DM2 TISD - M1 Pro
à rendre pour le Vendredi 27 novembre 2009

Les codes sont à envoyer au plus tard le 26 novembre soir à : `chi.tran@univ-lille1.fr`.
Merci de rendre une copie et un code par binôme.
Les tables **SAS** dont vous aurez besoin sont sur ma page web :
`http://math.univ-lille1.fr/~tran/enseignements.html`

Exercice 1 (Travail des femmes)
Nous reprenons les données `travailfemmes.sas7bdat` issues de l'enquête emploi 2005 de l'INSEE. Table et descriptif des variables sont accessibles sur ma page. Nous nous intéressons à l'activité des femmes (ACT6) et souhaitons la croiser avec d'autres variables (âge (AGE), type de ménage (TYPMEN5)) afin de voir si elles sont statistiquement liées.

1. En utilisant la `PROC FREQ` avec les options nécessaires, croiser la variable ACT6 avec la variable TYPMEN5 et réaliser un test du χ^2. Ces variables sont-elles liées ? Quelles sont les couples de modalités qui contribuent le plus à la liaison statistique ?

2. Définir une nouvelle variable AG6 en découpant la variable d'âge, quantitative continue, en 6 tranches ([16-25[, [25-30[, [30-40[, [40-50[, [50-60[, [60-65[) et croiser cette variable avec ACT6. Commenter.

3. Comparer les liaisons ACT6-TYPMEN5 et ACT6-AG6. Commenter.

Exercice 2 (EMV d'une borne (exercice de la fiche de TD n4))
Soient $X_1, \cdots X_n$ des variables aléatoires *iid* de loi uniforme sur l'intervalle $[a, b]$ où $a < b$ sont des paramètres réels inconnus que l'on cherche à estimer. Le code sera fait avec **R**.

1. Ecrire le modèle statistique qu'on étudie. Ce modèle est-il exponentiel ? Ce modèle est-il régulier ?

2. Ecrire la vraisemblance des observations $\ell(X_1, \cdots X_n\,;\, a, b)$.

3. Dessiner la log-vraisemblance renormalisée $(a, b) \mapsto \ln(\ell(X_1, \cdots X_n\,;\, a, b))/n$ pour un échantillon de $n = 100$ variables aléatoires que vous aurez simulé avec $a = 0$ et $b = 1$.

4. Déterminer les estimateurs du maximum de vraisemblance $(\widehat{a}_n, \widehat{b}_n)$ (graphiquement puis théoriquement).

5. Ces estimateurs sont-ils biaisés ? (répondre sans calcul).

6. Ecrire une fonction qui trace la courbe de la fonction aléatoire $n \mapsto \sqrt{n}(\widehat{a}_n - a)$ et qui prend en argument la valeur a, les valeurs minimales et maximales de n, ainsi que le pas de la sequence utilisée pour determiner les abscisses n dans le tracé de la courbe.

7. Utiliser cette fonction pour $a = 0$ et $n \in \{10, \ldots, 10\,000\}$.

8. Faire de même avec les fonctions $n \mapsto n^2(\widehat{a}_n - a)$ et $n \mapsto n(\widehat{a}_n - a)$. Qu'en déduire ?

9. (théorique) Déterminer la loi limite de $n\left(\widehat{a} - a,\, b - \widehat{b}\right)$. Pour cela, on pourra calculer

$$\mathbb{P}\left(\{n(\widehat{a}_n - a) > t\} \,\cap\, \{n(b - \widehat{b}_n) > u\}\right).$$

Que peut-on dire de la dépendance asymptotique entre $\widehat{a}$ et $\widehat{b}$?

10. Pour vérifier le résultat précédent :
- Générer $N = 1000$ échantillons de $n = 100$ observations *iid* de loi $\mathcal{U}[a, b]$, avec $a = 0$ et $b = 1$.
- Calculer l'EMV $(\widehat{a}_n, \widehat{b}_n)$ pour chacun d'eux. On dispose de $N = 1000$ réalisations de l'EMV.
- Tracer l'histogramme des $n\widehat{a}_n$.
- Faire un QQ-plot pour comparer la distribution des $n\widehat{a}_n$ avec la loi exponentielle de paramètre $1/(b - a)$. De même pour $n(1 - \widehat{b})$.
- Calculer la corrélation des $\widehat{a}_n$ avec les $\widehat{b}_n$.

