STT 1000 - STATISTIQUES

ARTHUR CHARPENTIER

Inférence Ponctuelle & Intervalle de Confiance

Estimation ponctuelle : $\widehat{\theta}(\boldsymbol{y}) \in \mathbb{R}$ (ou $\mathbb{R}^d$)
Intervalle de confiance: $IC_{\alpha=}\big[\widehat{a}(\boldsymbol{Y}), \widehat{b}(\boldsymbol{Y})\big] \subset \mathbb{R}$
Ici, on va se poser la question sur la valeur prise par θ.
Example: "*si la fréquentation moyenne par mois pendant la saison d'ouverture dépasse 3500, nous engagerons des moyens financiers pour ouvrir la piste cyclable à l'année*"
y_i : fréquentation moyenne par mois

- si $\mathbb{E}[Y] > 3500$ ouverture
- si $\mathbb{E}[Y] \leq 3500$ fermeture

Hypothèse 1: Y_i suit une loi normale $\mathcal{N}(\mu, \sigma^2)$, et les Y_i sont des variables indépendantes
Hypothèse 2: Y_i suit une loi de Poisson $\mathcal{P}(\lambda)$, et les Y_i sont des variables indépendantes
data : $\boldsymbol{y} = \{3400,\ 4204,\ 4224,\ 4255,\ 4111,\ 3005\}$

Inférence Ponctuelle & Intervalle de Confiance

Hypothèse 1-2: $\widehat{\mu}(\boldsymbol{y}) = \overline{y}$ et $\widehat{\lambda}(\boldsymbol{y}) = \overline{y}$
Pour trancher entre les deux hypothèses :

- si $\overline{y} > y^\star$, $\mathbb{E}[Y] > 3500$ est l'hypothèse la plus crédible
- si $\overline{y} \leq y^\star$, $\mathbb{E}[Y] \leq 3500$ est l'hypothèse la plus crédible

Comment choisir $y^\star$?

Test, rejet & valeur critique

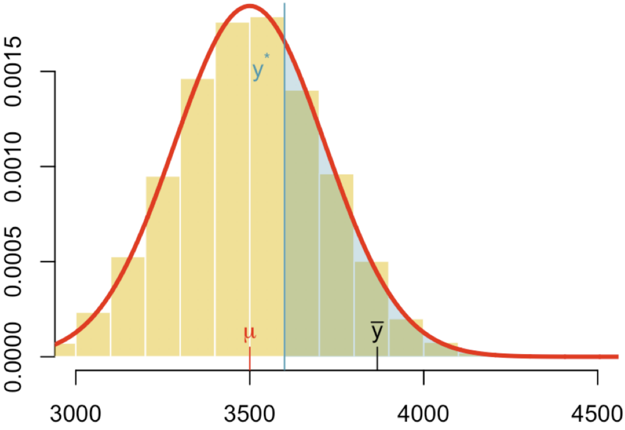
La région $\{\overline{y} \leq y^\star\}$ est la région de rejet,
$y^\star$ est appelée valeur critique.

Tentons quelques scénarios, par simulations (pour comprendre)

Valeur critique et Simulation

Example: $\overline{y} \sim 3866.5$, scenario $\mu = 3500$ ($\sigma = 530$), $y^\star = 3600$

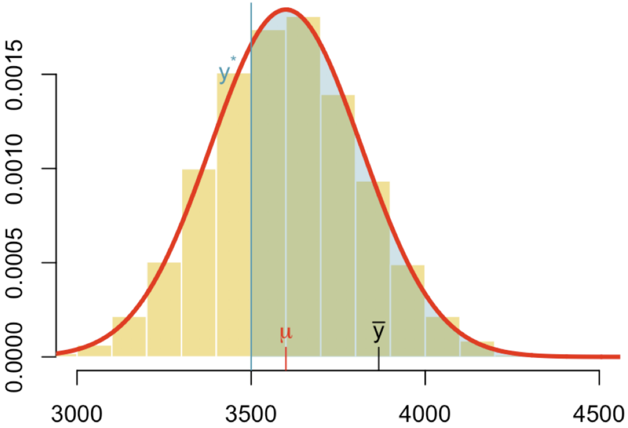
$\mathbb{P}[Z > y^\star] \sim 42.51\%$ avec $\overline{Z} \sim \mathcal{N}(\mu, \dfrac{\sigma}{\sqrt{n}})$



Valeur critique et Simulation

Example: $\overline{y} \sim 3866.5$, scenario $\mu = 3600$ ($\sigma = 530$), $y^\star = 3500$

