Statistique
Compétence 2
Ressource R4.04
Traitement numérique des données

David Xu

IUT de Sceaux GEA 2

08 janvier 2024

Organisation du cours

Traitements numériques des données : 2 parties

- Statistique (D. Xu & M. Ghrabli)
- Informatique (E. Bampas)

Statistique

- 2 CM de 2h : maintenant & *mardi 9 janvier (demain) de 10h30 à 12h30*.
- *8 TD de 2h* : vendredi à 11h-13h, 13h-15h ou 15h-17h.
- Un devoir final (1h30) : entre le 1er et le 7 avril.

Évaluations en stat :

- Un DM (non noté) pendant les vacances pour préparer ...
- ... un DS (noté) de 45min après les vacances.
- Le devoir final de 1h30.

Note finale :

- Note du devoir final, si meilleur que la note du devoir de DS post-vacances.
- 2/3 de la note du devoir final + 1/3 de la note du devoir du DS post-vacances.

Ressources

- Le site web de Jérôme Casse :
  https ://sites.google.com/view/jcasse/enseignement/s4
  Taper “Jérôme Casse math” sur google puis [Enseignement > Années précédentes > S4].
- Voir aussi e-campus “Traitements numériques des données”.
- Mon mail : david.xu@universite-paris-saclay.fr

Des questions sur l'organisation du cours ?

Sondage

WOOCLAP

La statistique

À votre avis, pourquoi faire des statistiques ?

WOOCLAP

La statistique

Un outil pour

- chiffrer avec précision,
  Quel est le CA annuel moyen d'une entreprise ?
- aider à prendre des décisions,
  Change-t-on de fournisseurs ?
- voir/quantifier des changements,
  Le marché automobile de 2020 ressemble-t-il à celui de 2000 ?
- établir des corrélations.
  La campagne marketing a-t-elle un impact ?

Plan du cours

Dans ce cours, on va voir 3 outils basiques de la statistique :

1. Intervalles de confiance.
2. Tests de la moyenne.
3. Tests du χ^2 (prononcé khi-2, χ est une lettre grecque).

Questions ?

Intervalles de confiance

Exemple

Test du kilométrage d'une nouvelle gamme de pneus.
Les pneus roulent jusqu'à dégradation.
On note le kilométrage de la dégradation :
81 200 pour le 1er pneu, 84 300 pour le 2ème, 78 100, ...

- Quelle est le kilométrage moyen d'un pneu ?
- De combien d'expérience a-t-on besoin pour avoir une précision de ce kilométrage moyen à 100km près ?
- Quel est la précision si on a que 3 expériences ? Si on a en 10 ? 100 ? ou 1000 ?

But : une réponse du type la moyenne est comprise dans l'intervalle $[79\,224; 81\,642]$ avec une confiance de 95%.

Formalisons

On a une loi inconnue de moyenne m (inconnue) d'écart-type σ (inconnue).
On a $X_1, \ldots, X_n$: n variables aléatoires indépendantes de cette loi.

Deux questions se posent :

1. On veut estimer m, i.e. trouver un nombre proche de m. Facile.
2. On veut savoir à quel point le nombre que l'on trouve est proche de la vraie valeur de m. Intervalle de confiance.
   Si on fait une expérience de plus, l'estimateur trouvé avant va changer.

Flashback S3 : moyenne et écart-type d'une loi discrète

Si la variable aléatoire X est discrète,

- sa loi est la donnée de tous les $P(X = i) = p_i$ (probabilités d'obtenir chaque valeur),
- sa moyenne m est
$$m = E[X] = \sum_i i \times p_i,$$
- sa variance σ^2 est
$$\sigma^2 = \mathrm{Var}(X) = E[X^2] - E[X]^2,$$
- son écart-type σ est
$$\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{E[X^2] - E[X]^2}.$$

Comment estimer m ?

L'expérience des pneus a donné :
81 200, 84 300, 78 100, 79 200, 80 600, 87 300, 81 200,
74 900, 81 700, 80 300, 79 100, 81 300, 76 500, 79 900.

WOOCLAP

Comment estimer m ?

