Fouille de données

Cours 2 - Exploration des données : cas d'une et de deux dimensions

NGUYỄN Thị Minh Huyền
Email: huyenntm@vnu.edu.vn
VNU University of Science, Hanoi

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)

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

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

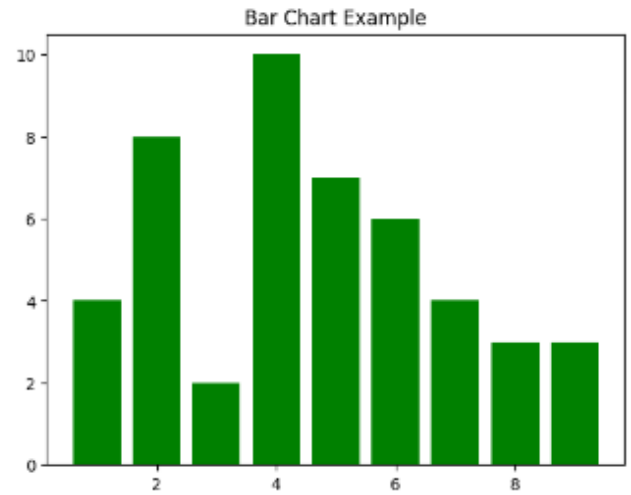
  $B_1, \cdots, B_n$ forment une partition de Ω.
- Indépendance stochastique : $P(A|B) = P(A)$ ou $P(B|A) = P(B)$ ou $P(A \cap B) = P(A)P(B)$.

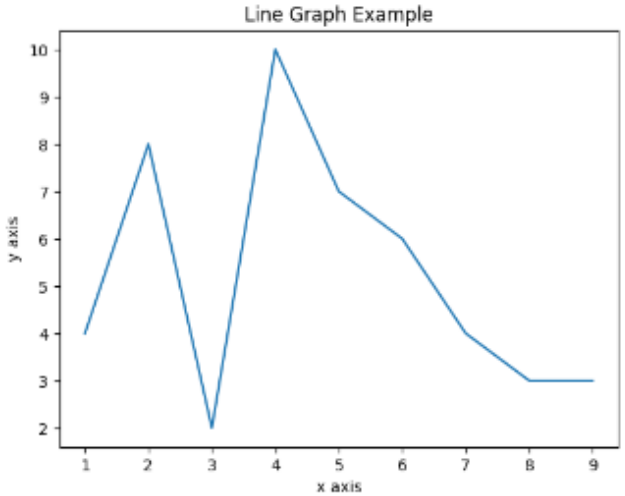
Variables aléatoires

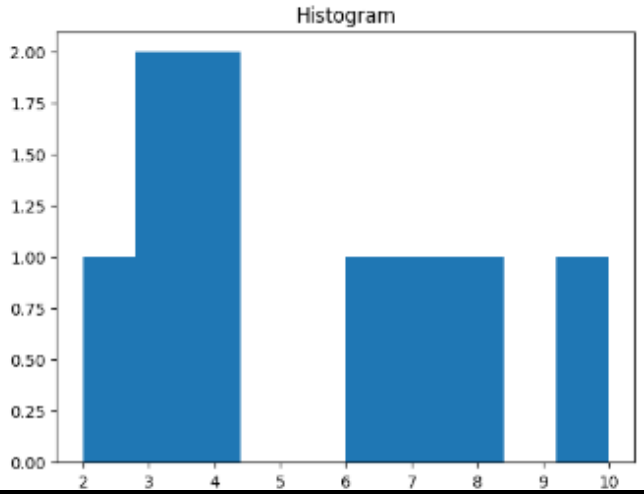
- $X(\Omega)$ fonction à valeurs réelles, discrètes ou continues
  Variable aléatoire à plusieurs dimensions : vecteur aléatoire
- Variables aléatoires discrètes : ensemble de valeurs fini ou dénombrable
  - Distribution de X : $P(X = x_k) = p_k$, $k = 1, 2, \cdots$
- Variables aléatoires continues
  - $f(x)$ 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 $F(x) = P(X \leq x)$

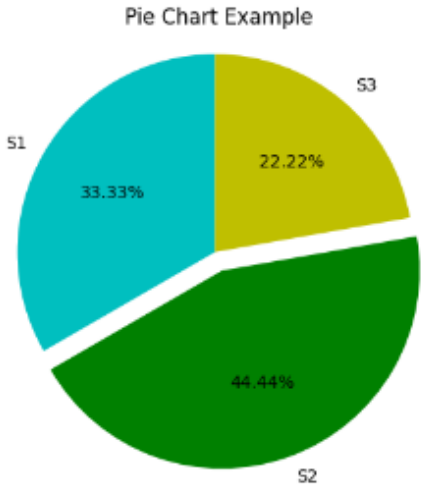
Représentation graphique des distributions

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









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 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 :

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

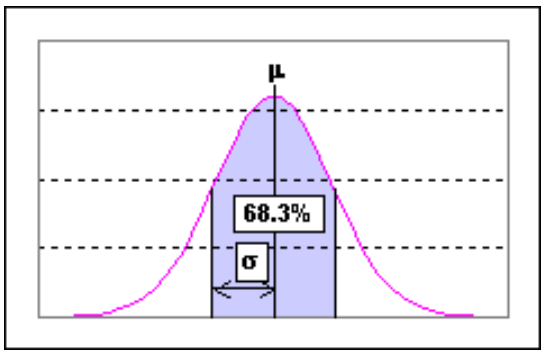
- $\mu = 1/\lambda$
- $\sigma^2 = 1/\lambda^2$
  - Souvent utilisée pour décrire le temps entre deux réalisations successives d'un événement suivant le processus Poisson ;

