# T.D. de Probabilités et Statistiques

# Licence Informatique

Gilles HUNAULT, 1998

## T.D. 8 : Approximations et Algorithmes

1. Intervalle de confiance et de variabilité
2. Intervalle de confiance à 5 %
3. Etalonnage de fréquence d'un spectrophotomètre
4. Approximation de $\mathcal{B}(n, p)$ par $\mathcal{N}(0, 1)$
5. Approximation de $\mathcal{P}(\lambda)$ par $\mathcal{N}(0, 1)$
6. Approximation de $\mathcal{B}(n, p)$ par $\mathcal{P}(\lambda)$
7. Reprise des exercices *montre*, *concentrateur* et *cinéma*
8. Algorithme du calcul de $m$, $\sigma$ pour une matrice de données
9. Algorithme de $H(N, D, n)$ (exam. mai 97)

# 1. Intervalle de confiance et de de variabilité

### Enoncé

Soit $X$ une série statistique de moyenne $m$, de variance $V$ et d'écart-type $\sigma$. On appelle intervalle de confiance à $\alpha\%$ l'intervalle centré $I_m = [m-t\sigma, m+t\sigma]$ où $t$ correspond $p(U > \alpha)$. On appelle intervalle de variabilité l'intervalle centré $I_V = [m-tV, m+tV]$ et intervalle de sûreté $I_s = [m-t\sigma/\sqrt{n}, m+t\sigma\sqrt{n}]$

Soit $X = (12, 15, 17, 50)$, $Y = (25, 31, 35, 101)$ et $Z = (144, 255, 289, 2500)$ trois séries statistiques correspondant à des variables quantitatives, dont les unités sont respectivement la minute, le kilomètre et la minute-carrée.

Donner la matrice des corrrélations de $X$, $Y$ et $Z$ ainsi que leur intervalle de confiance, leur intervalle de variabilité pour $t = 1.96$ et leur intervalle de sûreté.

On fournit, si cela peut aider les sommes suivantes

```
sx       sy       sz
94       192      3158
sx*sx    sy*sy    sz*sz
3158     13012    6404882
sx*sy    sx*sz    sy*sz
6410     135016   273190
```

### Solution

On trouve assez facilement $m_X = 23.5min$, $m_Y = 48\ km$, $m_Z = 789.5\ min^2$, $\sigma_X = 15.403\ min$, $\sigma_Y = 30.806\ km$ et $\sigma_Z = 988.893\ min^2$ soit des cdv respectifs de 70 %, 64 % et 125 %.

La matrice des corrélations est

```
        X        Y        Z

X       1.0000
Y       1.0000   1.0000
Z       0.9980   0.9980   1.0000
```

Et on on trouve donc que $Y = 2X + 1$ et donne des $km$ par $min$, $Z = 64.071 * X - 716.159$ (alors qu'en fait la "vraie" liaison est $Z = X^2$).

Pour la variable $X$, $I_m = [m - t\sigma, m + t\sigma] = [23.5 - 1.96 * 15.4, 23.5 + 1.96 * 15.4]$ est donc l'intervalle $[-6.68, 53.68]$. L'intervalle $I_V = [m - t.V, m + t.V]$ est stupide car $m$ et $V$ n'ont pas les mêmes unités.

Pour $I_s = [m - t\sigma/\sqrt{n}, m + t\sigma\sqrt{n}]$, comme $n$ vaut 4, on trouve $[8, 4, 38.6]$

# 2. Intervalle de confiance à 5 %

### Enoncé

Au risque de $\alpha = 5\%$, la valeur de $t$ pour l'intervalle de confiance est 1.96 ; comparer ces intervalles de confiance pour $\mathcal{B}(12; 0.3)$ et $\mathcal{P}(\lambda)$ pour $\lambda$ bien choisi.

### Solution

Le $\lambda$ bien est celui tel que les deux lois aient la même moyenne, d'où $\lambda = np = 12 * 0.3 = 3.6$. Pour la loi binomiale $\sigma = \sqrt{np(1-p)} = \sqrt{2.52} = 1.59$ alors que pour la loi de Poisson, $\sigma = \sqrt{\lambda} = \sqrt{3.6} = 1.90$. Pour la loi binomiale, $t\sigma/\sqrt{n} = 1.96 * 1.59/\sqrt{13} = 0.8629$ donc l'intervalle est $[2.7371, 4.4629]$. Pour la loi de Poisson, on suppose que la loi est tronquée aux 13 premières valeurs, d'où $t\sigma/\sqrt{n} = 1.96 * 1.9/\sqrt{13} = 1.0314$, $I = [2.5686, 4.6314]$.