- On fait la moyenne de nos expériences.
- Formellement : $\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1+X_2+\cdots+X_n}{n}$.
- On notera $\hat{m}$ la valeur ainsi obtenue.
- Cette valeur est appelé moyenne empirique ou estimateur de la moyenne.

$$\hat{m} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Pourquoi ça marche ? Probabilités

- Le bon sens.
- Formalisé par la loi des grands nombres en probabilité :

Théorème (Loi des Grands Nombres : LGN ou LLN)

Si $X_1, \ldots, X_n$ sont n variables aléatoires de même loi de moyenne m, alors

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = m.$$

- "Avec une infinité d'expérience, on aurait la valeur exacte."
- LLN : Law of Large Numbers.

La précision de cet estimateur ?

Estimateur de la moyenne

$$\hat{m} = \frac{X_1 + \cdots + X_n}{n}.$$

- Si vous faites l'expérience 1000 fois contre 1 fois, quelle valeur de $\hat{m}$ devrait-être la plus proche de m ?
- À quel point m et $\hat{m}$ sont proches ?
- À quel point est-on sûr que $m - \hat{m}$ est proche de 0 ?

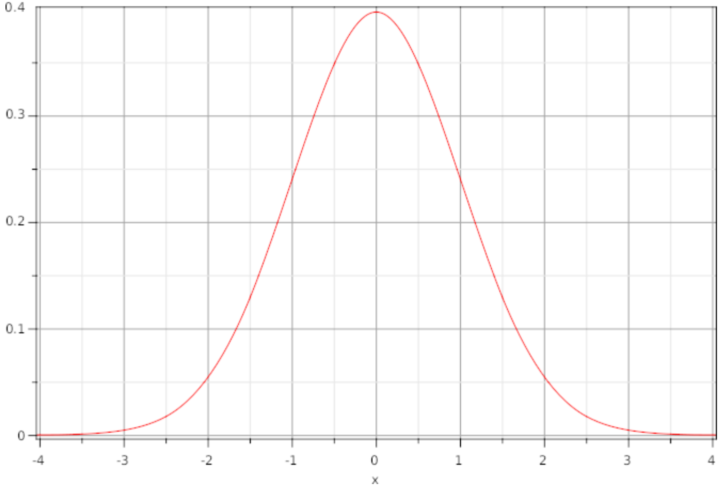
Digression : loi normale et théorème central limite

- Portrait de Carl Friedrich Gauss (1777-1855)
- par Christian Albrecht Jensen (1840)
- Source : wikipedia

La courbe en cloche

- $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.
- Aire sous la courbe est 1
  (via la théorie de l'intégration).
- Comme c'est 1, on a une loi de probabilité :
  la loi normale centrée réduite $\mathcal{N}(0,1)$.
- X variable aléatoire de loi $\mathcal{N}(0,1)$, si
  $P(a \leq X \leq b)$ est l'aire sous la courbe entre a et b.
- $E[X] = 0$ (centrée) et $\text{Var}(X) = 1$ (réduite).

La courbe en cloche



La loi normale de moyenne m et d'écart-type $\sigma > 0$.

- On la note $\mathcal{N}(m, \sigma)$.
- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$

Proposition

Si $X \sim \mathcal{N}(m, \sigma)$, *alors* $\frac{X-m}{\sigma} \sim \mathcal{N}(0, 1)$.

Le théorème central limite

Pourquoi la loi normale ?

Théorème

Soit $X_1, \ldots, X_n$ n variables aléatoires indépendantes de même loi inconnue de moyenne m et d'écart-type σ.
Alors, quand n est grand (ici $n \geq 30$), pour tous nombres réels a, b,

$$P\left(a \leq \frac{\frac{X_1+\cdots+X_n}{n} - m}{\frac{\sigma}{\sqrt{n}}} \leq b\right) \simeq P\left(a \leq \mathcal{N}(0,1) \leq b\right).$$

- Si on somme beaucoup de variables aléatoires indépendantes de même loi, cela ressemble à une loi normale.

Questions ?
Fin de la digression.

