Fouille de données

NGUYỄN Thị Minh Huyền ©2016

huyenntm@hus.edu.vn

Plan

Plan

Plan

Expérience stochastique/aléatoire - Evénement

- Ensemble de tous les résultats possibles/univers de l'expérience : ensemble fondamental Ω
- Evénement $A \subset \Omega$.
  - A est réalisé si le résultat $\omega \in A$.
  - $|A| = 1$: événement élémentaire
  - Opérations : $A \cup B$ (ou), $A \cap B$ (et), $\overline{A}$ (événement contraire)
- Incompatibilité : $A \cap B = \emptyset$ (A et B mutuellement exclusifs)

Expérience stochastique/aléatoire - Evénement

- Ensemble de tous les résultats possibles/univers de l'expérience : ensemble fondamental Ω
- Evénement $A \subset \Omega$.
  - A est réalisé si le résultat $\omega \in A$.
  - $|A| = 1$: événement élémentaire
  - Opérations : $A \cup B$ (ou), $A \cap B$ (et), $\overline{A}$ (événement contraire)
- Incompatibilité : $A \cap B = \emptyset$ (A et B mutuellement exclusifs)

Expérience stochastique/aléatoire - Evénement

- Ensemble de tous les résultats possibles/univers de l'expérience : ensemble fondamental Ω
- Evénement $A \subset \Omega$.
  - A est réalisé si le résultat $\omega \in A$.
  - $|A| = 1$: événement élémentaire
  - Opérations : $A \cup B$ (ou), $A \cap B$ (et), $\overline{A}$ (événement contraire)
- Incompatibilité : $A \cap B = \emptyset$ (A et B mutuellement exclusifs)

Probabilité

Espace probabilisé (Ω, P)

- P - loi de probabilité, en accord avec les axiomes :
  - $0 \leq P(A) \leq 1$ pour tout $A \subset \Omega$
  - $P(\Omega) = 1$
  - $P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$ pour toute suite finie d'événements incompatibles deux à deux
  - Si Ω infini, la formule ci-dessus peut être appliquée avec n infini.
- Loi uniforme (discrète ou continue) : tous les événements élémentaires sont équiprobables.
- Définition statistique de la probabilité : répéter l'expérience un grand nombre de fois - $P(A) = n_A/n$

Probabilité

Espace probabilisé (Ω, P)

- P - loi de probabilité, en accord avec les axiomes :
  - $0 \leq P(A) \leq 1$ pour tout $A \subset \Omega$
  - $P(\Omega) = 1$
  - $P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$ pour toute suite finie d'événements incompatibles deux à deux
  - Si Ω infini, la formule ci-dessus peut être appliquée avec n infini.
- Loi uniforme (discrète ou continue) : tous les événements élémentaires sont équiprobables.
- Définition statistique de la probabilité : répéter l'expérience un grand nombre de fois - $P(A) = n_A/n$

Probabilité

Espace probabilisé (Ω, P)

- P - loi de probabilité, en accord avec les axiomes :
  - $0 \leq P(A) \leq 1$ pour tout $A \subset \Omega$
  - $P(\Omega) = 1$
  - $P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$ pour toute suite finie d'événements incompatibles deux à deux
  - Si Ω infini, la formule ci-dessus peut être appliquée avec n infini.
- Loi uniforme (discrète ou continue) : tous les événements élémentaires sont équiprobables.
- Définition statistique de la probabilité : répéter l'expérience un grand nombre de fois - $P(A) = n_A/n$

Probabilité conditionnelle

- $P(A/B) = P(A \cap B)/P(B)$
  - probabilité de l'événement A sachant que B est réalisé,
  - probabilité conditionnelle de A étant donné B
- $P(A_1 \cap A_2 \cap \cdots \cap A_n) =$
  $P(A_1)P(A_2/A_1)P(A_3/A_1 \cap A_2) \cdots P(A_n/A_1 \cap \cdots \cap A_{n-1})$
- Formule de Bayes

$$P(B_k/A) = \frac{P(A \cap B_k)}{P(A)} = \frac{P(A/B_k)P(B_k)}{\sum_{i=1}^{n} P(A/B_i)P(B_i)}$$

  $B_1, \cdots, B_n$ forment une partition de Ω.
- Indépendance stochastique : $P(A/B) = P(A)$ ou $P(B/A) = P(B)$ ou $P(A \cap B) = P(A)P(B)$.

