L3 Linguistique Informatique
Université Paris 7

Marie Candito

# Probabilités et statistiques pour le TAL

## Examen Session 1

## Jeudi 7 janvier 2010

**Recommandations générales : nommez les propriétés / lois / formules que vous utilisez pour résoudre les exercices. Vous serez également évalués là-dessus.**

## 1 Cours (3 pts)

1.1. Dans le cas d'un ensemble fondamental S à évènements élémentaires équiprobables, si E est un événement quelconque sur S, combien vaut P(E) ?

**=> En notant N(E) le nombre d'éléments de l'événement E, on aura P(E) = N(E) / N(S)**

1.2. Soit E et F deux évènements sur un ensemble fondamental S. Donnez la définition de la probabilité conditionnelle P(E|F)

**=> P(E|F) = P(EF) / P(F) si P(F) ≠ 0 (pour F = Ø, P(E|F) n'est pas définie)**

1.3. Comment définit-on formellement l'indépendance de deux évènements E et F ?

**(plusieurs formulations possibles)**

**E et F sont indépendants ⇔ P(E|F) = P(E) ⇔ P(F|E) = P(F) ⇔ P(EF) = P(E)P(F)**

**Ne pas confondre avec le caractère mutuellement exclusif (EF = Ø)**

1.4. Donnez le théorème de Bayes.

**=> P(E|F) = P(F|E) P(E) / P(F)**

## 2 Exercices

2.1. **(1,5 pts)** Pour un jeu télévisé où vont participer 10 candidats (5 femmes et 5 hommes), on choisit au hasard un ordre de passage des 10 candidats.

a) Quelle est la probabilité d'une séquence de candidats donnée ?

**L'ensemble des résultats possibles est constitué des 10! permutations possibles des 10 candidats. Chaque séquence a donc la probabilité 1/10!**

b) Quelle est la probabilité que le 2ème candidat soit une candidate ?

**On a 5 façons de choisir une femme en 2ème position, et 9! façons d'ordonner les 9 autres candidats, d'où la probabilité 5 x 9! / 10! = ½**

**On peut aussi résoudre plus directement, en remarquant qu'il y a symétrie totale entre hommes et femmes, et donc qu'à une position donnée on a autant de chances d'avoir un homme qu'une femme, d'où le résultat ½.**

2.2. **(3 pts)** Formule des probabilités totales. On considère qu'une phrase « journalistique » a une probabilité de 0,05 d'être averbale. Par ailleurs une phrase journalistique verbale a 1% de chance d'être un titre, alors qu'une phrase averbale a 60% de chances d'être un titre. Si on considère une phrase journalistique au hasard, quelle est la probabilité qu'elle soit un titre ?

**L'expérience considérée est le « tirage » d'une phrase journalistique, qui peut être un titre ou pas, et verbale ou averbale. On note T l'événement « la phrase est un titre », A l'événement « la phrase est averbale ». L'énoncé nous donne :**

$P(A) = 0{,}05$

$P(T \mid A) = 0{,}6$

$P(T \mid \overline{A}) = 0{,}01$

**On calcule P(T) par la formule des probabilités totales (« pour obtenir l'événement « la phrase est un titre », on a deux chemins : le cas « la phrase est averbale » et le cas « la phrase est verbale ») :**

$P(T) = P(T \mid A)P(A) + P(T \mid \overline{A})P(\overline{A}) = 0{,}6 \times 0{,}05 + 0{,}01 \times 0{,}95 = 0{,}0395$

2.3. **(4 pts)** Variable binomiale. Un système de reconnaissance vocale reconnaît des mots isolés correctement dans 95% des cas. On teste le système par série de 5 mots à reconnaître. Quelle est la probabilité que le système fasse au plus une erreur sur les 5 mots ? (donnez l'expression mathématique sans faire le calcul !)