- Rappel : un estimateur de la moyenne m est $\hat{m} = \frac{X_1+\cdots+X_n}{n}$.
- Si $n \geq 30$, on en déduit que

$$P\left(a \leq \frac{\hat{m}-m}{\frac{\sigma}{\sqrt{n}}} \leq b\right) \simeq P\left(a \leq \mathcal{N}(0,1) \leq b\right).$$

Donc

$$P\left(a\frac{\sigma}{\sqrt{n}} \leq \underbrace{\hat{m}-m}_{\text{"précision"}} \leq b\frac{\sigma}{\sqrt{n}}\right) \simeq \underbrace{P\left(a \leq \mathcal{N}(0,1) \leq b\right)}_{\text{"confiance"}}.$$

- "précision" ou "fourchette" : la différence entre la vraie valeur m inconnue et la valeur estimée $\hat{m}$.
- "confiance" : la probabilité que la vraie valeur m soit dans l'intervalle

$$\left[\hat{m} - b\frac{\sigma}{\sqrt{n}} ; \hat{m} - a\frac{\sigma}{\sqrt{n}}\right].$$

Pour finir

- Estimer σ.
- Quel valeur pour la confiance $P\,(a \leq \mathcal{N}(0,1) \leq b)$?

Estimation de σ

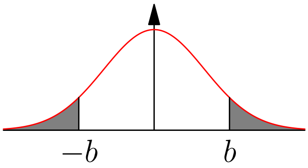
- Rappel : $\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{E[(x-m)^2]}$.
- Le bon sens : $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{m})^2}$.
- Mathématiquement, on lui préfère (car sans biais)

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{m})^2}.$$

La confiance

$P\left(a \leq \mathcal{N}(0,1) \leq b\right)$

- La confiance donnée par l'énoncé, notée α (souvent 95% ou 99%).
- Souvent $a = -b$, i.e. $P\left(-b \leq \mathcal{N}(0,1) \leq b\right)$.
- Trouver b via la table des quantiles.



Conclusion

On en déduit que la valeur moyenne m est compris dans l'intervalle

$$\left[\hat{m} - b\frac{\hat{\sigma}}{\sqrt{n}}; \hat{m} + b\frac{\hat{\sigma}}{\sqrt{n}}\right]$$

avec une confiance α.

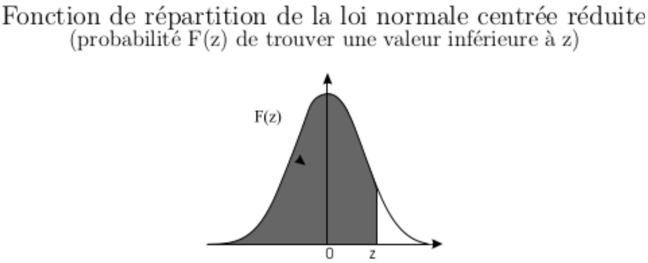
Test du kilométrage d'une nouvelle gamme de pneus.
Les pneus roulent jusqu'à dégradation.
On note le kilométrage de la dégradation.

Au bout de $n = 100$ expériences, on obtient une moyenne empirique $\hat{m} = 80\,425$ et un écart-type empirique $\hat{\sigma} = 2614$.

- À la confiance $\alpha = 95\%$, donner un intervalle de confiance du kilométrage moyen d'un pneu ?

PRENEZ DES NOTES

Lecture de la table



Dans la table on lit l'aire sous la courbe de $-\infty$ à z.

Lecture de la table

AFFICHAGE DE LA TABLE

Les variantes

Fait :

- Loi inconnue et $n \geq 30$.

On va voir :

- Loi Bernoulli ("pourcentage") et $n \geq 30$.
- Loi gaussienne et $n \leq 30$.

Les autres cas :

- Loi inconnue et $n \leq 30$: pas possible.
- Loi connue et $n \leq 30$: compliqué et spécifique à chaque loi.

À cette question vous préferez répondre “oui” ou “non” ?

WOOCLAP

Live-exemple

Donner un intervalle de confiance du pourcentage de “oui” au niveau de confiance 90%.
Probabilité estimée de “oui” : $\hat{p} = \frac{\text{nombre de oui}}{\text{nombre total de réponses}}$.

Avant de commencer

Il faut s'assurer que

- $n \geq 30$ et
- $n\hat{p} \geq 5$ et
- $n(1-\hat{p}) \geq 5$.

Il faut au moins 30 réponses et au moins 5 “oui” et au moins 5 “non”.
Si ce n'est pas le cas, réponse : impossible de faire un intervalle de confiance.

