MIASHS, Licence 3
Université Rennes 2
Logiciel R - Travaux dirigés - 2020/2021

Manipulations de données - fonctions usuelles de R

---

1 Données chd

Le fichier chd.csv renseigne, sur $n = 100$ individus, l'âge de l'individu (variable age) et le fait qu'il soit porteur d'une maladie ou non (variable nommée chd, codée 1 pour oui, 0 pour non). On s'intéresse à l'influence de l'âge sur le fait d'être porteur ou non de la maladie.
L'objet de l'exercice est en particulier de former des classes d'âge (sous forme d'intervalles contigus) et de calculer la proportion de malades dans chacune des classes. On cherche aussi à représenter graphiquement la situation et les résultats obtenus.

Plus généralement, ce type de données peut être utilisé pour essayer de voir si la variable âge (quantitative) permet d'expliquer l'appartenance à l'une des modalités d'une variable binaire (ici variable chd de modalités 1 ou 0). C'est un problème typique de la régression dite logistique. Nous ne rentrons pas ici dans la méthode. Cependant, précisons pour information que la dernière question consiste à appliquer les résultats d'un modèle logistique aux données.

1.1 Exploitation des données

1. Importer et calculer le résumé des données.
2. Calculer la moyenne de la variable chd. A quoi correspond cette valeur ?
3. La variable chd est interprétée comme quantitative : ajouter aux données une colonne appelée chd.quali qui contient la conversion de la variable chd en facteur (fonction factor()). Recalculer le résumé des données et retrouver ainsi que 43% des individus sont porteurs de la maladie.
4. Calculer la moyenne d'âge d'un individu malade, d'un individu sain.
5. On se propose de former des classes d'âge puis de calculer la proportion d'individus malades dans chaque classe.
   (a) Utiliser la fonction cut() pour regrouper les valeurs d'âge dans les intervalles suivants :

   $$[20, 29] \quad]29, 34] \quad]34, 39] \quad]39, 44] \quad]44, 49] \quad]49, 45] \quad]54, 59] \quad]59, 69]$$

   Indications :
   - On s'assurera que tous les individus sont bien associés à une classe d'âge.
   - On pourra, comme à la question 2, créer une colonne supplémentaire dans le jeu de données, appelée par exemple age.quali, qui recevra la classe d'appartenance de chaque individu.

   (b) Calculer la proportion d'individus malades dans chaque classe d'âge.

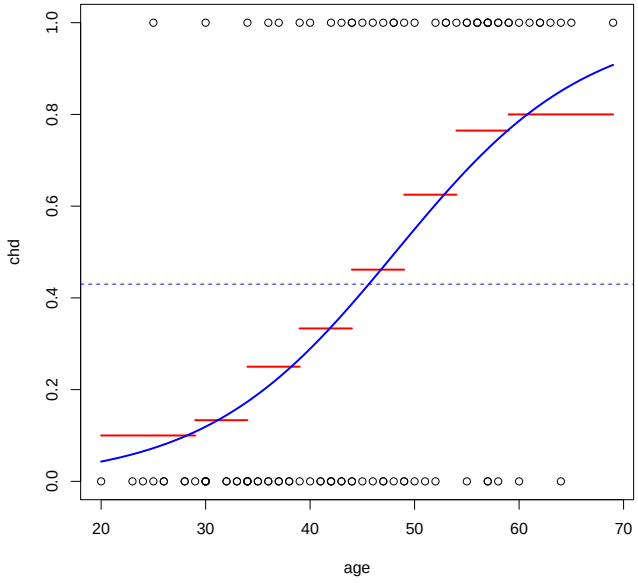
1.2 Graphe

Le graphe à obtenir est représenté en fin d'énoncé

1. Représenter dans un plan les données d'âge en abscisse et celle de la variable quantitative chd en ordonnée, sans oublier de légender les axes.
2. Ajouter par un trait horizontal en pointillés, la proportion des individus malades (sur l'ensemble des données).
3. On souhaite ajouter à ce graphe une représentation de la proportion d'individus malades par classe d'âge. Pour cela, on pourra tracer des segments horizontaux de couleur rouge, de bornes correspondants aux extrémités des classes d'âge, et les placer à la hauteur correspondant à la proportion d'individus malades dans la classe.
4. On considère la fonction $f : \mathbb{R} \to [0, 1]$ telle que