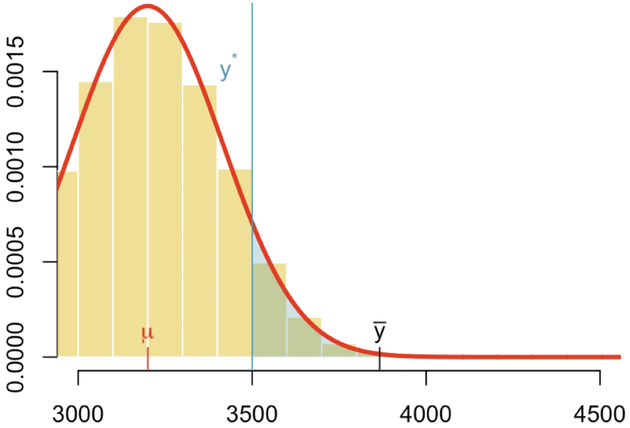
$\mathbb{P}[Z > y^\star] \sim 57.48\%$ avec $\overline{Z} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$



Valeur critique et Simulation

Example: $\overline{y} \sim 3866.5$, scenario $\mu = 3200$ ($\sigma = 530$), $y^\star = 3500$

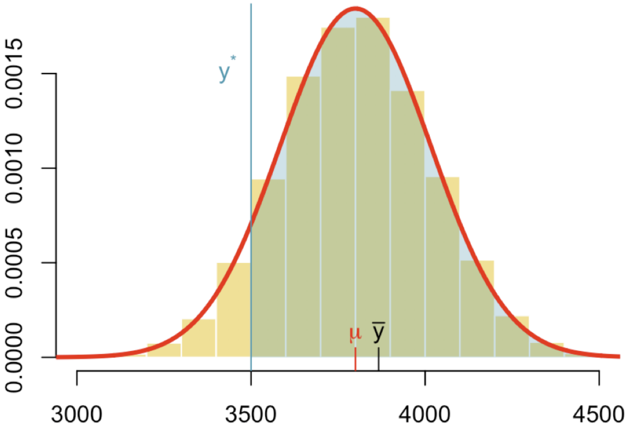
$\mathbb{P}[Z > y^\star] \sim 28.57\%$ avec $\overline{Z} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$



Valeur critique et Simulation

Example: $\overline{y} \sim 3866.5$, scenario $\mu = 3800$ $(\sigma = 530)$, $y^\star = 3500$

$\mathbb{P}[Z > y^\star] \sim 71.43\%$ avec $\overline{Z} \sim \mathcal{N}(\mu, \dfrac{\sigma}{\sqrt{n}})$



Valeur critique et Simulation

Pour une règle de décision fixée ($y^\star$), correspondant à une région de rejet $\{y > y^\star\}$, deux erreurs de décision sont possible, en fonction de la valeur de μ supposée.

- $\mu \notin \{y > y^\star\}$ et $\overline{y} \in \{y > y^\star\}$
- $\mu \in \{y > y^\star\}$ et $\overline{y} \notin \{y > y^\star\}$

On ne peut pas contrôler les deux erreurs en même temps...
On se donne $\alpha \in (0,1)$, et on cherche $y^\star$ tel que

$$\mathbb{P}\left[\overline{Y} > y^\star \middle| H_0 \text{ est vraie}\right] = \alpha$$

Valeur critique et Simulation

Comparer $\widehat{\mu}(\boldsymbol{y})$ à $y^\star$ peut se réécrire

$$\frac{\widehat{\mu}(\boldsymbol{y}) - y_0}{\sigma\sqrt{n}} \leftrightarrow u_\alpha$$

Inférence Ponctuelle & Intervalle de Confiance

- H_0: hypothèse nulle, $\theta \in \Theta_0$
- H_1: hypothèse alternative, $\theta \in \Theta_1$ (souvent $\theta \notin \Theta_0$)

Inférence Ponctuelle & Intervalle de Confiance

Probabilité critique (p-value)

La probabilité critique (p-value) associéee à une statistique de test est la probabilité d'observer des valeurs aussi ou plus extrêmes que la valeur observée dans l'échantillon sachant que H_0 est vraie.

Elle est le plus petit seuil auquel on peut rejeter H0. Pour faire simple, c'est le (plus petit) risque à encourir pour rejeter H0 et accepter H_1.
En fonction de celle-ci, la règle de décision peut se réécrire de la façon suivante : on rejete H_0 (ou on accepte H_1) si p-valeur$< \alpha$.

Inférence Ponctuelle & Intervalle de Confiance

Il est inévitable que des erreurs soient possibles dans la prise de décision suite à un test.
On contrôle ici

$$\mathbb{P}(\text{rejeter } H_0 | H_0 \text{ est vraie}) = \mathbb{P}(\text{accepter } H_1 | H_0 \text{ est vraie})$$

en définissant la région de rejet de telle sorte que cette probabilité α soit petite.
Mais un autre type d'erreur est possible :

$$\mathbb{P}(\text{ne pas rejeter } H_0 | H_0 \text{ est fausse}) = \mathbb{P}(\text{ne pas accepter } H_1 | H_0 \text{ est fausse})$$