L'estimation de σ est

$$\hat{\sigma} = \sqrt{\hat{p}(1-\hat{p})}.$$

- C'est l'unique changement par rapport à $n \geq 30$ et loi inconnue.

On utilise le Théorème Central Limite (TCL) pour dire que

$$P\left(-z \leq \frac{p - \hat{p}}{\hat{\sigma}/\sqrt{n}} \leq z\right) = P\left(-z \leq \mathcal{N}(0,1) \leq z\right) = \alpha = 0.90.$$

Par lecture dans la table, on trouve $z = 1.645$.
Donc l'intervalle de confiance à 90% est

$$\left[\hat{p} - 1.645 \times \frac{\hat{\sigma}}{\sqrt{n}}\ ;\ \hat{p} + 1.645 \times \frac{\hat{\sigma}}{\sqrt{n}}\right].$$

PRENEZ DES NOTES

Exemple : loi normale et $n \leq 30$

Pour avoir un idée du marché, un agent immobilier a relevé le prix au m^2 d'appartements en vente à Sceaux :

$$9520€ \; ; \; 10180€ \; ; \; 9630€ \; ; \; 9425€ \; ; \; 9500€.$$

Les prix sont supposés de loi normale.

- Donner un intervalle de confiance du prix au m^2 des appartements à Sceaux au niveau 90%.

Deux choses à remarquer

- $n = 5 \leq 30$, mais
- "les prix sont supposés de loi normale".

Donc, on peut faire.

Calcul des estimateurs

- Estimateur de la moyenne $\hat{m}$:

$$\hat{m} = \frac{9520 + 10180 + 9630 + 9425 + 9500}{5} = 9651.$$

- Estimateur de la variance $\hat{\sigma}$:

$$\begin{aligned}\hat{\sigma} &= \sqrt{\frac{(9520 - 9651)^2 + (10180 - 9651)^2 + \cdots + (9500 - 9651)^2}{5 - 1}} \\ &= \sqrt{\frac{131^2 + 529^2 + 21^2 + 226^2 + 151^2}{4}} \\ &= \sqrt{92830} \simeq 305.\end{aligned}$$

Digression : loi de Student

Théorème ("TCL pour petites valeurs dans le cas gaussiens")

Soit X_1, ..., X_n n variables aléatoires indépendantes de loi gaussiennes de moyenne m et de variance σ. Alors

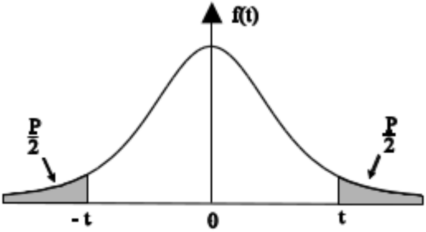
$$P\left(a \leq \frac{m - \hat{m}}{\hat{\sigma}/\sqrt{n}} \leq b\right) = P\left(a \leq \mathcal{T}(n-1) \leq b\right)$$

où $\mathcal{T}(n-1)$ est la loi de Student à $n-1$ degré de liberté.

- $\mathcal{T}(n-1) \simeq \mathcal{N}(0,1)$ si $n \geq 30$.

Table de la loi de Student

Valeurs de **T** ayant la probabilité **P** d'être dépassées en valeur absolue



Application

Par la loi de Student,

$$P\left(-z \leq \frac{m-\hat{m}}{\hat{\sigma}/\sqrt{n}} \leq z\right) = P\left(-z \leq \mathcal{T}(4) \leq z\right) = \alpha = 0.90.$$

Par lecture dans la table de Student, on trouve $z = 2.132$.
Donc, pour notre exemple,

$$P\left(-2.132 \leq \frac{m-9651}{305/\sqrt{5}} \leq 2.132\right) = 0.90.$$

Ainsi $P\left(9651 - 2.132 \times \frac{305}{\sqrt{5}} \leq m \leq 9651 - 2.132 \times \frac{305}{\sqrt{5}}\right) = 0.90.$
L'intervalle de confiance à 90% est $[9360; 9942]$.

Lecture dans la table de Student

AFFICHAGE DE LA TABLE

Des questions ?

À demain !

BONUS : Illustration du TCL