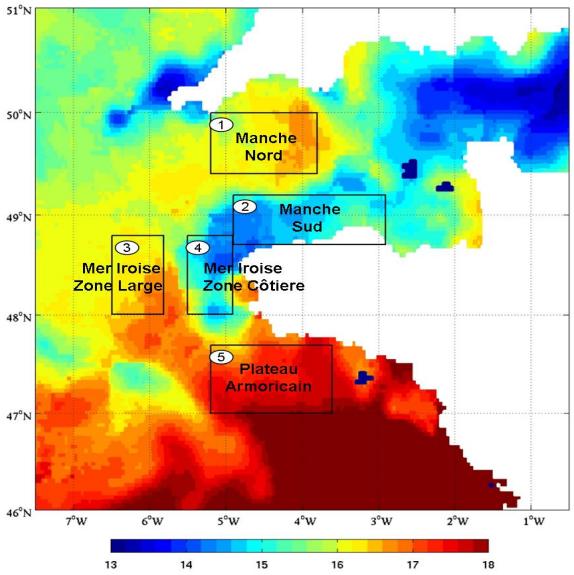
$$f(x) = \frac{\exp(\hat{\beta}_1 + \hat{\beta}_2 x)}{1 + \exp(\hat{\beta}_1 + \hat{\beta}_2 x)}$$

Représenter sur le graphe la courbe de f pour $\hat{\beta}_1 = -5,30$ et $\hat{\beta}_2 = 0,11$. Cette fonction est le résultat de l'application d'un modèle logistique aux données : $f(x)$ donne une estimation de la probabilité qu'un individu soit porteur de la maladie à l'âge x.



2 Températures de surface de l'océan

Vous disposez de 2 jeux de données de températures de surface de l'océan (en degré Celsius) : données observées et données modélisées. Elles se trouvent dans les fichiers observations.csv et modelise.csv respectivement.
Ces fichiers contiennent des données SST pour 5 régions océaniques proches de la Bretagne : dans la Manche au nord (Manche Nord), et au sud (Manche Sud), en Mer d'Iroise au large (Mer Iroise Zone Large), ouprès de la côte (MerIroise Zone Cotiere) et enfin dans l'Atlantique au large des côtes du Morbihan (Plateau Armoricain). Ces régions sont localisées sur la figure ci-dessous :



Données observées Chaque groupe possède les données de Mer Iroise Zone Large et celles d'une autre région. Il y a 5 colonnes :

- année,
- mois (numéroté de 1 à 12),
- jour (numéroté de 1 à 31),
- SST pour Mer Iroise Zone Large
- SST de l'autre région

Ce sont des données satellites à 5km de résolution moyennées spatialement, au pas de temps journalier de 1986 à 2013.

Données modélisées Ce sont des données issues de 3 modèles : CNRM ; MPILR MPIMR à différentes résolutions. Ce sont des valeurs, moyennées sur quelques points qui correspondent à la Mer IroiseZone Large dans chacun des modèles. Les données sont au pas de temps journalier de 1980 à 2005.

2.1 Importation et préparation des données

1. Importer dans R les données. On conservera les colonnes jour, mois et annee au format numérique.
2. Dans les données d'observations, changer le nom des deux dernières colonnes : SST MerIroiseZL en region1 et SST Plateau Armor en region2.
3. Dans les données modélisées, changer le nom des trois dernières colonnes en les nommant : model1, model2 et model3.
4. Ajouter aux deux data frame de données une colonne intitulée date résultant de la chaîne de caractères obtenue en "collant" les informations d'année, de mois, de jour, dans cet ordre.
   NB. La date servira de clé de jointure pour la fusion des tableaux ; il est préférable que le nom de la variable date soit identique dans les deux data frame
5. Dans chaque data frame créer une variable intitulée saison, à 4 modalités,
   - H associée aux mois de décembre, janvier et février
   - P associée aux mois de mars, d'avril et mai
   - E associée aux mois de juin, juillet et août
   - A associée aux mois de septembre, octobre et novembre

2.2 Fusion des tableaux

L'opération de fusion de tableau est toujours une opération délicate : il faut avancer avec prudence et autant que possible mettre en place quelques procédures de vérifications.

