Statistiques inférentielles

LI323

Hugues Richard
(notes de cours: Pierre-Henri Wuillemin)

Université Pierre et Marie Curie (UPMC)
Laboratoire génomique des microorganismes (LGM)

Introduction

- Soit une population de taille N sur laquelle on observe une propriété, dont on peut calculer moyenne μ et de variance σ^2.
- Soit un individu de cette population, que dire de sa propriété ?

C'est une v.a. X de moyenne m et la variance s^2.

Définition (Échantillon aléatoire)

Un échantillon aléatoire est un prélèvement aléatoire de n individus dans cette population.
$X_1, X_2, \cdots, X_n$ sont alors n v.a. **indépendantes, identiquement distribuées** *(i.i.d.) de moyenne m et la variance s^2.*

Définition (Statistique inférentielle)

La statistique inférentielle a pour but d'induire les caractéristiques inconnues d'une population à partir d'un échantillon issu de cette population. Les caractéristiques de l'échantillon, une fois connues, reflètent avec une certaine marge d'erreur possible celles de la population.

L'inférence statistique est donc un ensemble de méthodes permettant de tirer des conclusions fiables à partir de données d'échantillons statistiques.

Exemples

- Sondages électoraux,
- Tests de fiabilités, de qualité,
- etc.

Exemple 1

Une association de consommateurs veut déterminer si la quantité de vin est bien égale à 75cl dans les bouteilles de Bordeaux. À cette fin, elle examine un échantillon de 100 bouteilles.

Exemple 2

Enquête, sur un échantillon de 400 individus de la population active, pour savoir si le taux de chomage, qui était de 10% le mois dernier, s'est modifié.

Échantillon représentatif

Définition (Échantillon représentatif)

- *Il est nécessaire de s'assurer que l'échantillon est représentatif de la population.*
- *L'échantillonnage aléatoire est le meilleur moyen d'y parvenir.*
- *Un échantillon aléatoire est un échantillon tiré au hasard dans lequel tous les individus ont la même chance de se retrouver.*
- *Dans le cas contraire, l'échantillon est biaisé.*

Échantillon biaisé

Soit une population de 58 étudiants en informatique de taille moyenne $1m78$.

- On choisit un échantillon 'non sexiste' de 5 garçons et 5 filles.
- Moyenne de l'échantillon : $1m74$.
- Le biais : la population comporte 40 garçons pour 18 filles. Donc, chaque garçon avait une probabilité $\frac{5}{40}$ d'être dans l'échantillon, et pour chaque fille : $\frac{5}{18}$.

Statistique

- À partir des (X_i), on peut construire de nouvelle v.a. permettant de synthétiser une information.
- Soit $f(X_1, \cdots, X_n)$ une application définie sur l'échantillon. Nécessairement, c'est également une v.a. !

Définition

On appelle statistique toute application définie uniquement sur l'échantillon.

Statistiques usuelles

- $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$: moyenne d'échantillon
- $S^2 = \displaystyle\sum_{i=1}^{n} \dfrac{\left(X_i - \overline{X}\right)^2}{n-1}$: variance d'échantillon (ou variance corrigée)
- $W = \overline{X} - \mu$ n'est pas une statistique car μ n'est pas observable sur l'échantillon ! !

 $\overline{X}$ et S^2 sont bien des variables aléatoires ! 

Échantillonnage : analyse de $\overline{X}$

- Soit un caractère X de moyenne μ et de variance σ^2.
- Soit un échantillon représentatif $(X_1, \cdots, X_n)$ i.i.d.

$$\mathbb{E}\left(\overline{X}\right) = \mathbb{E}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\mathbb{E}\left(X_1\right) + \cdots + \mathbb{E}\left(X_n\right)}{n} = n\frac{\mathbb{E}\left(X\right)}{n} = \mathbb{E}\left(X\right)$$

$$\mathbb{E}\left(\overline{X}\right) = \mu$$

De même,

$$V\left(\overline{X}\right) = \frac{\sigma^2}{n}$$

Finalement :

Pour n suffisamment grand, $\overline{X}$ suit une loi normale $\mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$.

