Université des Sciences et Technologies de Lille
U.F.R. de Mathématiques Pures et Appliquées

CAPES-M1 **Année 2010-2011**

Fiche de T.D. n° 7

Ex 1. *Test sur les intentions de vote*

1) Soit S_n une variable aléatoire sur $(\Omega, \mathcal{F}, (\mathbf{P}_\theta)_{\theta\in[0,1]})$ de loi binomiale $\mathrm{Bin}(n,\theta)$ sous $\mathbf{P}_\theta$.

a) Montrer pour toute constante a, que l'application

$$\theta \mapsto \mathbf{P}_\theta(S_n \geq a)$$

est croissante sur $[0,1]$ (en θ). *Indication :* on utilisera une méthode de simulation de la loi binomiale.

b) Soit $\theta_0 \in [0,1]$. En déduire la construction d'un événement de la forme

$$R = \{S_n \geq a\}$$

de telle sorte que $\sup_{\theta\leq\theta_0} \mathbf{P}_\theta(R) \leq 0,05$.

2) Application à la construction d'un test. Avant une élection, un candidat A affirme qu'il est sûr de dépasser le score des 30%. Pour savoir s'il a raison, on interroge un nombre n suffisant de personnes et on note S_n le nombre de personnes parmi les n interrogées qui ont l'intention de voter pour le candidat A. Si le nombre S_n dépasse un certain seuil (à déterminer), le candidat A affirmera que son score dépasse 30% en souhaitant que le risque de se tromper n'excède pas 5%.

On introduit alors le paramètre θ représentant la probabilité (inconnue) de voter pour A sur l'ensemble des electeurs et un modèle statistique $(\Omega, \mathcal{F}, (\mathbf{P}_\theta)_{\theta\in]0,1[})$ dans lequel S_n suit une loi binomiale $\mathrm{Bin}(n,\theta)$ sous $\mathbf{P}_\theta$.

On observe une réalisation de S_n et à l'aide de cette observation, on veut décider l'une des deux hypothèses suivantes :

$$H_0 \quad : \quad \theta \leq 0,3 \qquad \text{ou bien si} \qquad H_1 \quad : \quad \theta > 0,3.$$

La règle de décision sera donc du type

- si on observe $S_n(\omega) \geq a$, alors on rejette H_0
- sinon, on accepte l'hypothèse H_0,

a étant un seuil à déterminer. Il est déterminé de telle sorte que pour tout θ vérifiant H_0, la probabilité P_θ(rejet de H_0) (donc ici , rejet à tort) soit contrôlée par un niveau donné (ici 5%).

a) Déterminer le seuil a (en fonction de n).

b) On recueille 32% d'intentions de votes auprès de 500 personnes, que pouvez-vous en conclure ?

Ex 2. *Application du théorème de Lindeberg à un test d'égalité de deux espérances.* On dispose d'un échantillon $X_1, \ldots, X_l$ de la loi inconnue P_X (les X_i sont i.i.d. de même loi qu'une v.a. aléatoire générique X) et d'un échantillon $Y_1, \ldots, Y_m$ de la loi P_Y (les Y_i sont i.i.d. de même loi qu'une v.a. aléatoire générique Y). On suppose de plus que ces deux échantillons sont indépendants. On voudrait tester l'hypothèse

$$(\mathcal{H}_0): \quad \mathbf{E}X = \mathbf{E}Y \qquad \text{contre} \qquad (\mathcal{H}_1): \quad \mathbf{E}X \neq \mathbf{E}Y.$$

Cette situation se présente notamment en médecine quand on veut tester l'efficacité d'un nouveau médicament et où l'on observe les durées de guérisons X_i d'un premier échantillon de patients auxquels on a administré le nouveau médicament, tandis que les Y_i sont les durées de guérison des patients d'un autre échantillon ayant reçu un médicament ancien (ou un placebo !). A priori on ne connaît pas les variances σ_X^2 et σ_Y^2 des lois P_X et P_Y. On construit une statistique de test

$$T := \frac{\overline{Y} - \overline{X}}{\sqrt{\frac{1}{l}S_X^2 + \frac{1}{m}S_Y^2}}, \tag{1}$$

où $\overline{X}$, $\overline{Y}$ sont les moyennes et S_X^2, S_Y^2 les variances empiriques :

$$\overline{X} := \frac{1}{l}\sum_{i=1}^{l} X_i, \quad S_X^2 := \frac{1}{l}\sum_{i=1}^{l}(X_i - \overline{X})^2, \quad \text{etc.}$$

Pour l et m « grands », on considère que sous $(\mathcal{H}_0)$, T doit être approximativement gaussienne $\mathfrak{N}(0,1)$. Ceci amène à définir un test de niveau ε en décidant de rejeter $(\mathcal{H}_0)$ si l'on observe $|T| > t_\varepsilon$, où t_ε est défini par $\mathbf{P}(|Z| > t_\varepsilon) = \varepsilon$, avec $Z \sim \mathfrak{N}(0,1)$.

La légitimation théorique de ce test est contenue dans le théorème suivant que l'on vous propose de démontrer.