Les dates de mesures ne sont pas identiques dans les deux tableaux. Cependant, les deux tableaux possèdent des dates communes. Nous souhaitons créer un tableau qui comporte toutes les colonnes (i.e. les colonnes des deux tableaux) avec comme informations les lignes correspondant aux dates communes de mesures.

1. L'opérateur %in% qui s'applique à deux vecteurs A et B selon la syntaxe A%in%B renvoie un vecteur de booléens, de longueur identique à la longueur du vecteur A, précisant l'appartenance ou non de chaque coordonnée de A à l'ensemble défini par les coordonnées de B.
   Exécuter par exemple :
   ```
   A <- c(1,2,5, 30)
   B <- 1:10
   A %in% B
   B %in% A
   ```
2. Déterminer le nombre de lignes dans chacun des deux tableaux. Utiliser l'opérateur défini en question précédente pour déterminer le nombre de dates communes dans les deux tableaux.
3. Faire fusionner les deux tableaux dans un tableau résultat appelé tab.merge en utlisant date comme clé de jointure. Assurez vous que le nombre de lignes est bien celui attendu.
4. Identifier quelques lignes des deux tableaux initiaux qui doivent être présentes dans le tableau fusionné. En revenant aux données, vérifier, sur quelques individus, que le tableau fusionné a pris les bonnes informations.
5. Certaines colonnes du tableau fusionné sont redondantes : vérifiez-le puis supprimez les redondances.

2.3 Statistiques descriptives

Nous travaillons bien sûr désormais sur le tableau fusionné.

1. Présenter les boxplots des 3 variables modélisées toutes dates confondues.
2. Présenter les boxplots des 3 variables modélisées, saison par saison.
3. Présenter les boxplots des 3 variables modélisées, mois par mois.
4. Calculer les résumés numériques classiques des variables modélisées toutes dates confondues.
5. Calculer les résumés numériques classiques des variables modélisées saison par saison.
6. Calculer les résumés numériques classiques des variables modélisées mois par mois.

3 Traitement de valeurs manquantes : données bebe

Le jeu de données bebe.txt renseigne des variables mesurées dans une maternité, les individus sont ici les naissances (ou les bébés). Certaines données ne sont pas renseignées. Il s'agit dans cet exercice de repérer les individus pour lesquels au moins une variable n'est pas renseignée dans le but d'éliminer ces individus avant analyse.

3.1 Importation, prise en main des données

1. Importer les données.
2. Calculer le résumé des données. Quelles informations donne la fonction summary() quant aux valeurs manquantes ?
3. Paramétrer convenablement l'appel de la fonction mean() de sorte à calculer la taille moyenne d'un bébé.

3.2 Repérage des individus non totalement renseignés

1. Combien y-a-t'il de données manquantes ?
2. En combinant les fonctions is.na(), which() (préciser le paramètre arr.ind) et unique(), déterminer les individus non totalement renseignés.
3. Retrouver le résultat précédent au moyen d'une boucle sur les colonnes du tableau.

3.3 Export des données "nettoyées"

Etant donné le grand nombre d'individus concernés, les enlever serait sans trop doute trop radical. Mais pour l'exercice, nous construisons un jeu de données "nettoyé" et l'exportons.

1. Créer le tableau regroupant les individus totalement renseignés en utilisant les résultats obtenus.
2. Obtenir le même tableau de manière directe en utilisant la fonction na.omit().
3. Exporter le tableau ainsi nettoyé au moyen de la fonction write.table().

Statistique inférentielle

4 Simulations avec R

En statistique inférentielle, on est amené à examiner les propriétés théoriques d'estimateurs (i.e. de variables aléatoires fonction des données) pour juger de leur capacité à bien estimer, dans le cadre de l'estimation paramétrique par exemple, le(s) paramètre(s) inconnu(s) de la loi postulée par le modèle.
Ainsi, si on prouve qu'un estimateur a de bonnes propriétés (consistance, faible biais, faible variance,...), on a une plus grande confiance en l'estimation du paramètre, estimation qui correspond à la seule réalisation de cette variable aléatoire dont on dispose.

Cependant, étudier les propriétés de telles variables n'est pas toujours simple. En outre, dans certains cas, il peut être utile de tenter de vérifier empiriquement ces propriétés avant de se lancer dans leur étude théorique.