# 3. Etalonnage de fréquence

### Enoncé

Soient $A$ et $B$ les mesures d'étalonnage prises respectivement pour 250 $m\mu$ et 260 $m\mu$.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 120 | 164 | 153 | 148 | 143 | 132 | 155 | 142 | 169 | 144 |
| B | 172 | 210 | 206 | 199 | 192 | 181 | 204 | 190 | 218 | 198 |

Comparer ces résultats à l'aide du test des rapport des variances $R = V_A/V_B$ où le seuil maximal est donné dans la table de *Snedecor*.

**Solution**

Pour nos deux séries de 10 valeurs, on trouve $m_A = 147$, $m_B = 197$, $V_A = 208.67$, $V_B = 188.89$. $R$ vaut donc 1.10 or le seuil $R_{max}$ lu dans la table au seuil de $\alpha$ =5 pour 9 ddl. On considére donc qu'il n'y pas de différence significative entre nos séries de valeurs...

# 4. Approximation de $\mathcal{B}(n, p)$ par $\mathcal{N}(0, 1)$

**Enoncé**

Soit $X = \mathcal{B}(15, 0.3)$. Effectuer un calcul direct de $p(X \in [3, 6])$. Calculer cette même probabilité de façon approchée en utilisant la loi normale.

**Solution**

La valeur exacte est $\Sigma_{k=3}^{6} C_{15}^{k} 0.3^k (1-0.3)^{15-k}$ soit (par exemple avec $Maple$) 0.7419. On a bien sûr $m_X = 4.5$, $V_X = 3.15$, $\sigma_X = 1.7748$. Si on utilise la loi normale, on assimile$[\![3, 6]\!]$ à $[2.5, 6.5]$ soit $U \in [(2.5 - m_X)/\sigma_X, (6.5 + m_X)/\sigma_X]$ c'est à dire $U \in [-1.126, +1.126]$. Utilisant la fonction de répartition $F_U$ de la loi normale $U = \mathcal{N}(0, 1)$ on trouve $p(U \in [-a, a]) = 2F_u(a) - 1$ et pour $a = 1.126$, $F_U(0.8699]$ donc $p = 0.7398$.

# 5. Approximation de $\mathcal{P}(\lambda)$ par $\mathcal{N}(0, 1)$

**Enoncé**

Soit $X = \mathcal{P}(20)$. Effectuer un calcul direct de $p(X \leq 10)$.

Calculer cette même probabilité de façon approchée en utilisant la loi normale.

**Solution**

Le calcul direct donne $\displaystyle\sum_{k=0}^{10} e^{-20} 20^k / k!$ soit 0.0165.

On a $m_X = 20 = V_X$,$\sigma_X = 4.4721$. Si on assimile $X \leq 10$ à $X \leq 10.5$ alors $p = p(\ (X - m)/\sigma \ \leq \ (10.5 - 20)/4.4721\ )$ vaut $F_U(-2.1243)$ soit encore $1 - F_U(2.1243) = 1 - 0.09832 = 0.0168$

# 6. Approximation de $\mathcal{B}(n, p)$ par $\mathcal{P}(\lambda)$

**Enoncé**

Un livre de 1000 pages contient 1500 erreurs. Donner une approximation de la probabilité qu'une page contienne moins de 2 erreurs en utilisant la loi normale. Donner la valeur exacte de cet événement.

Calculer la probabilité de cet événement si on remplace la loi binomiale sous-jacente par une loi de Poisson bien choisie.

**Solution**

Bien sûr, $X$ est la loi $\mathcal{B}(n = 1500, p = 1/1000)$ de moyenne $m = np$ =1.5 et de variance $V = np(1-p)$ =1.4985 soit un écart-type de 1.2241. $X < 2$ se traduit par l'union disjointe de $X = 0$ et $X = 1$ de probabilités respectives $C_n^0 p^0 (1-p)^n$ et $C_n^1 p^1 (1-p)^{n-1}$ soit 0.22296+0.33478 donc à peuprès 0.55774.

