

Cours 1

Paramètres, Estimation de paramètres, Erreur type, Intervalle de confiance et test statistiques

Modèle = prédiction de la donnée composé de paramètres (ex : b_0) - caractériser la population

Estimer des paramètres

1. Erreur totale
On fait la différence entre la vraie valeur et celle qu'on estime ($y - \hat{y}$)
prob : valeurs + et - s'annulent
2. Somme des carrés (SC)
Donne la valeur la plus représentative de l'échantillon
la valeur de b_0 qui minimise SC = moyenne
3. Carré moyen (donne la représentativité face à l'échantillon)
Carré moyen est l'équivalent de la variance (s^2)
Écart type = racine de la variance (donc $=s$)
deg de lib = $N - \#$ de paramètres

Erreur-type (SE)

représentativité de notre estimateur face à la population

+ N augmente, + écart-type surestime l'erreur type
Variabilité \uparrow dans la distribution normale
Mais la variabilité dans la distribution d'échantillonnage \downarrow

Intervalle de confiance

Distribution de probabilité

Distribution de probabilité totale
Distribution d'échantillonnage = distribution des probabilités d'obtenir tous les échantillons possibles (Obtenir toutes les moyennes possibles)

Cours 1 (cont)

À quoi ressemble une distribution d'échantillonnage

Théorie des erreurs

1. # de causes = très grand
2. chq cause peut réussir ou échouer
3. probabilité de succès ou échec n'est pas étre (0 ou 1)

À quoi sert SE

Nous savons maintenant que si la distribution d'échantillonnage est distribuée normalement, alors connaître l'erreur type nous permet de cibler un intervalle de valeurs à l'intérieur duquel 95% des moyennes d'échantillons se trouveront
Aussi, si on ne connaît pas la valeur réelle de la moyenne de la population, on sait alors néanmoins que si l'on tirait une infinité d'échantillons, 95 % de ces échantillons nous permettraient de calculer un intervalle de confiance incluant la valeur réelle de la moyenne de la population

Test d'hypothèse

Pour contrer à l'erreur d'échantillonnage on fait stats inférentielles (inductives)

Test stats inférentielles qui sont dites "inductives"

induction vs deduction

Induction : on part des observations pour déterminer c'est quoi la loi générale
Dédution : partir d'une loi générale pour déduire ce que je vais observer

Erreur de type 1

Rejeter H_0 alors qu'elle est vraie.



Cours 1 (cont)

pourquoi onf ait un test bilatéral et non unilatéral si on peut pas s'appuyer sur littérature :

unilatéral a droite et a gauche : l'erreur s'additionne et on fini avec une erreur à 10% (zone de rejet) au lieu de 5% (bilatéral est plus conservateur)

Cours 3

Interv- 1. autour de l'estimateur alors permet de savoir la représentativité de L'erreur type
altes 2. intervalle de confiance autour h0 pour tests statistiques

3.29 = 1. Petit échantillon : ok car très rare (20 particip alors 0.02)
0.001 Mais si grand chantillon va falloir que j'augmente 3.29 selon N on choisit un score Z

Impact Si N diminue , puissance diminue
test T Si on ramène une val extreme à 3.29 , on augmente la puissance stat (car diminue l'écart-type) - On rejette H0 plus facilement

Cours 3 (cont)

Inspection graphique des scores extrêmes :
Histogramme et boîte à moustache

Bas de la boîte: 1er quartile
Haut de la boîte : 3e quartile
Moustache du bas = Minimum (excluant valeurs aberrantes/extrêmes)
Moustache du haut = Maximum (excluant valeurs aberrantes/extrêmes)
Cercle (°) = Donnée aberrante (distance minimum de 1.5 boîtes de la médiane)
Astérisque (*) : val extreme (distance minimum de 3 boîtes de la médiane)

Scores Z dans un distribution normale

Dans une distribution normale, on s'attend à avoir
: A. 0.1% des données dont $z > 3.29$
B. 1.0% des données dont $z > 2.58$
C. 5.0% des données dont $z > 1.96$

Comment gérer les données extrêmes ?

1. Supprimer la donnée 2. Supprimer le participant 3. Remplacer par une valeur qui correspond à 3.29 (score z)

Puissance statistique

Probabilité de rejeter H0 si Ho est **fausse**

Comment les données extrêmes influencent l'erreur type (et tests statistiques)

1. Surestimer l'erreur type
2. Erreur type = bruit , donc diminue la puissance du test statistique (rejet H0 plus difficile)

Plus mon test t est fort...

plus c'est fort , plus jepeux rejeter facilement PLUS T EST PUISSANT PLUS C'EST FACILE DE DÉPASSER LA VALEUR CRITIQUE



Cours 3 (cont)

$2.39 < 2.78 \Rightarrow$ On ne rejette pas H_0 car on aurait une probabilité supérieure à 5% de se tromper si H_0 est vraie.

Données manquantes Éliminer de l'échantillon les sujets ayant des données manquantes. 2 Éliminer d'une analyse les sujets ayant des données manquantes. 3 Remplacer les données manquantes par la moyenne de l'échantillon.

Qu'est-ce qui se passe si je remplace une donnée manquante par la moy

1. ADiminue ecart-type, SE diminue, score du test augmente

Postulats de base

1. Additivité et linéarité

2. Normalité : SI la distribution des fréquences dans l'échantillon est normale, ALORS la somme des carrés de l'erreur (SC) permettra d'estimer les valeurs des paramètres de manière optimale

2.1 : Normalité (Asymétrie et aplatissement)

Asymétrie : Si asymétrie = 0 => parfaitement symétrique \emptyset Si asymétrie < 0 => asymétrie négative (queue plus longue à gauche) \emptyset Si asymétrie > 0 => asymétrie positive (queue plus longue à droite)

Si kurtosis = 0 => aplatissement normal (mésokurtique) Si kurtosis < 0 => aplatissement négatif (platykurtique) => variance élevée Si kurtosis > 0 => aplatissement positif (leptokurtique) => variance faible

Cours 3 (cont)

2.4 : Problèmes: SI la taille de l'échantillon est faible, ALORS le test est rarement assez puissant pour détecter la non-normalité.

Normalité (test de normalité) SI la taille de l'échantillon est très grande, ALORS le test est trop sensible et rejette l'hypothèse nulle (la normalité) trop facilement. \emptyset Or, le théorème central limite suggère de toute façon qu'avec un grand échantillon, la distribution d'échantillonnage, elle, est normale.

H_0 : test est trop sensible et rejette l'hypothèse nulle (la normalité) trop facilement. \emptyset Or, le théorème central limite suggère de toute façon qu'avec un grand échantillon, la distribution d'échantillonnage, elle, est normale.

Asymétrie = 0, Kurtosis = 0 En général, on n'utilise donc pas ces tests (ex. Test de Kolmogorov-Smirnov).

Prob avec hétésceda Biase estimation de l'erreur type