Le logiciel R intègre ainsi un générateur de nombres aléatoires. Plusieurs fonctions prédéfinies permettent ainsi de générer des nombres aléatoires selon la loi usuelle souhaitée. Nous les utilisons dans cette section pour illustrer deux résultats fondateurs en probabilité (Loi des grands nombres, Théorème central limite) et revenir sur le sens de quelques éléments classiques en estimation (estimation par intervalles, biais-variance,...).

4.1 Fonctions de base pour générer des nombres aléatoires

1. Commenter les instructions suivantes :

```
#fonction sample()
sample(100)
table(sample(100))
table(sample(100,replace=TRUE))
table(sample(x=1:10, size=100, replace =TRUE))

#fonction rbinom()
rbinom(n=10,size=3,prob=0.5)
rbinom(n=10,size=5,prob=0.2)
rbinom(n=10,size=5,prob=0.9)



#fonction runif()
U <- runif(n=1000,min=0,max=5)
hist(U, freq=FALSE)
```

2. Générer des nombres aléatoires en utilisant les fonctions rnorm(), rpois() puis rexp().
3. Générer $n = 1000$ nombres aléatoires suivant une loi $\mathcal{N}(0, 1)$.
   (a) Représenter un histogramme de la distribution des valeurs.
   (b) Au graphe précédent, ajouter la courbe de densité de la loi $\mathcal{N}(0, 1)$ (fonctions dnorm() et lines()).
   (c) Représenter la fonction de répartition empirique de ces valeurs aléatoires (fonction ecdf()).
   (d) Sur le même graphe, ajouter la courbe de la fonction de répartition de la loi $\mathcal{N}(0, 1)$ (fonction pnorm()).

4.2 Génération d'échantillons

Nous venons de voir comment générer un échantillon (par définition un échantillon est aléatoire) selon une loi particulière. On est souvent intéressé par la moyenne d'un échantillon. En composant la fonction mean() avec les fonctions de génération des nombres aléatoires vues au dessus, cela est très simple. Par exemple :

```
mean(rnorm(10,mean=0,sd=1))
```

On propose dans la suite de générer K échantillons de taille n selon une loi commune et calculer chacune des K moyennes associées, au moyen d'une boucle tout d'abord, puis de manière plus appropriée ensuite.

1. Au moyen d'une boucle
   (a) Initialisation : créer une vecteur M qui recueillera ces moyennes, en l'initialisant de la manière suivante :
   ```
   M <- NULL
   ```
   (b) En vous aidant de l'exemple de boucle élémentaire ci-dessous, construire de manière itérative le vecteur M de longueur K qui contient K moyennes d'échantillons de taille n suivant la loi $\mathcal{N}(0,1)$
   ```
   for(i in 1:5){
     A <- i^2
     print(A)
   }
   ```
2. En combinant les fonctions matrix() et apply()
   (a) Construire une matrice de dimension $K \times n$ qui reçoit $K \times n$ réalisations d'une variable $\mathcal{N}(0,1)$.
   (b) Dans la matrice précédente, chaque ligne représente un échantillon. Appliquer par ligne la fonction mean pour obtenir K moyennes de tels échantillons.

4.3 Loi des grands nombres

On rappelle la **loi faible des grands nombres** :
Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires d'espérance commune $\mathbb{E}[X] = \mu$ et de variance commune $\text{var}(X) = \sigma^2$. Alors

$$\forall \varepsilon > 0, \qquad \lim_{n\to\infty} \mathbb{P}\left(| \frac{X_1 + \ldots + X_n}{n} - \mu| > \varepsilon \right) = 0$$

Autrement-dit, pour un échantillon i.i.d. (c'est le cas courant), la probabilité que la (variable aléatoire) moyenne d'échantillon s'éloigne de plus de ε de l'espérance commune $\mathbb{E}[X]$ converge vers 0 quand $n \to \infty$.
Il ne s'agit pas ici de calculer la probabilité en question, mais d'en obtenir une valeur approchée en comptant la proportion des moyennes d'échantillon. $\bar{x} = \frac{x_1+\ldots+x_n}{n}$ telles que

$$|\bar{x} - \mu| > \varepsilon.$$

On prendra les X_i i.i.d. selon une loi normale $\mathcal{N}(\mu, \sigma^2)$.