11. En déduire un intervalle asymptotique de confiance de niveau 95% pour a.

Exercice 3 (Sondages avec SAS)
On s'intéresse aux procédures `surveyselect` et `surveymeans` de **SAS** concernant la sélection d'échantillons, et l'estimation de moyennes, totaux et ratios. Les données, issues du recensement de 1999, sont dans la table `rec99htegne` disponible sur ma page. Ces données correspondent aux 554 communes de moins de 10000 habitants de la Haute-Garonne. Pour chacune de ces communes, nous nous intéressons à l'estimation du nombre total de logements vacants. L'ensemble des communes est partitionné en 32 Bassins de vie quotidienne (BVQ). La variable nombre de logements (LOG) est considérée comme une information auxiliaire. Les communes sont réparties en 4 strates (stratlog) constituée d'après cette variable auxiliaire.
Dans la suite de l'énoncé, `sondage` désigne la librairie dans laquelle vous avez copié votre table `rec99htegne`.

1. Calculer (avec **SAS**) la moyenne et l'écart-type du nombre de logements par strate. Commenter.

2. On cherche à créer un échantillon de 70 communes par un **sondage aléatoire simple** (simple random sampling en anglais). Pour cela utiliser le code :

```
proc surveyselect data=sondage.rec99htegne method=srs n=70 stats
   out=sondage.logsi1;
run;
```

A quoi correspondent les deux variables qui ont été ajoutée dans la table de sortie ? Vérifier que les données correspondent aux valeurs théoriques attendues.

3. Pour estimer le total à partir de la table des 70 communes tirées dans `logsi1`, on utilise le code suivant :

```
proc surveymeans data=sondage.logsi1 total=554 sum varsum clsum;
    var logvac;
weight Samplingweight;
 run;
```

Les commandes `sum varsum clsum` permettent d'obtenir les estimateurs du total, de sa variance, et l'intervalle de confiance pour ce total. Comparer l'estimateur du total et sa variance aux valeurs "théoriques".

4. On souhaite recommencer avec un **sondage stratifié**, en constituant un échantillon de 70 communes mais en tirant respectivement 5, 10, 21 et 34 communes dans chacune des strates déterminée par `stratlog`. Pour cela, utiliser le code suivant :

```
proc surveyselect data=sondage.rec99htegne method=srs
n=(5 10 21 34)
out=sondage.logsi2;
strata stratlog;
run;
```

Vérifier qu'il y a bien le bon nombre de communes tirées par strate.

5. Récupérer dans une table `sondage.stratsize` le nombre de communes par strate dans la table `rec99htegne`. On appellera cette variable `_TOTAL_`.

6. Reprendre la question 3 pour l'estimation du total, qui se fait ici avec le code suivant :

```
proc surveymeans data=sondage.logsi2 total=sondage.stratsize sum;
 strata stratlog;
 var logvac;
 weight Samplingweight;
run;
```

Comparer aux valeurs théoriques. Commenter.

Exercice 4 (Exo bonus : Test du poker)
Nous revenons sur les tests du générateur de réalisations de la loi uniforme de **R**. Le test du poker étudie l'homogénéité de la distribution des chiffres dans les nombres.
1. Simuler $n = 10000$ entiers aléatoires compris entre 0 et 9 999.

2. Compter la fréquence des nombres formés de 4 chiffres différents, de 2 chiffres identiques et 2 différents (paire), de 3 chiffres identiques (brelan), de 2 paires, de 4 chiffres identiques.

3. Tester avec ce découpage l'hypothèse nulle H_0 : *la loi des observations est uniforme sur* $[\![0, 9\,999]\!]$.