TP Statistiques 1

Charantonis Anastase & Brunel Nicolas & Julien Floquet

19 fevrier 2018

Rappel: on considere que vous avez suivi l'introduction à R: http://tryr.codeschool.com/ dans son integralité (d'ou le QCM dans 10 minutes)

- Utilitaires et informations importantes:
  nettoyer son espace de travail: rm(list=ls())
  nettoyer sa console: Ctrl +L
  retrouver le répertoire de travail: getwd()
  changer d'emplacement de travail: setwd()
- Biblio pratique:
  http://www3.jouy.inra.fr/miaj/public/formation/initiationRv10.pdf
- On travaillera sous Rmarkdown. Utilisez RStudio, et créez un fichier RMarkdown. Pour les rapports on attendera un pdf et un RMarkdown avec le même nom, du type `TPSTAT1_NOM_PRENOM_GROUPE_NUMERO.format` ou les formats sont Rmd et pdf. Le depot pédagogique serrat ouvert à partir de la semaine prochaine. Nous allons évaluer, pour chaque étudiant, 2/5 des rendus de TP choisis aléatoirement.

Générer des donnÃ©es et les enregistrer

Dans cette partie on va apprendre à générer des echantillons issus d'une loi de probabilités.

Un echantillon d'une loi de probabilité est une suite de réalisations de cette loi. Il est trés utile en statistique de pouvoir générer des variables aléatoires selon diverses lois de probabilité.

R peut le faire pour un grand nombre de lois via les fonctions de la forme r*func*(n,p1,p2,...) où *func* indique la loi de probabilité, n est le nombre de variables à générer et p1, p2, ... sont les paramétres de la loi. Pour ce faire on aura besoin de utiliser `help()` pour les fonctions suivantes:

Lois	Nom sous R
Gaussienne	rnorm(n,mean=0,std=1)
Uniforme	runif(n,min=0,max=1)
Poisson	rpois(n,lambda)
Exponentielle	rexp(n,rate=1)
χ^2	rchisq(n,df)
Binomiale	rbinom(n,size,prob)
Cauchy	rcauchy(n,location=0,scale=1)

Retrouvez ces fonctions dans vos notes de probabilités (ou sur internet), ça va vous être utile.

Pour chaque une de ces fonctions générer une echantillon de 40 données i.i.d. (independantes et identiquement distribuées), associez les à un vecteur inclus dans un `data.frame`, puis utilisez les fonctions `write.csv` et `write.table` pour les enregistrer. Il serrait intelligent de noter les paramétres utilisés (moyenne,std,...) dans le nom de votre variable/fichier enregistré.

Charger des données depuis un fichier txt (texte) et csv (comma separated variables)

Nettoyez votre espace de travail. Utilisez les fonctions `read.csv` et `read.table`, pour charger la distribution Gaussienne que vous avez généré. Que remarquez-vous?

Pensez à utiliser header=TRUE.

Tracer les données

Générez un vecteur qui contient 10 réalisations de la loi normale N(0,1). Tracez les points obtenus en utilisant 'plot', et m?ttant sur l'axe des x un vecteur sequentiel de la taille de votre vecteur.

Que remarquez-vous? (Utiliez la commande 'abline(h=0)')

Tracez également les lignes horizontales 1 et -1. Que remarquez-vous? Combien de points sont en dehors de ces lignes? La même chose avec les lignes horizontales 2 et -2, 3 et -3. Que remarquez vous?

Effectuer la même chose avec des vecteurs contenant 100 et 1000 valeurs. Que remarquez vous?

Chargez le fichier 'distribution_inconue_1_100_realisations.csv' que vous pouvez trouver dans le même emplacement que ce fichier.

Est-ce que vous pouvez conclure quelque chose sur cette distribution, à partir d'une visualisation?

Testez avec d'autres distributions. Que remarquez-vous?

Histogrammes

La visualisation des résultats precedents nous donnent certaines informations sur la distribution dont ils sont issus.

Les histogrammes sont une autre façon d'évaluer visuellement les données d'un echantillon. Ils representent la densité de distribution de valeurs de réalisations de notre echantillon par segements.

Utilisez `help()` pour la fonction `hist()`.