1. Proposer une programmation qui permettent de faire varier les constantes (μ, σ, n, ε et K le nombre d'échantillons) au clavier et qui compte, parmi les K échantillons, la proportion de ceux dont les moyennes $\bar{x} = \frac{x_1+\ldots+x_n}{n}$ vérifient
   $$|\bar{x} - \mu| > \varepsilon.$$
2. Faire varier les constantes de sorte à illustrer la loi des grands nombres.
3. Représenter graphiquement la situation.

Remarque Dans cet exercice, nous admettons que la proportion des moyennes d'échantillons telles que $|\bar{x} - \mu| > \varepsilon$ est proche de $\mathbb{P}\left(| \bar{X} - \mu| > \varepsilon\right)$. Nous l'admettons car c'est un résultat conforme à l'intuition. Cependant sa justification relève... de la loi des grands nombres...

4.4 Théorème Central Limite

On rappelle le théorème : Soit (X_n) une suite de variables aléatoires i.i.d. selon une loi commune X d'espérance μ et de variance σ^2. Alors :

$$\frac{1}{\sqrt{n}}\left(\frac{X_1 + \ldots + X_n - n\mu}{\sigma}\right) \to_{\mathcal{L}} \mathcal{N}(0,1). \tag{1}$$

1. Vérifier que la variable aléatoire dans l'équation (1) peut s'écrire $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$.
2. En pratique, pour n assez grand, par quelle loi peut-on approcher la loi de $\bar{X}$?

Illustration avec $X \sim \mathcal{U}_{[a,b]}$

1. Rappeler les expressions de $\mathbb{E}[X]$ et $\text{var}(X)$ pour $X \sim \mathcal{U}_{[a,b]}$.
2. Générer K moyennes d'échantillon $X_1, \ldots, X_n$ avec les X_i i.i.d. selon la loi $\mathcal{U}_{[a,b]}$.
3. Sur un même graphe représenter la distribution de ces K moyennes d'échantillon et la distribution théorique par laquelle on peut l'approcher.

Illustration avec $X \sim \mathcal{B}(p)$

Reprendre la démarche de l'exercice précédent avec les X_i suivant une loi $\mathcal{B}(p)$.

Illustration avec $X \sim \mathcal{N}(\mu, \sigma^2)$

Reprendre le TCL en considérant cette fois que les X_i suivent une loi $\mathcal{N}(\mu, \sigma^2)$. Illustrer.

4.5 Notion d'intervalle de confiance

Données Nile

Les données Nile sont disponibles sous R. Elles donnent la valeur du débit annuel du Nil, entre 1871 et 1970, à Assouan.

1. Quel est le débit moyen, mesuré sur ces 100 années ? Quel est l'écart-type ?
2. En utilisant la fonction t.test(), obtenir un intervalle de confiance pour le débit moyen, au niveau de confiance 95%, au niveau de confiance 99%.
3. Vérifier que les bornes des intervalles de confiance sont données par

$$\bar{x} \pm \frac{\hat{\sigma}}{\sqrt{n}} t_{n-1}(1-\alpha/2).$$

4. Quelles conditions supposent l'emploi de ces formules ?

Données simulées

1. Générer K échantillons de taille n, i.i.d. selon une loi $\mathcal{N}(\mu, \sigma^2)$.
2. Pour chacun d'eux, calculer et stocker les bornes de l'intervalle de confiance pour μ au niveau de confiance 95%.
3. Calculer la proportion de ces intervalles qui contiennent μ.
4. Représenter la situation.

4.6 Quelques tests

Introduction par la simulation

A l'approche d'une élection mettant en concurrence 2 candidats, un sondage a été réalisé sur 1000 personnes : celui-ci donne $48,2\%$ d'intentions de votes pour le candidat A. En notant p_A la proportion d'électeurs à voter pour A dans la population, On se propose d'examiner l'hypothèse

$$\mathcal{H}_0 : p_A = 0.5$$

1. Commenter les instructions suivantes :

```
rbinom(n=1000, size=1, prob = 0.8)
table(rbinom(n=1000, size=1, prob = 0.8))
table(rbinom(n=1000, size=1, prob = 0.8))/1000
table(rbinom(n=1000, size=1, prob = 0.8))[2]/1000
```

2. En vous aidant des instructions précédentes, simuler les résultats d'un sondage effectué sur 1000 personnes où, dans la population des électeurs, l'hypothèse $\mathcal{H}_0 : p_A = 0.5$ est vraie.
3. Au moyen d'une boucle (par exemple), simuler (et conserver) les résultats de 100 sondages effectués sur 1000 personnes où, dans la population des électeurs, l'hypothèse $\mathcal{H}_0 : p_A = 0.5$ est vraie.
4. Evaluer, à l'aide de ces résultats,
$$\mathbb{P}(\hat{p}_A \leq 0.482 | \mathcal{H}_0)$$
5. Comparer ce résultat à la probabilité critique donnée par l'instruction

```
prop.test(x=482,n=1000,p=0.5,alternative = "less")
```

Données Nile

On reprend les données Nile.