Probabilité conditionnelle

- $P(A/B) = P(A \cap B)/P(B)$
  - probabilité de l'événement A sachant que B est réalisé,
  - probabilité conditionnelle de A étant donné B
- $P(A_1 \cap A_2 \cap \cdots \cap A_n) =$
  $P(A_1)P(A_2/A_1)P(A_3/A_1 \cap A_2) \cdots P(A_n/A_1 \cap \cdots \cap A_{n-1})$
- Formule de Bayes

$$P(B_k/A) = \frac{P(A \cap B_k)}{P(A)} = \frac{P(A/B_k)P(B_k)}{\sum_{i=1}^{n} P(A/B_i)P(B_i)}$$

  $B_1, \cdots, B_n$ forment une partition de Ω.
- Indépendance stochastique : $P(A/B) = P(A)$ ou $P(B/A) = P(B)$ ou $P(A \cap B) = P(A)P(B)$.

Probabilité conditionnelle

- $P(A/B) = P(A \cap B)/P(B)$
  - probabilité de l'événement A sachant que B est réalisé,
  - probabilité conditionnelle de A étant donné B
- $P(A_1 \cap A_2 \cap \cdots \cap A_n) =$
  $P(A_1)P(A_2/A_1)P(A_3/A_1 \cap A_2) \cdots P(A_n/A_1 \cap \cdots \cap A_{n-1})$
- Formule de Bayes

$$P(B_k/A) = \frac{P(A \cap B_k)}{P(A)} = \frac{P(A/B_k)P(B_k)}{\sum_{i=1}^{n} P(A/B_i)P(B_i)}$$

  $B_1, \cdots, B_n$ forment une partition de Ω.
- Indépendance stochastique : $P(A/B) = P(A)$ ou $P(B/A) = P(B)$ ou $P(A \cap B) = P(A)P(B)$.

Probabilité conditionnelle

- $P(A/B) = P(A \cap B)/P(B)$
  - probabilité de l'événement A sachant que B est réalisé,
  - probabilité conditionnelle de A étant donné B
- $P(A_1 \cap A_2 \cap \cdots \cap A_n) =$
  $P(A_1)P(A_2/A_1)P(A_3/A_1 \cap A_2) \cdots P(A_n/A_1 \cap \cdots \cap A_{n-1})$
- Formule de Bayes

$$P(B_k/A) = \frac{P(A \cap B_k)}{P(A)} = \frac{P(A/B_k)P(B_k)}{\sum_{i=1}^{n} P(A/B_i)P(B_i)}$$

  $B_1, \cdots, B_n$ forment une partition de Ω.
- Indépendance stochastique : $P(A/B) = P(A)$ ou $P(B/A) = P(B)$ ou $P(A \cap B) = P(A)P(B)$.

Variables aléatoires

- $X(\Omega)$ fonction à valeurs réelles, discrètes ou continues
  Variable aléatoire à plusieurs dimensions : vecteur aléatoire
- Variables aléatoires discrètes : ensemble de valeurs fini ou dénombrable
  - Distribution de X : $P(X = x_k) = p_k$, $k = 1, 2, \cdots$
- Variables aléatoires continues
  - $f(x)$ fonction de densité de la variable aléatoire X : $f(x) \geq 0$, $\int_{-\infty}^{+\infty} f(x)dx = 1$
  - $P(u \leq X \leq v) = \int_u^v f(x)dx$ (surface sous la courbe de $f(x)$)
- Fonction de répartition $F(x) = P(X \leq x)$

Variables aléatoires

- $X(\Omega)$ fonction à valeurs réelles, discrètes ou continues
  Variable aléatoire à plusieurs dimensions : vecteur aléatoire
- Variables aléatoires discrètes : ensemble de valeurs fini ou dénombrable
  - Distribution de X : $P(X = x_k) = p_k$, $k = 1, 2, \cdots$
- Variables aléatoires continues
  - $f(x)$ fonction de densité de la variable aléatoire X : $f(x) \geq 0$, $\int_{-\infty}^{+\infty} f(x)dx = 1$
  - $P(u \leq X \leq v) = \int_u^v f(x)dx$ (surface sous la courbe de $f(x)$)
- Fonction de répartition $F(x) = P(X \leq x)$