**On considère l'expérience $\mathcal{E}$ « reconnaître 5 mots » comme la réalisation de 5 épreuves indépendantes $\mathcal{X}$ = « reconnaître un mot ».**

**Y la variable de Bernouilli sur l'expérience $\mathcal{X}$ qui vaut 1 si le mot est bien reconnu, et 0 sinon.**

**X la variable sur l'expérience $\mathcal{E}$ qui compte le nombre de succès (le nombre de mots bien reconnus) : c'est bien une variable binomiale, de paramètres n =5, et p=0,95**

**On cherche alors** $P(X=4) + P(X=5) = \binom{5}{4} p^4 (1-p)^1 + \binom{5}{5} p^5 (1-p)^0 = 5p^4(1-p) + p^5$

2.4. **(2 pts)** Soit la série statistique (10, 9, 6, 14, 17, 9, 10, 9, 12, 16) des longueurs d'un échantillon de 10 phrases. Définissez et calculez la moyenne et l'écart-type de cette série.

**moyenne** $= \bar{x} = \dfrac{\sum_{i=1}^{N} x_i}{N} = \dfrac{112}{10} = 11{,}2$

**écart-type** $= \sigma = \sqrt{\sigma^2} = \sqrt{\dfrac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}}$

2.5. **(3 pts)** Estimation. On considère un corpus de 100000 phrases du journal Le Monde, contenant 1481 phrases averbales.

a) On cherche la probabilité pour une phrase quelconque du Monde d'être averbale. Appliquez l'estimation vue en cours pour cette probabilité.

b) Quelle propriété a cette estimation (précisez mathématiquement)

**A partir du corpus de 100000 phrases, on peut dériver une séquence de 100000 résultats « averbal » ou « non averbal », et considérer cela comme la réalisation de 100000 épreuves indépendantes « tirer une phrase du Monde, et voir si elle est averbale ou pas ». Si on nomme p la probabilité cherchée (probabilité d'être averbale), alors la séquence de caractères « averbal » ou « non averbal » dérivée du corpus a comme probabilité** $p^{1481}(1-p)^{100000-1481}$

**On peut montrer que l'estimation de p par fréquence relative, ici 1481/100000 = 0,01481 est la valeur de p qui maximise cette probabilité :**

$p^{1481}(1-p)^{100000-1481}$ **est maximale pour p=1481/100000**

6. **(3,5 pts)** Règle de multiplication.

   **a)** Soient E1 et E2 deux évènements. Exprimez P(E1E2) à l'aide de probabilités conditionnelles.

   **P(E1E2) = P(E1)P(E2|E1)**

   **b)** Généralisez au cas de l'intersection de n évènements $E_1\ E_2 \ldots E_n$

   **$P(E_1E_2 \ldots E_n) = P(E_1)\ P(E_2|E_1)\ P(E_3|\ E_1E_2) \ldots P(En|\ E_1E_2 \ldots E_{n-1})$**

   **c)** Appliquez au cas suivant : On considère un système générant des phrases de 20 mots, pris dans un vocabulaire de 500 mots $m_1\ m_2 \ldots m_{500}$. On note $T_i = m_{ji}$ l'événement « le ième mot de la phrase générée est $m_{ji}$ ».

   **c1)** Exprimez sous forme d'intersection l'événement E correspondant à la génération de la phrase $m_{j_1}m_{j_2}m_{j_3}..m_{j_{20}}$

   **Avec les notations introduites, on peut écrire :**

   $E = T_1 = m_{j_1} \cap T_2 = m_{j_2} \cap T_3 = m_{j_3} \ldots \cap T_{20} = m_{j_{20}}$

   **On peut simplifier la notation en notant directement $M_{j_i}$ l'événement $T_i = m_{j_i}$**

   **c2)** Utilisez la règle de multiplication pour exprimer P(E)

   $P(E) = P(M_{j_1})P(M_{j_2} \mid M_{j_1})P(M_{j_3} \mid M_{j_1}M_{j_2})..P(M_{j_{20}} \mid M_{j_1}M_{j_2}..M_{j_{19}})$

   **c1)** Si on fait l'hypothèse simplificatrice que l'occurrence d'un mot ne dépend au plus que des deux mots précédents (s'ils existent), comment pouvez-vous réécrire P(E) ?

   $P(E) = P(M_{j_1})P(M_{j_2} \mid M_{j_1})P(M_{j_3} \mid M_{j_1}M_{j_2})..P(M_{j_{20}} \mid M_{j_{18}}M_{j_{19}})$