1. Tester la normalité des données.
2. On appelle μ l'espérance du débit. Tester, au seuil 95% puis 99%, l'hypothèse $\mathcal{H}_0 : \mu = 950$ contre $\mathcal{H}_1 : \mu \neq 950$.
3. Tester, au seuil 95% puis 99%, l'hypothèse $\mathcal{H}_0 : \mu = 950$ contre $\mathcal{H}_1 : \mu > 950$.
4. Tester, au seuil 95% puis 99%, l'hypothèse $\mathcal{H}_0 : \mu = 950$ contre $\mathcal{H}_1 : \mu < 950$.

Indépendance de deux variables qualitatives

L'instruction suivante permet d'obtenir le tableau croisant la couleur des cheveux et la couleur des yeux pour 592 étudiants :

```
don <- margin.table(HairEyeColor,c(1,2))
```

1. Obtenir les distributions marginales.
2. Obtenir les distributions conditionnelles.
3. Tester l'indépendance des deux variables en présence.
4. Obtenir le tableau des valeurs attendues sous l'hypothèse d'indépendance.
5. La composante residuals de l'objet résultant de la fonction chisq.test() donne une indication sur les croisements contribuant le plus aux écarts à l'indépendance : quel croisement est le plus représenté par rapport à l'indépendance ? Le moins représenté ?

4.7 Biais - Variance

L'absence de biais (ou du moins un biais faible) est une des bonnes propriétés que l'on souhaite avoir en estimation. Un estimateur $\hat{\theta}$ d'un paramètre θ est dit sans biais si $\mathbb{E}[\hat{\theta}] = \theta$. L'espérance mathématique de l'estimateur $\mathbb{E}[\hat{\theta}]$ est la moyenne des estimations, sur tous les échantillons possibles. Ainsi, un estimateur est sans biais lorsque, en moyenne, il tombe sur le paramètre. Un autre propriété souhaitable d'un estimateur est une petite variance. La variance d'un estimateur $\hat{\theta}$ est $\text{var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$. Cette quantité mesure, sur tous les échantillons possibles, la moyenne des carrés des écarts entre les estimations et l'espérance de l'estimateur. Ainsi, lorsqu'un estimateur a une petite variance, c'est qu'il y a peu de variabilité entre les estimations possibles : ainsi l'estimation dont on dispose est proche de celles que l'on aurait avec d'autres jeux de données.
Examiner les propriétés de biais et variance des estimateurs suppose donc de calculer des espérances, espérances dont le calcul théorique n'est pas toujours simple. La simulation permettant de générer un grand nombre d'estimations, elle autorise à visualiser la distribution de ces estimations donc à illustrer les propriétés des estimateurs et calculer de manière approchée des espérances.

deux estimateurs de variance

Pour une suite de variable $X_1, \ldots, X_n$ i.i.d. selon une loi d'espérance commune μ et de variance σ^2. On considère deux estimateurs de la variance commune σ^2 :

$$S_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \text{ et } \tilde{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Comparer les propriétés de biais de ces deux estimateurs de la variance, en prenant la loi commune de votre choix.

Maximum de lois uniformes

Soit $U_1, \ldots, U_n$ des variables aléatoires indépendantes de même loi uniforme sur $[0, \theta]$. On envisage deux estimateurs de la borne supérieure θ de l'intervalle : L'estimateur du maximum de vraisemblance $T_n^{(1)} = 2\bar{X}$ et la plus grande valeur de l'échantillon $T_n^{(2)} = \sup(U_1, \ldots, U_n)$.