Théorème. *On suppose que X et Y sont de carré intégrable (avec $\sigma_X, \sigma_Y > 0$) et que $l = l(n)$ et $m = m(n)$ tendent vers l'infini avec n. Alors sous $(\mathcal{H}_0)$ la statistique de test $T = T_n$ définie par (1) converge en loi vers $\mathfrak{N}(0,1)$ lorsque n tend vers l'infini. Par contre, sous $(\mathcal{H}_1)$, $|T_n|$ tend p.s. vers l'infini.*

1) Sous $(\mathcal{H}_0)$ ou $(\mathcal{H}_1)$, donner les limites presque sûres du numérateur et du dénominateur de T_n. En déduire la limite p.s. de $|T_n|$ sous $(\mathcal{H}_1)$.

2) On se place désormais sous $(\mathcal{H}_0)$. Comme on a une forme indéterminée du type « $0/0$ » pour la convergence p.s. de T_n, on étudie plutôt la convergence en loi. Considérons pour cela le tableau triangulaire de n-ième ligne (rappelons que $l = l(n)$ et $m = m(n)$) :

$$\frac{-X_1'}{l}, \ldots, \frac{-X_l'}{l}, \frac{Y_1'}{m}, \ldots, \frac{Y_m'}{m}, \tag{2}$$

où l'on a posé $X_i' := X_i - \mathbf{E}X_i$ et $Y_i' = Y_i - \mathbf{E}Y_i$. La somme de la ligne est $S_n' = \overline{Y'} - \overline{X'} = \overline{Y} - \overline{X} + \mathbf{E}X - \mathbf{E}Y$, mais comme on est sous $(\mathcal{H}_0)$, $\mathbf{E}X - \mathbf{E}Y = 0$, d'où $S_n' = S_n$. Grâce à l'indépendance, sa variance est

$$s_n^2 = \operatorname{Var} S_n = \frac{1}{l^2}\sum_{i=1}^{l} \operatorname{Var} X_i + \frac{1}{m^2}\sum_{j=1}^{m} \operatorname{Var} Y_j = \frac{1}{l}\sigma_X^2 + \frac{1}{m}\sigma_Y^2.$$

Justifiez les égalités $S_X^2 = S_{X'}^2$, $S_Y^2 = S_{Y'}^2$, d'où $T'_n = T_n$, en notant T'_n la statistique obtenue en centrant toutes les observations.

3) Vérifiez la convergence

$$\frac{\frac{1}{l(n)}\sigma_X^2 + \frac{1}{m(n)}\sigma_Y^2}{\frac{1}{l(n)}S_X^2 + \frac{1}{m(n)}S_Y^2} \xrightarrow[n\to+\infty]{\text{p.s.}} 1.$$

En écrivant T_n sous la forme

$$T_n = \left(\frac{\frac{1}{l}\sigma_X^2 + \frac{1}{m}\sigma_Y^2}{\frac{1}{l}S_X^2 + \frac{1}{m}S_Y^2}\right)^{1/2} \times \frac{S'_n}{s_n}$$

on réduit la preuve du théorème à la vérification de la condition de Lindeberg pour le tableau triangulaire (2), pourquoi ?

4) Il s'agit donc de montrer pour tout $\varepsilon > 0$ la convergence vers 0 de

$$R_n(\varepsilon) = \frac{1}{s_n^2}\left(\sum_{i=1}^{l}\mathbf{E}\left(\left(\frac{-X'_i}{l}\right)^2\mathbf{1}_{\{|X'_i/l|>\varepsilon s_n\}}\right) + \sum_{j=1}^{m}\mathbf{E}\left(\left(\frac{Y'_j}{m}\right)^2\mathbf{1}_{\{|Y'_j/m|>\varepsilon s_n\}}\right)\right).$$

Vérifiez que

$$R_n(\varepsilon) = \frac{1}{ls_n^2}\mathbf{E}\left(X_1'^2\mathbf{1}_{\{X_1'^2>\varepsilon^2 l^2 s_n^2\}}\right) + \frac{1}{ms_n^2}\mathbf{E}\left(Y_1'^2\mathbf{1}_{\{Y_1'^2>\varepsilon^2 m^2 s_n^2\}}\right).$$

puis, que

$$R_n(\varepsilon) \le \frac{1}{\sigma_X^2}\mathbf{E}\left(X_1'^2\mathbf{1}_{\{X_1'^2>\varepsilon^2 l^2 s_n^2\}}\right) + \frac{1}{\sigma_Y^2}\mathbf{E}\left(Y_1'^2\mathbf{1}_{\{Y_1'^2>\varepsilon^2 m^2 s_n^2\}}\right).$$

5) Expliquez pourquoi $l(n)^2 s_n^2$ tend vers l'infini quand n tend vers l'infini et en déduire que

$$\mathbf{E}\left(X_1'^2\mathbf{1}_{\{X_1'^2>\varepsilon^2 l^2 s_n^2\}}\right) \xrightarrow[n\to+\infty]{} 0.$$

Conclure.

Ex 3. *Dé de Weldon*

Weldon a lancé $n = 315\,672$ fois un dé en notant pour chaque lancer la réalisation ou non de l'évènement E « obtention du 5 ou du 6 ». Il a observé 106 602 réalisations de E.

1) Au vu de ces observations, peut-on accepter au niveau $\alpha = 5\%$ l'hypothèse nulle $P(E) = \frac{1}{3}$?

2) À quel niveau maximal α peut-on accepter cette hypothèse ?