Appliquez la fonction pour l'echantillon de 100 réalisations que vous avez créé, et pour 'distribution_inconue_1_100_realisations.csv'. Que remarquez vous?

Testez les differents paramÃ¨trages de la fonction: breaks et freq.

Effectuez la même chose pour des distributions de Cauchy avec des parametrages differents.

Par ailleurs, regardez les fonctions de type d*func*(n,p1,p2,...). Elles peuvent vous donner la distribution théorique que vous devriez obtenir. Superposez deux plots en utilisant `par(new=TRUE)` puis en plottant la distribution correspondante au histogramme que vous visualisez.

Moments d'ordre

Les moments d'ordre élévé pour une distribution nous donnent des informations liés à la forme des écart à la moyenne. Si on connait notre loi analytiquement, on peut calculer ses moments. Mais quand on a seulement un échantillon i.i.d. d'une loi inconnue, nous devons les estimer.

- Empiriquement:
  Skewness négatif —> plus notre densité est dissymétrique vers la gauche.
  Kurtosis petit —> Plus l'extremité de la densité va tendre rapidement vers 0.

Sous R il existe les fonctions `skewness()` et `kurtosis()`. Calculez les moments des 4 premiers ordres pours les échantillons que vous avez généré et stockez les résultats dans une matrice. Commentez les resultats obtenus et comparez les valeurs théoriques de ces distributions.\

Moment	Ordre	Formule	Estimateur
Moyenne	1	$E[X] = \int_{-\infty}^{\infty} xdF(x)$	$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
Variance	2	$E[X^2] = \int_{-\infty}^{\infty} x^2dF(x)$	$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^{n^2} (x_i - \bar{x})^2$
Skewness	3	$E[X^3] = \int_{-\infty}^{\infty} x^3dF(x)$	$b_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^3}{[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2]^{3/2}}$
Kurtosis	4	$E[X^4] = \int_{-\infty}^{\infty} x^4dF(x)$	$g_2 = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^4}{[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2]^2} - 3$

Quantiles et Boxplot

Les moments (surtout de premier et second ordre) peuvent nous donner beaucoup d'informations sur les lois dont sont issus nos échantillons. Une autre façon de considérer cela correspond à ordonner nos données dans l'échantillon et de les évaluer en estimant quelle quantité de données sont inferieures ou superieures à une valeur.

q-Quantile: si on segmente notre distribution de densité de probabilités en q parts de volume égal, la valeur en dessous de la quelle se situent p/q des données est nommée p-éme quantile. Typiquement on travaille avec des segmentations de notre distribution en quatre ou cent morceaux. Formellement :

Le quantile $x_{\frac{p}{q}}$ d'un variable aléatoire X est défini comme: $P(X \leq x_{\frac{p}{q}}) = \frac{p}{q}$
où de façon equivalente: $P(X \geq x_{\frac{p}{q}}) = 1 - \frac{p}{q}$.

Comme avant, entre connaitre la distribution réelle et essayer de "faire parler les données", il y a une grande différence. On s'appuie sur notre echantillon pour essayer d'avoir plus d'informations sur nos distributions.

- Quantiles spéciaux:
  Q_1: La valeur en dessous de la quelle on a le quart des valeurs de notre echantillon.
  Q_2: La valeur en dessous de la quelle on a la moitié des valeurs de notre echantillion, aussi connue sous le nom de médiane.
  Q_3: La valeur en dessous de la quelle on a les trois-quarts des valeurs de notre échantillon.

Le boxplot nous permet de voir les valeurs entre Q_1 et Q_2, et Q_2 et Q_3, ainsi que la moyenne et l'étendue de $+/-3\sigma$. Toute valeur en dehors de ces $+/-3\sigma$ est marqué avec des points individuels.

Regardez l'aide de la fonction `boxplot()` et appliquez la sur les different ensembles que vous avez généré. Pour le tableau precedent, contenant les moments de ordre 1 à 4, ajoutez 3 colonnes qui contiennent les 3 quantiles.

Interprétation visuelle

Générez 3 ensembles de 100 individus avec la loi de Cauchy avec des paramétrisations differentes. Effectuez toutes les demarches vues dans ce TP. Que remarquez-vous?