1. Générer K échantillons de taille $n = 10$ en donnant une valeur à θ.
2. Visualiser la distribution des K estimations de θ avec chacun des deux estimateurs.
3. Au regard des estimations simulées, comparer les propriétés de biais et de variance des deux estimateurs.
4. L'erreur quadratique moyenne d'un estimateur est une mesure de qualité d'un estimateur qui résulte d'un compromis biais-variance. Elle est définie pour un estimateur $\hat{\theta}$ par

   $$\mathbb{E}[(\hat{\theta} - \theta)^2] = \text{biais}^2(\hat{\theta}) + \text{var}(\hat{\theta})$$

   Calculer pour les estimateurs $T_n^{(1)}$ et $T_n^{(2)}$ une valeur approchée de cette quantité.

5 Régression linéaire simple

5.1 Simulation d'un modèle de régression linéaire simple

On considère le modèle de régression

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

avec les x_i non aléatoires équirépartis sur $[0, 1]$ et les variables aléatoires ε_i i.i.d. selon une loi $\mathcal{N}(0, \sigma^2)$.

Les paramètres (β_0, β_1) et σ^2 sont inconnus en pratique. Pour illustrer on prend

$$(\beta_0, \beta_1) = (1, 2) \qquad \sigma^2 = 0.1$$

1. Simuler des données selon ce modèle.
2. Représenter les points (x_i, Y_i).
3. Sur le même graphe représenter la droite inconnue de paramètres (β_0, β_1).

5.2 Un estimateur : la droite des points extrêmes

On se propose d'ajuster au nuage la droite passant par les points (x_1, y_1) et (x_n, y_n).

1. Calculer les paramètres de la droite ajustée par cette méthode.
2. Représenter la droite sur le graphe précédent.
3. Calculer l'erreur quadratique moyenne d'ajustement

$$S(\tilde{\beta}_0, \tilde{\beta}_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

où $\tilde{y}_i$ est la valeur ajustée en x_i par cette méthode.

5.3 Un autre estimateur

On considère maintenant la méthode consistant à prendre la droite passant par le point moyen $(\bar{x}, \bar{y})$ et ayant comme pente la pente définie par les points extrêmes. Reprendre les questions de l'exercice précédent avec ce nouvel estimateur.

Au regard des données, quel ajustement est préférable (au sens quadratique moyen) ?

5.4 Estimateur des moindres carrés ordinaires

On rappelle que l'estimateur des moindres carrés est défini par

$$\hat{\beta}_1 = \frac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

où $\mathrm{cov}(X, Y)$ et $\mathrm{var}(X)$ désignent respectivement covariance et variance empirique.

Reprendre la démarche proposée dans les exercices précédents avec l'estimateur des moindres carrés ordinaires. Quel ajustement est préférable (au sens quadratique moyen) ?

5.5 Fonction lm()

L'utilisation de la fonction lm() (pour linear model) permet de retrouver les résultats de la section précédente, et de faire bien d'autres choses encore.

1. Lancer la commande

   ```
   lm(Y ~ X)
   ```

   avec Y et X désignant les vecteurs des observations de X et Y. Retrouver ainsi l'estimationdes paramètres par les moindres carrés.
2. Stocker le résultat de la fonction lm() dans un objet puis appliquer à cet objet les fonctions : mode(), names(), summary(), plot().
3. **Utilisation standard de la fonction lm()** :
   — Construire un data frame regroupant les observations de X et Y.
   — Retrouver les résultats précédents en utilisant le paramètre data de la fonction lm().

5.6 Comparaison d'estimateurs en régression linéaire simple

Dans cet exercice, on se propose de comparer les 3 estimateurs vus précédemment de la pente de la droite de régression dans un modèle de régression linéaire simple

1. Les instructions qui suivent permettent de simuler $n = 100$ données (x_i, y_i) selon le modèle $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ avec $\beta_0 = 1$, $\beta_1 = 2$ et les ε_i i.i.d. selon une loi $\mathcal{N}(0, \sigma)$ où $\sigma = 0,5$:

```
beta0 <- 1 ; beta1 <- 2
n<-100
sigma <- 0.5
x<-seq(0,1,length=n)
eps <- rnorm(n,mean=0,sd=sigma)
y <- beta0+beta1*x+eps
```

   (a) Représenter les données.

   (b) Calculer l'estimation de (β_0, β_1) par les 3 méthodes.

   (c) Tracer les droites associées.

2. Au moyen d'une boucle, simuler $K = 1000$ jeux de données suivant le modèle précédent et stocker les valeurs de pente les 3 estimateurs.