Variables aléatoires

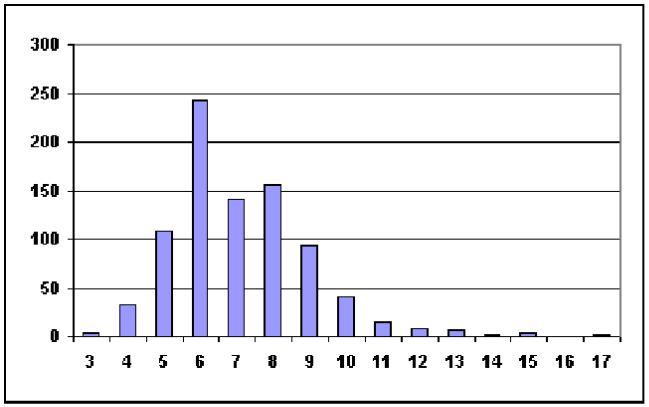
- $X(\Omega)$ fonction à valeurs réelles, discrètes ou continues
  Variable aléatoire à plusieurs dimensions : vecteur aléatoire
- Variables aléatoires discrètes : ensemble de valeurs fini ou dénombrable
  - Distribution de X : $P(X = x_k) = p_k$, $k = 1, 2, \cdots$
- Variables aléatoires continues
  - $f(x)$ fonction de densité de la variable aléatoire X :
    $f(x) \geq 0$, $\int_{-\infty}^{+\infty} f(x)dx = 1$
  - $P(u \leq X \leq v) = \int_u^v f(x)dx$ (surface sous la courbe de $f(x)$)
- Fonction de répartition $F(x) = P(X \leq x)$

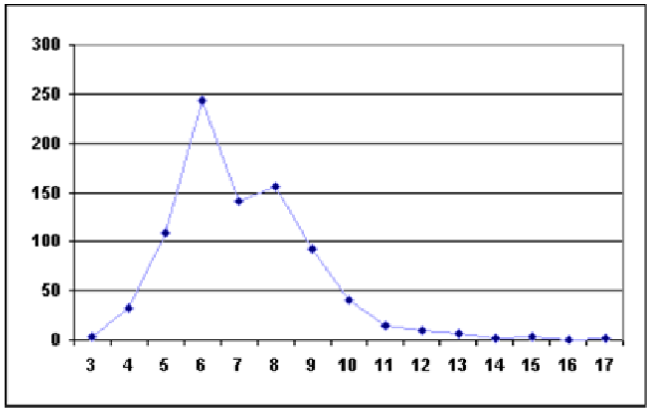
Variables aléatoires

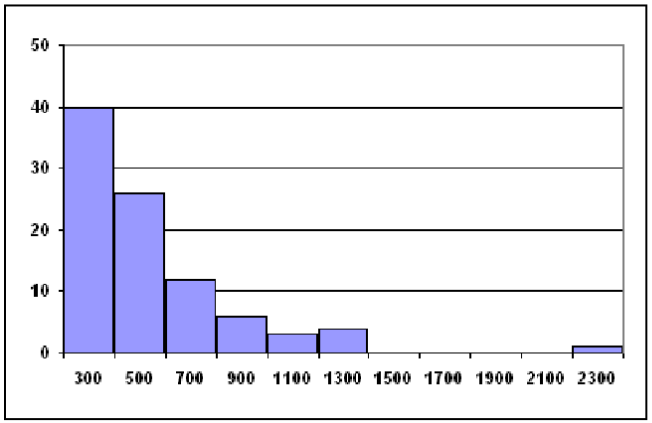
- $X(\Omega)$ fonction à valeurs réelles, discrètes ou continues
  Variable aléatoire à plusieurs dimensions : vecteur aléatoire
- Variables aléatoires discrètes : ensemble de valeurs fini ou dénombrable
  - Distribution de X : $P(X = x_k) = p_k$, $k = 1, 2, \cdots$
- Variables aléatoires continues
  - $f(x)$ fonction de densité de la variable aléatoire X :
    $f(x) \geq 0$, $\int_{-\infty}^{+\infty} f(x)dx = 1$
  - $P(u \leq X \leq v) = \int_u^v f(x)dx$ (surface sous la courbe de $f(x)$)
- Fonction de répartition $F(x) = P(X \leq x)$

