Examen de Statistique

$2^{\text{ème}}$ année Informatique et Mathématiques Appliquées

mardi 24 novembre 2009

documents autorisés : 1 feuille au format A4, recto-verso.

durée : 2 heures.

Les trois exercices sont indépendants. L'exercice 1 (estimation) sera rédigé sur une première copie. Les exercices 2 et 3 (tests) seront rédigés sur une autre copie.

1 Estimation

Le nombre de pannes N d'une unité de production est observé sur une année. On effectue un test de fiabilité consistant à contrôler $n = 10000$ unités pendant une année et à noter leur nombre de pannes respectif N_i, $\{1, ..., n\}$. Le tableau suivant donne un récapitulatif des résultats de ce test.

Nombre de pannes	Nombre d'unités de production concernées
0	1695
1	3566
2	3131
3	1301
4	285
5	22

$$\begin{cases} n & = 1695 + 3566 + 3131 + 1301 + 285 + 22 = 10000, \\ \sum_{i=1}^{n} N_i & = 0 \times 1695 + 1 \times 3566 + 2 \times 3131 + 3 \times 1301 + 4 \times 285 + 5 \times 22 = 14981. \end{cases}$$

1. Dans un premier temps, supposons que les N_i $(i = 1, ..., n)$ sont des variables aléatoires indépendantes et identiquement distribuées selon une loi de Poisson de paramètre λ. Déterminer $\widehat{\lambda}_{\mathrm{MM}}$ l'estimateur de λ par la méthode des moments. Déterminer son biais et sa variance. Est-il convergent ?
2. Dans un deuxième temps, supposons que les N_i $(i = 1, ..., n)$ sont des variables aléatoires indépendantes et identiquement distribuées selon une loi de binomiale $\mathcal{B}(5, p)$. Déterminer $\widehat{p}_{\mathrm{MV}}$ l'estimateur du maximum de vraisemblance de p. Est-il sans biais ? convergent ? efficace ?
3. En supposant que le théorème Central-Limite s'applique pour $n = 10000$, déterminer un intervalle de confiance à 95% pour $\widehat{p}_{\mathrm{MV}}$ c'est-à-dire calculer t tel que $P\left[p \in [\widehat{p}_{MV} - t; \widehat{p}_{MV} + t]\right] = 0.95$.
4. Des mesures effectuées dans un autre lieu de production ont montré que le paramètre p peut être modélisé par une variable aléatoire distribuée suivant une loi Beta $\mathcal{B}e(\alpha, \beta)$ où les paramètres α et β sont connus.
   (a) Déterminer $\widehat{p}_{\mathrm{MAP}}$ l'estimateur du maximum *a posteriori* de p en fonction de α et β.
   (b) Etudier le comportement asymptotique de $\widehat{p}_{\mathrm{MAP}}$ pour le cas $\alpha = \beta \approx 1$. Interpréter.

2 Test d'ajustement

D'après les mesures, on souhaite vérifier si le nombre de pannes par an pour une unité de production suit une loi binomiale. Quel test d'ajustement est le plus approprié. Effectuer le test. On donne la table de la loi Binomiale(5, $\widehat{p}_{\mathrm{MV}}$) :

k	$\mathrm{P}[X=k]$
0	0.1685
1	0.3605
2	0.3084
3	0.1319
4	0.0282
5	0.0024

3 Test paramétrique

On suppose dans la suite que les variables aléatoires N_i $(i = 1, ..., n)$ sont indépendantes et identiquement distribuées selon une loi binomiale $\mathcal{B}(5, p)$ .On désire choisir entre deux hypothèses :

$$\begin{aligned} H_0 &: p = p_0 = 0.3 \\ H_1 &: p = p_1 = 0.4 \end{aligned}$$

1. Construire le test de Neyman-Pearson : expliciter la statistique de test et préciser la règle de décision.
2. Déterminer la loi exacte de la statistique de test sous chacune des hypothèses. Etant donné le nombre d'observations disponibles, fournir une loi approchée de la statistique de test.
3. Donner l'expression du seuil de décision pour un risque $\alpha = 5\%$ en utilisant la loi approchée de la statistique de test.
4. Calculer la puissance du test $\pi = 1 - \beta$ pour un risque $\alpha = 5\%$.
5. Effectuer le test sur les observations fournies.

Lois utiles :

- Loi Beta $\mathcal{B}e(\alpha, \beta)$: $f(x) = kx^{\alpha-1}(1-x)^{\beta-1}, k = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}, \alpha > 0, \beta > 0, x \in]0, 1[$.
  $E[X] = \frac{\alpha}{\alpha+\beta}, var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
- Loi Binomiale $B(n,p)$: $P[X = k] = C_n^k p^k q^{n-k}$, $p \in [0,1], q = 1-p$
  $E[X] = np, var[X] = npq$. Fonction caractéristique $\varphi_X(t) = \left(pe^{it} + q\right)^n$.
- Loi de Poisson $P(\lambda)$: $P[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}$, $\lambda > 0, k \in \mathbb{N}^*$
  $E[X] = \lambda, var[X] = \lambda$.

Table du Test de Kolmogorov

La table suivante donne l'écart maximal théorique, Δ_{kolmo}, entre fonctions de répartition empirique et théorique pour accepter l'hypothèse H_0 avec un risque α de 5 ou 1%, en fonction de Nc (nombre de points de calcul de la fonction de répartition, c'est-à-dire nombre de classes de l'histogramme). Si l'écart maximal mesuré $\Delta_{\max}$ est inférieur à Δ_{kolmo}, on accepte l'hypothèse H_0 avec un risque α.

N_c	$\alpha = 5\%$	$\alpha = 1\%$
5	0.5633	0.6685
10	0.4087	0.4864
15	0.3375	0.4042
20	0.2939	0.3524
25	0.2639	0.3165
30	0.2417	0.2898
40	0.2101	0.2521
50	0.1884	0.2260
60	0.1723	0.2067
70	0.1597	0.1917
80	0.1496	0.1795
90	0.1412	
100	0.1340	

Cette table n'est valable que pour N_c inférieur à 100. Pour N_c plus grand, on sait calculer $P(\max|\widehat{F} - F| > \Delta_{kolmo})$ sous H_0 (Vladimir N.Vapnik, "The Nature of Statistical Learning Theory", Springer Ed. p87). Or :

$$\alpha = P\left[\text{rejeter } H_0 \mid H_0 \text{ vraie}\right] = P(\max|\widehat{F} - F| > \Delta_{kolmo}) \simeq \sum_{k=1}^{k=+\infty} 2(-1)^{k-1} \exp(-2k^2\, N_c\, \Delta_{komo}^2)$$

On note $\Delta_{\max} = \max|\widehat{F} - F|$ l'écart maximal entre la fonction de répartition théorique et la fonction de répartition estimée :

$$P_{\Delta_{\max}} = P(\max|\widehat{F} - F| > \Delta_{\max}) \simeq \sum_{k=1}^{k=+\infty} 2(-1)^{k-1} \exp(-2k^2\, N_c\, \Delta_{\max}^2)$$

Cette probabilité est une fonction décroissante de $\Delta_{\max}$. Donc si $P_{\Delta_{\max}} > \alpha$, on peut en déduire que $\Delta_{\max} < \Delta_{kolmo}$. Il n'est donc pas nécessaire de calculer Δ_{kolmo} pour prendre la décision. En effet, on accepte l'hypothèse H_0 si :

$$\sum_{k=1}^{k=+\infty} 2(-1)^{k-1} \exp(-2k^2\, N_c\, \Delta_{\max}^2) > \alpha$$