L'approximation gaussienne dit que

$$p(X < 2) = p\left(\frac{X-m}{\sigma} \;<\; \frac{2-np}{\sqrt{np(1-p)}}\right) \text{ vaut } F_U\left(\frac{1.5-np}{\sqrt{np(1-p)}}\right)$$

soit ici $F_U(0) = 0.5$. L'erreur commise entre le 0.5 d'approximation et les 0.5577 véritable provient des conditions d'applications qui ne sont pas respectées ($n > 30$ est vérifiée mais $np > 5$ ne l'est pas car $np$ vaut 1.5).

Prenant $\lambda = np = 1.5$ et remarquant que les conditions d'application $n > 36$ et $np < 5$ sont respectées, $p(X = 0) + p(X = 1)$ devient $e^{-\lambda}\lambda^0/0! + e^{-\lambda}\lambda^1/1!$ soit, puisque $e^{-\lambda} = e^{-1.5} = 0.22313$, 0.22313+0.33469 ce qui aboutit à la valeur 0.55782 qui elle, est très proche du résultat exact 0.55774.

# 7. Reprise : *montre*, *concentrateur* et ***cinéma***

**Enoncé**

Une montre fait une erreur d'au plus 30 secondes par jour. Calculer la probabilité que l'erreur soit inférieure à 15 minutes au bout d'un an.

Un serveur concentrateur dessert 1000 postes via 50 lignes à haut débit. Aux heures de pointe, chaque poste est occupé en moyenne pendant 2.5 secondes.

Quelle est la probabilité de saturation de l'ensemble des lignes à un instant donné pendant une minute de pointe ?

Un cinéma comporte 2 salles de $n$ places. Si $N$ personnes se présentent, en admettant que les choix des spectateurs sont indépendants, chaque spectateur a une chance sur 2 d'aller dans la prmière salle. Comment choisir $n$ pour que la probabilité qu'un spectateur ne puisse pas voir le film qu'il (elle) a choisi soit inférieure à $\varepsilon$ ? Application numérique : $N = 1000$, $\varepsilon = 0.01$.

**Solution**

Pour la montre, soit $X_i$ l'erreur en minute au jour $i$. $X_i$ peut être modélisée par une loi réelle uniformément répartie sur $[-1/2, 1/2]$. Supposant "les jours indépendants", l'erreur $S$ au bout d'un an est la somme des $X_i$ pour $i$ de 1 à 365 (on ignore "poliment" les années bisextiles) : $S = \sum_{i=1}^{365} X_i$. La moyenne de $S$ est donc $\mu = 0$, $\sigma = 1/\sqrt{12}$. Le théorème de limite centrale permet d'assimiler à

$$\frac{S - n\mu}{\sigma\sqrt{n}} = S\sqrt{\frac{12}{365}} \text{ à } \mathcal{N}(0,1)$$

$p(|S| \leq 15)$ est donc assimilable à $p\left( \ |U| \leq 15 \leq \sqrt{15/365}\ \right)$ soit $p(|U| \leq 2.72)$. Cette probabilité vaut donc $2F_U(2.72) = -1 = 2 * 0.99674 - 1$ soit à peu près 0.99.

Pour le concentrateur, à partir de la variable aléatoire binaire "un poste est occupé", modélisé par $b(p = 2,5/60)$, "le nombre de postes occupés désirant communiquer avec le serveur". correspond alors à la loi $Y$ somme de $n = 1000$ lois $b(p)$ (considérées comme indépendantes). C'est donc $\mathcal{B}(n,p) = \mathcal{B}(1000, 0.0417)$ La saturation correspond à l'évènement $Y > 50$. Sa probabilité exacte est

$$p(Y \geq 51) = \sum_{k=51}^{1000} C_{1000}^{k}\, p^k (1-p)^{1000-k}$$

et elle vaut 0.08413259 (avec *Maple* par exemple).

Si au lieu de la calculer directement on effectue une approximation par la loi normale, puisque $n = 1000 > 36$, $np = 41.6 > 5$ et $np(16p) = 39.869 > 5$, on remplace

$$p(Y > a) = p\left(\frac{Y-m}{\sigma} > \frac{a-np}{\sqrt{np(1-p)}}\right) \text{ par } \int_a^{+\infty} \frac{1}{\sqrt{2\pi}e^{-t^2/2}}dt$$