- n suffisamment grand quand $n \geq 30$.
- Si $X \sim \mathcal{N}\left(\mu; \sigma\right)$ alors ce résultat est vrai même pour n petit.

Théorème Central Limite

Théorème (TCL)

Quelle que soit la distribution d'une variable aléatoire X, la moyenne m d'un échantillon de taille n suit symptotiquement une loi normale $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Théorème (TCL généralisée)

Soit $S_n = \sum_{i=1}^{n} X_i$, *avec les* X_i *v.a. indépendante, à variance finie. Alors*

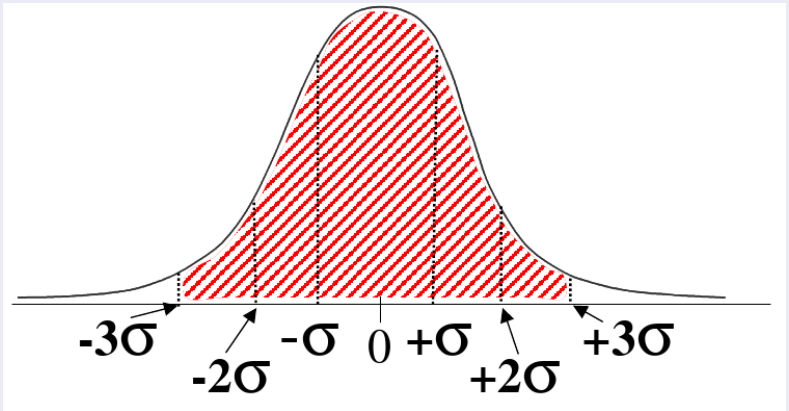
$$S_n \underset{n\to\infty}{\longrightarrow} \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Retour sur la loi normale

Rappel

$$\text{Si } X \sim \mathcal{N}(\mu;\sigma) \text{ alors } \frac{X-\mu}{\sigma} \sim \mathcal{N}(0;1)$$

$\mathcal{N}(\mu;\sigma)$



$P(x \in]\mu-\sigma, \mu+\sigma[$	68,0%
$P(x \in]\mu-2\cdot\sigma, \mu+2\cdot\sigma[$	95,0%
$P(x \in]\mu-3\cdot\sigma, \mu+3\cdot\sigma[$	99.7%

Estimation ponctuelle

Dans l'échantillonnage, on a pu évaluer la distribution de $\overline{X}$ à partir des données de la distribution de la population (μ et σ).
Dans l'estimation, on se pose le problème inverse : étant donnée un échantillon de moyenne m et de variance s^2, que peut-on dire de μ et σ^2 ?

Estimations ponctuelles

- La moyenne m de l'échantillon est la meilleure estimation ponctuelle de μ :

$$\widehat{\mu} = m$$

- La variance corrigée $\frac{n}{n-1} \cdot s^2$ est la meilleure estimation ponctuelle de σ^2 :

$$\widehat{\sigma} = \sqrt{\frac{n}{n-1}} \cdot s$$

Une population peut être décrite par la fréquence $p \geq 1$ d'occurrence de la propriété étudié. On peut alors calculer f la v.a. de cette fréquence dans l'échantillon.

Estimation ponctuelle de p

- la fréquence f est la meilleure estimation ponctuelle de p.

Estimation par intervalle de confiance

Soit P la population, de moyenne μ, de variance σ^2.
Soit un échantillon de taille n de moyenne m et de variance s^2,

μ pour σ, m et $n \geq 30$ sont connues

On sait que (sous les bonnes conditions)

$$\overline{X} \sim \mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right) \text{ et donc } Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0;1)$$

On cherche un intervalle I_α tel que la $P(\mu \in I_\alpha) = 1-\alpha$ avec $\alpha \in [0,1]$.
I_α est l'intervalle de confiance avec le rique α.