   (a) Représenter les distributions de ces pentes.

   (b) Evaluer le biais des estimateurs de pente pour les 3 méthodes.

   (c) Evaluer la variance des estimateurs de pente pour les 3 méthodes.

   (d) Selon vous, quel estimateur est préférable ?

5.7 Tests de nullité d'un coefficient en régression linéaire

Dans un modèle de régression linéaire simple

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

les hypothèses sur $\mathbb{E}[\varepsilon] = 0$ et $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ pour $i \neq j$ permettent de justifier de manière théorique les propriétés d'abscence de biais et de faible variance de l'estimateur des moindres carrés. Ces hypothèses ont été implicitement respectées en générant les réalisations de ε dans le modèle simulé en section précédente et les propriétés en découlant ont ainsi été observées.

L'hypothèse de normalité des résidus $\varepsilon_i = \mathcal{N}(0, \sigma^2)$ qui est traditionnellement fomulée ensuite permet de faire de l'inférence sur les paramètres β_0 et β_1 du modèle : tests, calculs d'intervalles de confiance,...

Nous nous focalisons en particulier sur les tests de nullité des paramètres. Par défaut, la colonne p-value de la sortie fournie par la fonction lm() donne les probabilités critiques associées aux tests

$$\mathcal{H}_0 : \beta_j = 0 \qquad \text{vs} \qquad \mathcal{H}_1 : \beta_j \neq 0 \qquad \text{pour } j = 0, 1$$

1. Quel est l'intérêt de tester l'hypothèse $\beta_1 = 0$ dans un tel modèle ?
2. Simuler un modèle tel qu'à la section précédente. Faire varier n, σ^2, ainsi que la valeur du paramètre β_1 et commenter les probabilités critiques sur le test de nullité de β_1.

5.8 Régression linéaire sur données réelles : données ozone

Les données ozone proviennent d'une étude sur les pics de pollution en Bretagne en 2001. Le receuil de ces données avait pour but d'expliquer la pollution de l'air, en construisant en particulier des modèles de prévisions du pic d'ozone du jour (variable de nom maxO3). Le jeu de données renseigne également plusieurs variables potentiellement explicatives, en particulier des variables météo. On s'intéresse en particulier à l'influence des variables

- T9, T12 et T15 qui donnent les températures du jour à 9h, 12h, 15h ;
- Ne9, Ne12 et Ne15 qui donnent les indices de nébulosité du jour à 9h, 12h, 15h ;
- Vx9, Vx12 et Vx15 qui sont des indicateurs de force du vent du jour à 9h, 12h, 15h

1. Importer et résumer les données. On précise que les identifiants sont les dates de mesures.
2. Calculer la matrice de corrélation entre les variables quantitatives en présence.
3. Proposer une méthodologie pour déterminer la "meilleure" variable explicative quantitative dans un modèle de régression linéaire simple. Mettre en œuvre la méthodologie.

5.9 Une mesure de la qualité de l'ajustement : le R^2

Dans un modèle de régression linéaire, si l'on note Y_i les observations de la variable à expliquer et $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ les valeurs ajustées, le coefficient dit de détermination, noté R^2, est défini par

$$R^2 = \frac{\text{var}(\hat{Y})}{\text{var}(Y)}$$

Ce coefficient, compris entre 0 et 1 mesure la qualité d'ajustement du modèle : plus il est proche de 1 et plus les points du nuage sont proches de la droite ajustée.

On poursuit avec les données ozone.

1. Retrouver les sorties associées au modèle linéaire liant maxO3 et T12.
2. Retrouver dans les sorties la valeur de R^2.
3. La sortie est associée au test

$$\mathcal{H}_0 : R^2 = 0 \qquad \text{vs} \qquad \mathcal{H}_1 : R^2 > 0$$

   Vérifier que la probabilité critique de ce test est la même que celle associée au test de nullité de β_1 (c'est un résultat général en régression linéaire simple).

5.10 Données ozone : influence de T12 selon les conditions de pluie

Le fichier ozone renseigne également la variable pluie, variable qualitative qui indique s'il pleuvait ou pas le jour de la mesure.

1. Représenter dans un même nuage de points les variables maxO3 et T12 en introduisant un code couleur permettant de distinguer les observations de la variable pluie.
2. Ajuster les droites de régression obtenues selon les conditions de pluie.