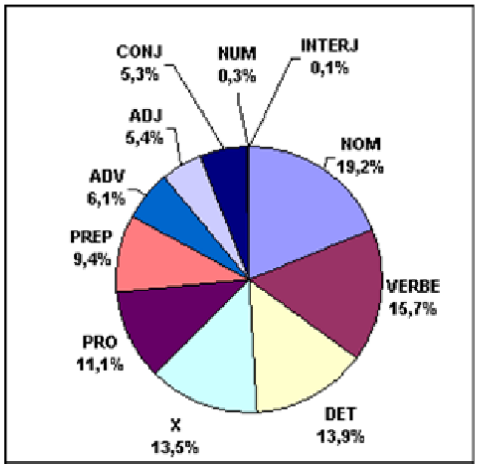
Représentation graphique des distributions

Diagramme en bâtons (ou en tuyau d'orgue), histogramme, polygone, ou encore diagramme en secteurs (camembert).









(extrait du cours d'Informatique et Statistique de Jean Véronis)

Espérance mathématique (moyenne) et variance

- Variable aléatoire discrète :
  - Espérance math.

$$E(X) = \mu = \sum_k x_k p_k$$

  - Variance

$$Var(X) = \sigma^2 = \sum_k (x_k - \mu)^2 p_k = \sum_k x_k^2 p_k - \mu^2$$

σ écart-type

- Variable aléatoire continue :
  - Espérance math.

$$E(X) = \mu = \int_{-\infty}^{+\infty} x f(x) dx$$

  - Variance

$$Var(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{+\infty} x^2 f(x) dx - \mu^2$$

Espérance mathématique (moyenne) et variance

- Variable aléatoire discrète :
  - Espérance math.

$$E(X) = \mu = \sum_k x_k p_k$$

  - Variance

$$Var(X) = \sigma^2 = \sum_k (x_k - \mu)^2 p_k = \sum_k x_k^2 p_k - \mu^2$$

σ écart-type

- Variable aléatoire continue :
  - Espérance math.

$$E(X) = \mu = \int_{-\infty}^{+\infty} x f(x) dx$$

  - Variance

$$Var(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{+\infty} x^2 f(x) dx - \mu^2$$

Distribution binomiale

- Distribution de Bernoulli : variable à deux valeurs (modalités), notées 0 (échec) et 1 (succès)
  - $P(X = 1) = p$, $P(X = 0) = q = 1 - p$.
  - $\mu = p$
  - $\sigma^2 = p(1 - p)$
- Distribution binomiale : nombre de succès rencontrés en effectuant n répétitions d'expérience de Bernoulli $B(n, p)$
  - $B(k; n, p) = P(X = k) = C_n^k p^k (1 - p)^{n-k}$
  - $\mu = np$
  - $\sigma^2 = np(1 - p)$

Distribution binomiale

- Distribution de Bernoulli : variable à deux valeurs (modalités), notées 0 (échec) et 1 (succès)
  - $P(X = 1) = p$, $P(X = 0) = q = 1 - p$.
  - $\mu = p$
  - $\sigma^2 = p(1 - p)$
- Distribution binomiale : nombre de succès rencontrés en effectuant n répétitions d'expérience de Bernoulli $B(n, p)$
  - $B(k; n, p) = P(X = k) = C_n^k p^k (1 - p)^{n-k}$
  - $\mu = np$
  - $\sigma^2 = np(1 - p)$

Distribution de Poisson

- Loi de probabilité notée $P(\lambda)$: $P(X = k) = e^{-\lambda}\lambda^k/k$
- $\mu = \lambda$, $\sigma^2 = \lambda$
  - Souvent utilisée pour décrire le nombre de réalisations d'un événement dans un intervalle de temps donné t, sachant le nombre moyen de réalisations α par unité de temps ($\lambda = \alpha t$);
  - Pour $\lambda \leq 10$, on utilise une table pour consulter les probabilités ;
  - Pour $\lambda > 10$, X obéit approximativement à une loi normale.

