moyenne

Si σ est connu

On a l'intervalle de confiance suivant :

$$\hat{\mu} - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \hat{\mu} + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

où $u_{\alpha/2}$ est la valeur de la table de la loi normale pour laquelle $p(X > u_{\alpha/2}) = \frac{\alpha}{2}$. Puisqu'on choisit souvent $\alpha = 5\%$, on a $u_{\alpha/2} = 1.96$.

Si σ n'est pas connu

On utilise ici l'intervalle de confiance suivant ($t_{\alpha/2}$ est la valeur de la table de la loi de Student pour laquelle $p(X > u_{\alpha/2}) = \frac{\alpha}{2}$) :

$$\hat{m} - t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} < m < \hat{m} + t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$$

Si n est grand, on retrouve la valeur 1.96 pour $t_{\alpha/2}$ avec $\alpha = 5\%$.

Intervalle de confiance d'une proportion

On se base sur le fait que si n est grand, alors la variable aléatoire de la proportion suit approximativement une loi normale. On obtient donc l'intervalle suivant

$$\hat{p} - u_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < X < \hat{p} + u_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

On a toujours $u_{\alpha/2} = 1.96$ pour $\alpha = 5\%$.

Exemple d'estimations

- On a mesuré le niveau de pluie pendant 9 ans, et on a obtenu les valeurs suivantes :
  - $\bar{x} = 610.2222$ et $s = 111.5289$
- X : niveau de pluie de la région
  - Suit une loi Normale $N(\mu, \sigma)$
- Estimation de la moyenne μ par $\bar{x} = 610.2222$
- Estimation par intervalle de confiance à 5% : $[537.3567; 683.0877]$

Section 2

Tests statistiques

Notions générales sur les tests statistiques

- Idée : niveau de pluie en augmentation
- Niveau de pluie suit une loi $N(600, 100)$ (étude précédente)
- Mesure du niveau de pluie pendant 9 ans
  - $\bar{x} = 610.2222$ et $s = 111.5289$
- Que peut-on conclure ?
- Opposer deux hypothèses contradictoires :
  - $[H_0]$ le niveau de pluie n'a pas augmenté, donc $\mu = 600$
  - $[H_1]$ le niveau de pluie a augmenté, donc $\mu > 600$.
- Choix d'une règle de décision

Notions générales sur les tests statistiques

Comment tester ces hypothèses ?

- Intérêt naturel porté à $\hat{\mu}$, moyenne des observations, et donc estimation du niveau de pluie
- Variable considérée comme la **variable de décision**
- Si H_0 vrai, $\hat{\mu}$ suit une loi $N(600, \frac{100}{\sqrt{9}})$

Règle de décision

- Si $\hat{\mu}$ est trop grand, choix de l'hypothèse H_1
  - Donc si $p(\hat{\mu} > k) = 0.05$
  - 5% de chance de se tromper
- Sinon, conservation de H_0

Notions générales sur les tests statistiques

- Test avec $k = 600 + \frac{100}{\sqrt{9}} \times 1.64 = 655$
  - Si $\hat{\mu} > 655$, alors on rejette H_0 pour conserver H_1
  - Si $\hat{\mu} \leq 655$, alors on conserve H_0

Ensemble des évènements

- $\{\hat{\mu} > 655\}$: **région critique** ou région de rejet
- $\{\hat{\mu} \leq 655\}$: **région d'acceptation**

Sur les données

$\hat{\mu} = 610.2$

$\rightarrow$ Conservation de H_0 (pas d'augmentation du niveau de pluie)

Notions générales sur les tests statistiques

- Mais il existe une possibilité de se tromper
  - Croire le chercheur alors qu'il avait tort
  - Ne pas croire ce chercheur alors qu'il avait raison
- Test présentant une forte probabilité d'être inexact
- Si augmentation de la pluie, le niveau suit une loi $N(650, \frac{100}{\sqrt{9}})$

Erreur commise quand $\hat{\mu}$ inférieur à 655

- Probabilité $\beta = p(\hat{\mu} < 655)$
- $u = \frac{\hat{\mu}-650}{100/\sqrt{9}}$ suit une loi $N(0,1)$
- $\beta = p(u < \frac{655-650}{100/\sqrt{9}}) = p(u < 0.15)$
- $\beta = 0.56$, ce qui est effectivement considérable

Notions générales sur les tests statistiques

- Deux probabilités d'erreur
  - α : risque de première espèce
  - β : risque de seconde espèce

	H_0 vraie	H_1 vraie
Choix H_0	$1-\alpha$	β
Choix H_1	α	$1-\beta$

- Dans la pratique, plus d'importance à l'hypothèse nulle
- Calcul de β souvent impossible
- $1-\beta$ est appelé **puissance du test**
- Choix des probabilités d'erreur α de 5%, 1% ou 10%

Notions générales sur les tests statistiques

Pour effectuer un test, voici les étapes à suivre

1. 'Etablir deux hypothèses contradictoires,
2. Déterminer la variable de décision,
3. Calculer la région critique en fonction de α,
4. Calculer si possible la puissance $1 - \beta$,
5. Calculer la valeur expérimentale de la variable de décision,
6. Conclure : rejet ou acceptation de H_0.

Types de test

- Unilatéral : on cherche à tester si une variable a une moyenne supérieure (ou inférieure) à une certaine valeur
  - risque sur un seul côté
- Bilatéral : on cherche à tester si une variable a une moyenne égale à une certaine valeur
  - risque des deux côtés

Exemple de test

- Niveau de pluie suit une loi $N(600, 100)$ (étude précédente)
- Est-ce toujours le même ?
- On a mesuré le niveau de pluie pendant 9 ans, et on a obtenu les valeurs suivantes :
  - $\bar{x} = 610.2222$ et $s = 111.5289$
  - Intervalle de confiance à 5% : $[537.3567; 683.0877]$
- Hypothèses
  - $[H_0]$: le niveau n'a pas changé
  - $[H_1]$: le niveau a changé
- Région critique : en dehors de l'intervalle de confiance

$\rightarrow$ Conservation de H_0 (pas de changement du niveau de pluie)