M2 recherche: DM2: Lois et tests de Student

Emeline Schmisser, `emeline.schmisser@math.univ-lille1.fr`, bureau 314 (bâtiment M3).

1 Lois du χ^2 et de Student

Exercice 1 (Loi du χ^2)

$$X = \sum_{k=1}^{n} X_k^2 \sim \chi^2(n) \quad \text{si} \quad X_k \, idd \text{ de loi } \mathcal{N}(0,1)$$

1. Construire une fonction `rchi(n,k)` qui simule n variables indépendantes de loi $\chi^2(k)$.
2. (théorique). Quelle est la moyenne d'une loi $\chi^2(k)$? Sa variance ?
   Indication: : Si $X \sim \mathcal{N}(0,1)$, $\mathbb{E}((X^2-1)^2) = 2$.
3. Tracer sur un même graphique les densités du *chi*2 à 1,2,3 et 10 degrés de liberté en utilisant la fonction *dchisq*.
4. Simuler $n = 4000$ variables X_i de loi $\chi^2(40)$. Centrer et réduire ces variables (en utilisant la moyenne et la variance théoriques). Superposer la densité de la loi normale centrée réduite. Conclure.
5. Simuler 1000 variables aléatoires de loi $\chi^2(k)$ et calculer la moyenne et la variance empirique. Faire varier k de 1 à 100 et tracer l'évolution de la moyenne et de la variance en fonction de k. Vérifier que la moyenne et la variance suivent bien les lois prévues.

Exercice 2 (Loi de Student)

$$S(k) \sim \frac{X}{\sqrt{Z}} \quad X \sim \mathcal{N}(0,1) \quad Z \sim \chi^2(k)$$

avec X et Y indépendantes.

1. Construire une fonction `rstu(n,k)` qui simule n variables indépendantes de loi de Student à k degrés de liberté.
2. (théorique). Quelle est la moyenne d'une loi de Student à k degrés de liberté ?
3. La loi de Student est notée `t` dans R. Superposer sur un même graphique les densités de Student pour $k = 1, 2, 3$ et 20.
4. Simuler $n = 40$ variables qui suivent une loi de Student à 3 degrés de liberté et construire l'histogramme. Superposer la densité théorique (utiliser la fonction de R).
5. Simuler 4000 variables de loi χ^2 pour $k = 20$. Superposer la densité de la loi normale centrée réduite et conclure.
6. Simuler 1000 variables aléatoires de loi $\chi^2(k)$ et calculer la moyenne et la variance empirique. Faire varier k de 3 à 100 (on ne part pas de 1 car la moyenne n'est pas définie dans ce cas, ni de 2 car la variance n'est pas définie) et tracer l'évolution de la moyenne et de la variance en fonction de k.
7. Superposer la courbe $k \to k/(k-2)$. Conclure.

2 Tests de Student

On rappelle le principe du test : on a n variables aléatoires indépendantes identiquement distribuées $X_1, X_2 \ldots, X_n$, de variance inconnue et on veut tester si leur moyenne est égale à m. On utilise la statistique

$$\frac{\bar{X}_n - m}{\sqrt{V/n}} \sim T(n-1) \quad \text{où} \quad \bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \quad \text{et} \quad V = \frac{1}{n-1}\sum_{k=1}^{n}(X_i - \bar{X}_n)^2$$

qui suit une loi de Student sous l'hypothèse nulle $\mathbb{E}(X) = m$.
La justification de ce test est la suivante : sous l'hypothèse nulle :

- $\sqrt{n}\bar{X}_n \sim \mathcal{N}(0, \sigma^2)$ où σ^2 est la variance de X.
- $(n-1)V = \sum_{k=1}^{n}(X_i - \bar{X}_n)^2 \sim \sigma^2\chi^2(n-1)$ (la somme des $X_i - \bar{X}_n$ est fixée et vaut 0).
- D'après le théorème de Cochran, ces deux quantités sont indépendantes.
- Donc

$$\frac{\bar{X}_n}{\sqrt{V/n}} \sim \frac{\sigma}{\sigma} T(n-1)$$

Exercice 3 (Test de Student à deux échantillons, taille différente)

1. Télécharger le fichier `salaires.csv` et utiliser la fonction `attach`. Construire ensuite les vecteurs `salh<-Salaire[Sexe==1 & Csp==1]/1000` et `salf<-Salaire[Sexe==2 & Csp==1]/1000`. On divise les salaires par 1000 car R a du mal à gérer les grands entiers.
   Le but est de tester si les femmes de la catégorie socio-professionnelle 1 gagnent moins que les hommes.
   On note n_1 le nombre d'hommes, n_2 le nombre de femmes.
2. Calculer les moyennes $\bar{X}_1$ et $\bar{X}_2$ et les variances S_1 et S_2 en fonction du sexe.
   Les variances sont à peu près égales. On suppose donc (hypothèse nulle) que les salaires des hommes et des femmes suivent une loi normale de même moyenne m, et de même variance σ^2. L'hypothèse alternative est que la moyenne des salaires des hommes $m2$ est plus grande que celle des femmes $m1$. On va pouvoir appliquer un test de Student. Si les variances ne sont pas égales, il existe un autre test, le test de Welch.
3. (préliminaire) Soit $X \sim \mathcal{N}(m_1, \sigma_1)$ et $Y \sim \mathcal{N}(m_2, \sigma_2)$. Quelle est la loi de $X + Y$?
4. Quelle est la loi de $\bar{X}_1 - \bar{X}_2$?
5. Quelle est la loi de $S = \sum_{k=1}^{n_1}(X_i - \bar{X}_1)^2 + \sum_{k=1}^{n_2}(X_i - \bar{X}_2)^2$?
   On admet que les variables $\bar{X}_1 - \bar{X}_2$ et S sont indépendantes.
6. On considère la statistique

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S/(n-2)}}$$

   Quelle est sa loi.
7. Calculer T pour notre exemple (comparaison des salaires des hommes et des femmes pour la catégorie 1.
8. Donner la p-value du test (quelle est la probabilité de trouver ce résultat si l'hypothèse nulle est vérifiée). Conclure.
9. Si on ne suppose plus que la loi des salaires est normale, est-ce que ce test marche encore ?
10. Pourquoi ne peut-on pas appliquer ce test indépendemment de la catégorie professionnelle (pour tous les hommes et toutes les femmes) ?

On va maintenant utiliser la fonction `t.test` de R. Regarder l'aide pour cette commande (et surtout comment l'appeler). Il y a plusieurs options pour cette commande. Les options qui nous interessent ici sont :

- x
- y
- mu
- alternative
- var.equal

On ne touche pas aux autres options.

11. Faire le test de Student pour les hommes et les femmes de catégorie socio-professionnelle 1, en considérant que les variances sont égales. Conclure.
12. Faire le test de Welch (variances non égales) pour l'ensemble des hommes et des femmes. Conclure.