Soit t_α, $P(-t_\alpha \leq Z \leq t_\alpha) = 1-\alpha$.

$$\iff P(-t_\alpha \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq t_\alpha) = 1-\alpha$$

$$\iff P(-t_\alpha \cdot \frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq t_\alpha \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$\iff P(\overline{X} - t_\alpha \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + t_\alpha \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

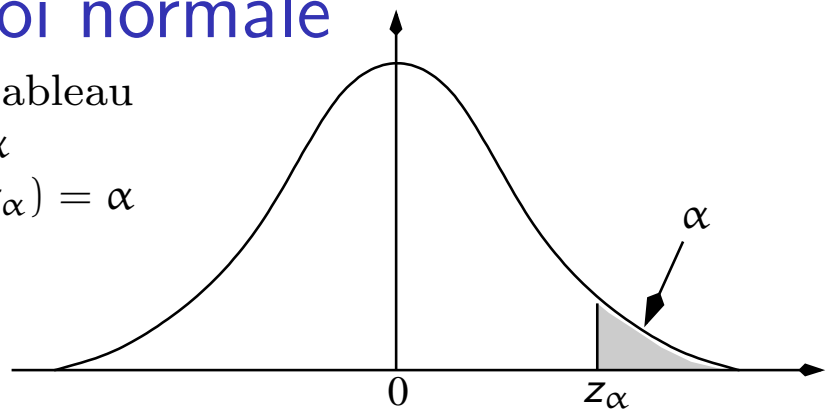
Définition (Intervalle de confiance à risque α)

$$I_\alpha = \left[m - t_\alpha \cdot \frac{\sigma}{\sqrt{n}}, m + t_\alpha \cdot \frac{\sigma}{\sqrt{n}}\right] \textit{ avec } t_\alpha \textit{ tel que } P(-t_\alpha \leq Z \leq t_\alpha) = 1-\alpha.$$

PS- cf. page suivante : $t_\alpha = z_{\frac{\alpha}{2}}$.

Extrait de la table de la loi normale

valeurs dans le tableau ci-dessous : les α tels que $\P(Z > z_\alpha) = \alpha$



z_α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2297	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0859	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0722	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0466	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233

Intervalle de confiance – suite

μ pour σ inconnu, m et $n \geq 30$ connues

En notant $S^2 = \frac{n}{n-1} \cdot s^2$ la variance corrigée de l'échantillon :

$$\overline{X} \sim \mathcal{N}\left(\mu; \frac{S}{\sqrt{n-1}}\right) \text{ et donc } Z = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n-1}}} \sim \mathcal{N}(0;1)$$

p pour f et $n \geq 30$ connues

$$f \sim \mathcal{N}\left(p; \sqrt{\frac{f(1-f)}{n}}\right)$$

σ^2 pour m, s^2 et $n \geq 30$ connues

En notant $S^2 = \frac{n}{n-1} \cdot s^2$ la variance corrigée de l'échantillon :

- $\mathbb{E}\left(S^2\right) = \sigma^2$
- $V(S^2) \to 0$

Calculs d'intervalles

Exemple 1

Une entreprise reçoit un stock important de pièces. L'entreprise n'accepte la livraison que si la proportion de pièces défectueuses est inférieur à 5%. On extrait du stock 200 pièces et on en dénombre 15 défectueuses. L'entreprise doit-elle accepter cette livraison ?

1. $f = \frac{15}{200} = 0.075$.
2. variance $s^2 = p(1-p) = 0.069$
3. Intervalle de confiance à 95% pour p :
   - $t_{5\%} = z_{2.5\%} = 1.96$
   - $\Rightarrow p \in I_{5\%} = \left[0.075 - 1.96 \cdot \sqrt{\frac{0.069}{200}}, 0.075 + 1.96 \cdot \sqrt{\frac{0.069}{200}}\right]$

     $I_{5\%} = [0.038, 0.111]$

Exemple 2

Confitures

Les poids en grammes de 1000 pots de confiture sortis successivement d'une machine à conditionner ont été les suivants (les résultats sont donnés par classes de longueur 2, l'origine de la première étant 2000 et l'extrémité de la dernière 2022) :

classe	1	2	3	4	5	6	7	8	9	10	11
effectif	9	21	58	131	204	213	185	110	50	16	3

- $m = 2010.73$ et $S = 3.58$.
- À 95%, $m - 1.96\frac{S}{\sqrt{n}} \leq \mu \leq m + 1.96\frac{S}{\sqrt{n}}$ donc $2010.51 \leq \mu \leq 2010.95$.