Distribution de Poisson

- Loi de probabilité notée $P(\lambda)$: $P(X = k) = e^{-\lambda}\lambda^k/k$
- $\mu = \lambda$, $\sigma^2 = \lambda$
  - Souvent utilisée pour décrire le nombre de réalisations d'un événement dans un intervalle de temps donné t, sachant le nombre moyen de réalisations α par unité de temps ($\lambda = \alpha t$);
  - Pour $\lambda \leq 10$, on utilise une table pour consulter les probabilités ;
  - Pour $\lambda > 10$, X obéit approximativement à une loi normale.

Distribution de Poisson

- Loi de probabilité notée $P(\lambda)$: $P(X = k) = e^{-\lambda}\lambda^k/k$
- $\mu = \lambda$, $\sigma^2 = \lambda$
  - Souvent utilisée pour décrire le nombre de réalisations d'un événement dans un intervalle de temps donné t, sachant le nombre moyen de réalisations α par unité de temps ($\lambda = \alpha t$);
  - Pour $\lambda \leq 10$, on utilise une table pour consulter les probabilités ;
  - Pour $\lambda > 10$, X obéit approximativement à une loi normale.

Distribution exponentielle

- Densité de probabilité :

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{ailleurs} \end{cases}$$

- $\mu = 1/\lambda$
- $\sigma^2 = 1/\lambda^2$
  - Souvent utilisée pour décrire le temps entre deux réalisations successives d'un événement suivant le processus Poisson ;

Distribution exponentielle

- Densité de probabilité :

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{ailleurs} \end{cases}$$

- $\mu = 1/\lambda$
- $\sigma^2 = 1/\lambda^2$
  - Souvent utilisée pour décrire le temps entre deux réalisations successives d'un événement suivant le processus Poisson ;

Distribution normale

- Loi normale (gaussienne) réduite/standard $N(0,1)$ (moyenne = 0, variance = 1)

$$f(x) = \frac{1}{\sqrt{2\pi}} exp[-x^2/2] \qquad (-\infty < x < \infty)$$

- Loi normale (gaussienne) $N(\mu, \sigma^2)$ (moyenne = μ, variance = σ^2)
  $\Rightarrow$ variable aléatoire $Y = \sigma X + \mu$, où X est une variable normale réduite

Distribution normale

- Loi normale (gaussienne) réduite/standard $N(0,1)$ (moyenne = 0, variance = 1)

$$f(x) = \frac{1}{\sqrt{2\pi}} exp[-x^2/2] \qquad (-\infty < x < \infty)$$

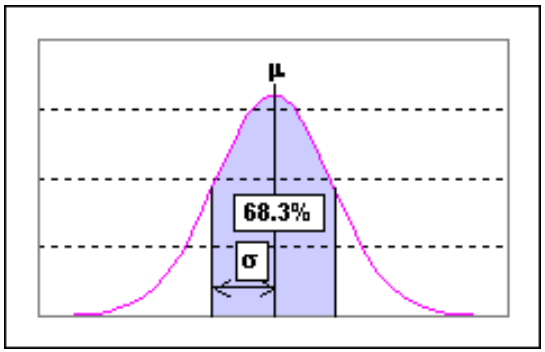
- Loi normale (gaussienne) $N(\mu, \sigma^2)$ (moyenne = μ, variance = σ^2)
  $\Rightarrow$ variable aléatoire $Y = \sigma X + \mu$, où X est une variable normale réduite

