



Statistique:

Apprendre à partir de données.

Répondre à des questions à partir de données.

Méthode statistique: les données sont des réalisations de variables aléatoires.

Modèle probabiliste pour les v.a., information cherchée dans le modèle.

Ex: Sondage (Gauchers dans une grande population?)

Soit p la proportion (dans la grande population)

p inconnue

On tire au hasard n personnes, on compte la proportion de gauchers.

Modèle: $X_i = \begin{cases} 1 & \text{si le } i^{\text{ème}} \text{ individu tiré au hasard est gaucher} \\ 0 & \text{sinon} \end{cases}$

$X_1, \dots, X_n$ indépendantes de loi $\mathcal{B}(p)$

$\bar{X}_n$ proportion de gauchers dans l'échantillon des n individus

Loi de grands nombres

La LGN dit que $\bar{X}_n$ donne l'information sur p .

Déf: Un estimateur est une v.a. qui ne dépend que des réalisations $X_1, \dots, X_n$

Sondage: $\bar{X}_n$ est un estimateur de p

Déf: On dit que T_n est un estimateur sans biais d'un paramètre θ si $E[T_n] = \theta$

Sondage: $\bar{X}_n$ est un estimateur sans biais de p .

Déf: On dit que T_n est un estimateur consistant de θ s'il converge en probabilité vers θ

$$\forall \varepsilon > 0, \lim_{n \rightarrow \infty} P(|T_n - \theta| \geq \varepsilon) = 0$$

Sondage: $\bar{X}_n$ est un estimateur consistant de p

LGN: une moyenne empirique (de v.a.s indépendantes et de même loi)
est un estimateur sans biais et consistant de leur espérance.

Méthode empirique: Utiliser une moyenne empirique pour estimer ce
qui s'écrit comme une espérance

Ex: Si $X_1, \dots, X_n$ indépendantes et de loi $g(p)$

$$E[X_i] = \frac{1}{p}, \quad p = \frac{1}{E[X_i]}$$

$$\text{Avec la méthode empirique: } \hat{p}_n = \frac{1}{n(X_1 + \dots + X_n)}$$

$\frac{1}{\hat{p}_n}$ est un estimateur sans biais et consistant de $\frac{1}{p}$

$\hat{p}_n$ est un estimateur consistant de p (exercice: le démontrer)
mais pas sans biais $E[\hat{p}_n] \neq p$

Test statistique: procédure statistique pour répondre oui/non

Ex: Il y a 50 ans, on sait qu'il y avait 11% de gauchers.
Est-ce qu'il y en a plus maintenant?

Sondage: $\hat{p}_n = \bar{X}_n$ est un estimateur sans biais et consistant de p ,
avec $X_1, \dots, X_n$ indépendantes de loi $\mathcal{B}(p)$

Question: Est-ce que $p = 0.11$ ou bien est-ce que $p > 0.11$?

Données: $x_1, \dots, x_n$: valeurs observées de $X_1, \dots, X_n$

Modèle: loi de $X_1, \dots, X_n$. θ est un paramètre de la loi, $\theta \in \mathcal{H}$

Sondage: Modèle: $(X_1, \dots, X_n) \sim (\mathcal{B}(p))^{\otimes n}$ $\leftarrow$: v.a. indépendantes
et de loi $\mathcal{B}(p)$

Tester: $\mathcal{H}_0: "p = 0.11"$ contre $\mathcal{H}_1: "p > 0.11"$ $(\mathcal{H}_0 = \{0.11\})$
ou $\mathcal{H}_0: "p \leq 0.11"$ contre $\mathcal{H}_1: "p > 0.11"$ $(\mathcal{H}_0 = [0, 0.11] \text{ et } \mathcal{H}_1 =]0.11, 1])$

$$\theta \in \mathcal{H}, \quad (\mathcal{H}_0) \subset (\mathcal{H}), \quad (\mathcal{H}_1) \subset (\mathcal{H}), \quad (\mathcal{H}_0) \cap (\mathcal{H}_1) = \emptyset$$

$$\mathcal{H}_0: "\theta \in (\mathcal{H}_0)" \text{ contre } \mathcal{H}_1: "\theta \in (\mathcal{H}_1)"$$

Idee : on regarde si $\bar{X}_n \leq 0,11$ ou si $\bar{X}_n > 0,11$
on répond H_0 vrai si $\bar{X}_n \leq 0,11$ et H_1 vrai si $\bar{X}_n > 0,11$

Evidemment, il y a des erreurs possibles !

Erreur: décider H_1 alors que H_0 vrai
 H_0 ————— H_1 vrai

Contrôle de la probabilité de la 1^{ère} erreur:
lorsque $p = 0,11$, $P(\bar{X}_n > 0,11) \approx \frac{1}{2}$ qd n grand

Pour avoir P (1^{ère} erreur plus faible), décider H_0 si $\bar{X}_n \leq s$
avec $s > 0,11$, choisir s pour contrôler P (1^{ère} erreur)

Statistique de test: estimateur de θ , v.a. à partir de laquelle on construit
une règle de décision Dans ex: $T_n = \bar{X}_n = \hat{p}_n$
règle de rejet: $T_n > s$

Règle de décision: construire une forme de région de rejet

- * si la région de rejet est réalisée on décide H_1 (on rejette H_0)
- * sinon: on décide H_0 (on accepte H_0)

On calibre le ou les seuils pour contrôler l'erreur
de 1^{ère} espèce.

Véritable Décision \	H_0	H_1
H_0		Erreur de 2 ^{ème} espèce
H_1	Erreur de 1 ^{ère} espèce	

Déf: On dit que le test est de niveau α
si l'erreur de 1^{ère} espèce
est $\leq \alpha$

Si on augmente s , on diminue
l'erreur de 1^{ère} espèce,
mais on augmente l'erreur de
2^{ème} espèce.

Dans ex: On se fixe un niveau
 α , et on choisit pour que
le test soit de niveau α .
Si on choisit α de 5%

On choisit $s \in \gamma$ avec $p = 0.11$,
 $P(\frac{x_1 + \dots + x_n}{n} > s) \leq 0.05$

Soit $P(x_1 + \dots + x_n > ns) \leq 0.05$

$\sim \mathcal{B}(n, 0.11)$ table des lois

binomiales

l'entier k^* tq $P(U > k^*) \leq 0.05$
 $\sim \mathcal{B}(n, 0.11)$

Si $U \sim \mathcal{B}(n, 0.11)$,

et $k < n$

$$P(U > k) = \sum_{l=k+1}^n P(U = l)$$

$$= \sum_{l=k+1}^n \binom{n}{l} (0.11)^l (1-0.11)^{n-l}$$