soit, pour $a = 50$, $p = 1 - F_U(1.319)$=1-0.9066 donc 0.0934

Pour les salles de cinéma, à partir de l'évènement $S_i$ "le spectateur $i$ veut aller dans la salle 1" modélisé par $b(1/2)$, pour $S$ le nombre de spectateurs ayant choisi la salle 1. et $T$ le nombre de spectateurs ayant choisi la salle 2 on a $T = N - S$ et $S = \sum_{i=1}^{N} S_i = \mathcal{B}(N, 1/2)$ de moyenne $\mu = 1/2$, de variance $V = 1/4$ et d'écart-type $\sigma = 1/2$. L'évènement "un spectateur ne peut pas voir le film qu'il a choisi" correspond aux évènements "$S > n$" ou "$T > n$" si chaque salle comporte $n$ places ce qui s'écrit encore "$|S-N/2| > n-N/2$" au cas où $N \leq 2n$. On veut donc calculer $n$ pour que $p(|S-N/2| > n-N/2 \leq \varepsilon)$. Utilisant l'approximation par la loi normale, et remplaçant $|S - N\mu| > \alpha$ par $|\mathcal{N}(0,1)|/\sigma\sqrt{N} > \alpha$ on calcule $p\Big(\ |\mathcal{N}(0,1)| > (n - N/2).2/\sqrt{N}\ \Big)$. On trouve $p(|U| > (2n - N)/\sqrt{N}) = 2\Big(\ 1 - F_U((2n - N)/\sqrt{N})\ \Big) < \varepsilon$. Pour l'application numérique : $p < 0.01$ donne $F_U(y) > 0.995 = F_U(2.58)$ ce qui implique $y \geq 2.58$, soit $(2n - N)/\sqrt{N} \geq 2.58$ c'est à dire finalement $n \geq (\sqrt{N}2.58 + N)/2$ donc $n \geq 540,8$ pour $N = 1000$, $\sqrt{N} = 31.62$.

# 8. Algorithme de $m$, $\sigma$

### Enoncé

Soit $X$ un tableau de $n$ valeurs notées $X[1]$, $X[2]$... $X[n]$. Donner un algorithme du calcul de $m_X$, $\sigma_X$ et $\sigma_X/m_X$.

Soit $D$ un tableau de données dont les $n$ lignes sont repérées par un premier $i$ et les $p$ colonnes sont repérées par un second indice $j$ : $D[i,j]$ désigne donc la valeur à l'intersection de la ligne $i$ et de la colonne $j$. On admettra que les colonnes 1 à $t$ de $D$ correspondent à des variables quantitatives et que la colonne $t+1$ contient les codes d'une variable qualitative avec $q$ codes notés $1, 2...q$.

Sachant que le tableau $D$ est déjà constitué, que les valeurs $n, p, t, q$ sont toutes définies et valides, on veut calculer la moyenne $moy[k, z]$, l'écart-type $sig[k, z]$ et le coefficient de variation $cdv[k, z]$ de chacune des sous-populations (ici correspondant au code $k$) pour la variable $z$. On supposera les tableaux $moy$, $sig$ et $cdv$ déjà déclarés. On utilisera un tableau $eff[k]$ que l'on remplira avec l'effectif de la population $k$.

a) Si les données sont

```
1     1      1   |   1
2     2      4   |   1
1     3     25   |   2
3     4      9   |   1
1     5     16   |   2
```

compléter le tableau suivant des résultats à 0,1 près

```
n = 5 lignes ; p = 3 colonnes ; q = 2 sous-populations
pop.  1  eff[1]   =  3
         moy[1,1] =    2.0 sig[1,1] =    0.8 cdv[1,1] =    0.4
         moy[1,2] =    2.3 sig[1,2] =    1.2 cdv[1,2] =    0.5
         moy[1,3] =    4.7 sig[1,3] =    3.3 cdv[1,3] =    0.7
pop.  2  eff[2]   =  ????
         moy[2,1] =    ??? sig[2,1] =    0.0 cdv[2,1] =    0.0
         moy[2,2] =    4.0 sig[2,2] =    ??? cdv[2,2] =    0.3
         moy[2,3] =   20.5 sig[2,3] =    4.5 cdv[2,3] =    ???
```

b) Ecrire un algorithme qui calcule ces résultats. Vous n'utiliserez aucun autre tableau que $eff$, $moy$, $sig$ et $cdv$. On ne demande aucun affichage.