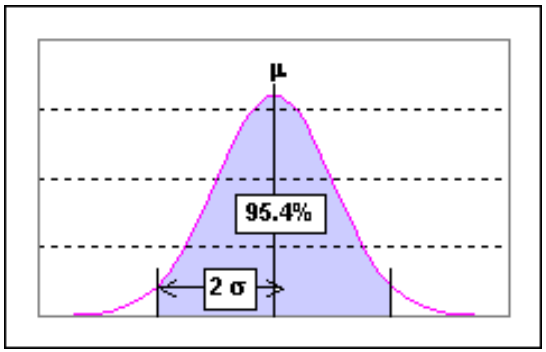
Distribution normale

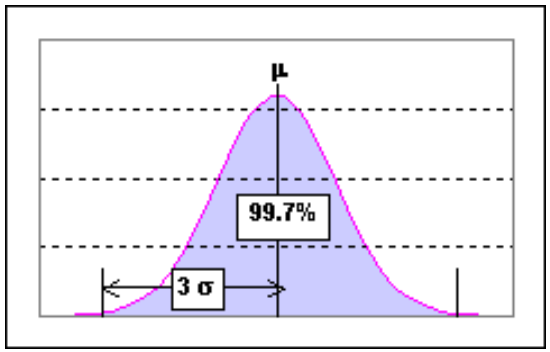
- Loi normale (gaussienne) réduite/standard $N(0,1)$ (moyenne = 0, variance = 1)

$$f(x) = \frac{1}{\sqrt{2\pi}} exp[-x^2/2] \qquad (-\infty < x < \infty)$$

- Loi normale (gaussienne) $N(\mu, \sigma^2)$ (moyenne = μ, variance = σ^2)
  $\Rightarrow$ variable aléatoire $Y = \sigma X + \mu$, où X est une variable normale réduite







Statistique descriptive et inférentielle

- Stat. descriptive : explorer les données, en tirer un certain nombre de mesures et d'indices, ou des représentations graphiques
  faire appraître des hypothèses
- Stat. inférentielle : tester des hypothèses,
  faire des prédictions à partir des données.

Statistique descriptive et inférentielle

- Stat. descriptive : explorer les données, en tirer un certain nombre de mesures et d'indices, ou des représentations graphiques
  faire appraître des hypothèses
- Stat. inférentielle : tester des hypothèses,
  faire des prédictions à partir des données.

Concepts généraux

- Variable (attribut, caractère) : propriété d'un ensemble d'objets ou d'événements à étudier
- Domaine d'une variable : ensemble de modalités ou valeurs
- Echelles de mesure : var. nominales (catégorielle), ordinales ou numériques (discrètes/continues)
- Population : ensemble de tous les objets ou événements qu'on veut étudier
  $\Rightarrow$ paramètres à estimer
- Echantillon : sous-ensemble permettant d'estimer une propriété de la population
  observations permettant de tester des hypothèses

Concepts généraux

- Variable (attribut, caractère) : propriété d'un ensemble d'objets ou d'événements à étudier
- Domaine d'une variable : ensemble de modalités ou valeurs
- Echelles de mesure : var. nominales (catégorielle), ordinales ou numériques (discrètes/continues)
- Population : ensemble de tous les objets ou événements qu'on veut étudier
  $\Rightarrow$ paramètres à estimer
- Echantillon : sous-ensemble permettant d'estimer une propriété de la population
  observations permettant de tester des hypothèses

Concepts généraux

- Variable (attribut, caractère) : propriété d'un ensemble d'objets ou d'événements à étudier
- Domaine d'une variable : ensemble de modalités ou valeurs
- Echelles de mesure : var. nominales (catégorielle), ordinales ou numériques (discrètes/continues)
- Population : ensemble de tous les objets ou événements qu'on veut étudier
  $\Rightarrow$ paramètres à estimer
- Echantillon : sous-ensemble permettant d'estimer une propriété de la population
  observations permettant de tester des hypothèses

Concepts généraux

- Variable (attribut, caractère) : propriété d'un ensemble d'objets ou d'événements à étudier
- Domaine d'une variable : ensemble de modalités ou valeurs
- Echelles de mesure : var. nominales (catégorielle), ordinales ou numériques (discrètes/continues)
- Population : ensemble de tous les objets ou événements qu'on veut étudier
  $\Rightarrow$ paramètres à estimer
- Echantillon : sous-ensemble permettant d'estimer une propriété de la population
  observations permettant de tester des hypothèses

Plan

Plan

Plan

Plan