**Solution** Les valeurs numériques sont

```
pop.  1  eff[1]   =  3
         moy[1,1] =    2.0 sig[1,1] =    0.8 cdv[1,1] =    0.4
         moy[1,2] =    2.3 sig[1,2] =    1.2 cdv[1,2] =    0.5
         moy[1,3] =    4.7 sig[1,3] =    3.3 cdv[1,3] =    0.7
pop.  2  eff[2]   =  2
         moy[2,1] =    1.0 sig[2,1] =    0.0 cdv[2,1] =    0.0
         moy[2,2] =    4.0 sig[2,2] =    1.0 cdv[2,2] =    0.3
         moy[2,3] =   20.5 sig[2,3] =    4.5 cdv[2,3] =    0.2
```

Un algorithme possible est

```
# initialisation des tableaux

pour ind de1a q
  eff[ind] ← 0
  pour jnd de1a t
     moy[ind,jnd]  ← 0
     sig[ind,jnd]  ← 0
  finpour jnd de1a t
finpour ind de1a q

# calcul des effectifs, des sommes et de leur carre

pour lig de1a n
  pour col de1a t
       ind          ← D[lig,t+1]
       si col = 1 alors  eff[ind] ← eff[ind] + 1 finsi
       valr         ← D[lig,col]
       moy[ind,col] ← moy[ind,col] + valr
       sig[ind,col] ← sig[ind,col] + valr*valr
  finpour col de1a t
finpour lig de1a n

# calcul des moyennes etc.

pour ind de1a q
  pour jnd de1a t
     moy[ind,jnd]  ← moy[ind,jnd]  / eff[ind]
     sig[ind,jnd]  ← sig[ind,jnd]  / eff[ind]
     sig[ind,jnd]  ← sig[ind,jnd]  - moy[ind,jnd]*moy[ind,jnd]
     sig[ind,jnd]  ← racine( sig[ind,jnd]  )
     si moy[ind,jnd]<> 0
        alors cdv[ind,jnd]  ← abs( sig[ind,jnd] ) / moy[ind,jnd]
        sinon cdv[ind,jnd]  ← -1
     finsi moy[ind,jnd]<> 0
  finpour jnd de1a t
finpour col de1a t
```

# 9. Algorithme de $H(N, D, n)$ (exam. mai 97)

**Enoncé**

Soit $X$ la *v.a.* "loi hypergéométrique" $\mathcal{H}(N, D, n)$ : on tire n valeurs sans remise dans N boules dont D sont défectueux ; $X$ est la loi "nombre de défectueux". $X$ prend les valeurs entières $k = a, a+1, a+2$ ... $b$ où $a = max(0, D+n-N)$ et $b = min(D, n)$. La valeur $k$ a pour probabilité $p_k = C_D^k.C_{N-D}^{n-k}/C_N^n$. Expliciter les valeurs de $k$ et $p_k$ pour $N = 6$, $D = 3$ et $n = 2$.

Donner, en respectant la syntaxe algorithmique du cours, un algorithme qui calcule la moyenne de $X$ pour $n$, $N$ et $D$ donnés ; on supposera connue la fonction $C(n, p)$ qui calcule $C_n^p$. On calculera au passage la somme des probabilités, la moyenne, la variance, l'écart-type et le coefficient de variation de $X$.

**Solution**

Si $N = 6$, $D = 3$ et $n = 2$ alors $a = max(0, D+n-N) = max(0, 3+2-6) = 0$ et $b = min(D, n) = min(3, 2) = 2$. $X$ prend donc les valeurs 0, 1 et 2. Leur probabilité respective est $p_0 = C_3^0 C_3^2/C_6^2 = 3/15$, $p_1 = C_3^1 C_3^1/C_6^2 = 9/15$ et $p_2 = C_3^2 C_3^0/C_6^2 = 3/15$.

Pour le calcul des valeurs, des probabilités, on pourra utiliser l'algorithme suivant

```
Algorithme du calcul de H(N,D,n)

# calcul des bornes des valeurs de X
# nt est mis pour n, NB pour N

  a ← D+nt-NB
  si 0 > a alors
     a ← 0
  finsi
  b ← D
  si nt < b alors
     b ← nt
  finsi
```

```
# valeurs de X, somme et somme des carrés

  u  ← 0       # nombre de valeurs
  s  ← 0       # somme des valeurs
  sc ← 0       # somme des carrés des valeurs
  sp ← 0       # somme des probabilités
  de ← c(nt,N) # dénominateur constant pour k
  di ← N-D     # différence  constante pour k
  pour k de a à b
       p  ← c(k,D).c(nt-k,di) / de
       u  ← u  + 1
       s  ← s  + k*p
       sc ← sc + k*k*p
       sp ← sp + p
  fin pour k

# moyenne, variance

  m    ← s/u
  v    ← sc/u - m*m
  e    ← racine(v)
  cdv  ← 100*e/m
```