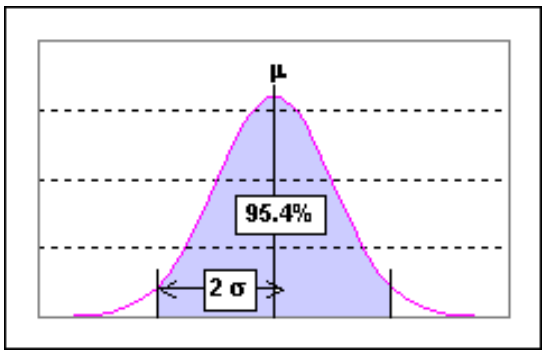
Distribution normale

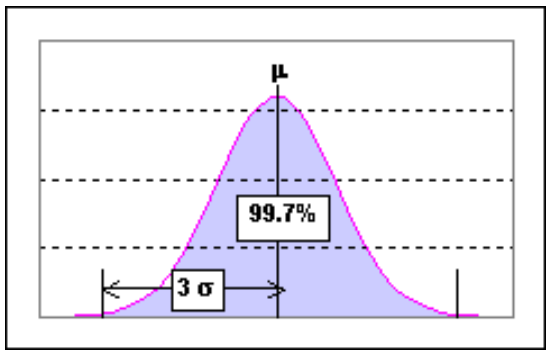
- Loi normale (gaussienne) réduite/standard $N(0,1)$ (moyenne = 0, variance = 1)

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

- 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.

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

Préparation de données

- Type de données
- Nettoyage du fichier (qualité des données)
- Distribution des variables
- Détection de valeurs aberrantes, extrêmes, rares, manquantes... et traitement
- Caractérisation des variables
- Création de nouvelles variables, transformation de variables

Exemple

- Exemple du Rola Cola de B.L. BOWERMAN / R.T. O'CONNELL (données fournies sur la page de Tenenhaus)
- Objectif: le département Marketing de Rola-Cola souhaite étudier les attitudes et les préférences des consommateurs envers Rola-Cola par rapport à Koca-Cola : pour cela, on réalise un test de goût avec les deux boissons avec des clients choisis au hasard.

Questions

1. Quelle boisson préférez-vous ?
   - Rola-Cola
   - Koka-Cola
2. Avez-vous déjà acheté Rola-Cola ?
   - Oui
   - Non
3. Entourez la réponse décrivant au mieux votre réaction à la phrase : J'aime mes boissons au Cola sucrées
   - D'accord
   - Je ne suis pas sûr
   - Pas d'accord
4. Combien de litres de boisson au Cola votre famille a-t-elle consommée au cours du mois dernier ?
5. Combien de paquets de chips avez-vous consommé le mois dernier ?

Données

- Fichier rola_cola.xls
- Echantillon : $n = 40$ personnes
- Codage :
  - Boisson préférée :
    1 = Rola-Cola 2 = Koka-Cola
  - Achat préalable :
    1 = oui 2 = non
  - Goût sucre :
    1 = oui 2 = indifférent 3 = non

Représentation de données

- Tableau
- Graphiques : diagramme circulaire (en secteurs), diagramme en bâtons, polygone de fréquence, histogramme, etc.

Etude d'une variable qualitative

- Etude d'une proportion
- Exemple : Boisson préférée entre Rola-Cola et Koca-Cola
  Feuille rola_cola.Proportion1

Etude d'une variable quantitative (numérique)

- Une variable numérique X prend des valeurs réelles $x_1, \cdots, x_i, \cdots, x_N$ sur une population et $x_1, \cdots, x_i, \cdots, x_n$ sur un échantillon.
- Elle est résumée par des indicateurs statistiques :
  - Tendance centrale : moyenne, médiane, mode
  - Dispersion : étendues, écart-type, écart absolu moyen à la médiane...
  - Forme :
    Asymétrie (coefficient d'asymétrie : 0 - symétrique, > 0 - étalée à gauche , < 0 - étalée à droite)
    Aplatissement (coefficient d'aplatissement ou kurtosis : = 0 distribution normale, > 0 - concentration élevée, < 0 - concentration faible)

Tendance centrale : Mode, médiane

- Mode : valeur qui apparaît le plus fréquemment
- Médiane : M divise l'échantillon ordonné $x_1 \leq x_2 \leq \cdots \leq x_n$ en 2 parties égales
  - $n = 2k + 1 : M = x_{k+1}$
  - $n = 2k : M = (x_k + x_{k+1})/2$

Tendance centrale et dispersion : Moyenne et écart-type

	Population	Echantillon
Effectif	N	n
Moyenne	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$
Ecart-type	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

- $\overline{x}$ - estimation de μ,
- s^2 - estimation de σ^2.

Dispersion : Etendue, Quantiles

- Etendue = max - min
- Notion : Division de l'échantillon ordonné en n parties égales (quantiles)
  - $n = 4 \Rightarrow$ Quartiles Q_1, Q_2, Q_3 : charnières entre quatre parties.
    $Q_2 = M$, $Q_3 - Q_1$: étendue interquartile
  - $n = 10 \Rightarrow$ Déciles $D_1, \cdots, D_9$
    $D_9 - D_1$: étendue interdécile
  - $n = 100 \Rightarrow$ Centiles

Représentation graphique

- Proportion de la variable X : Diagrammes (en tuyaux d'orgue, en secteurs, en tige et feuilles), histogramme
- La dispersion de X est visualisée par la boîte-à-moustaches et l'histogramme.
  Boîte à moustaches : minimum, $[D_1]$, Q_1, médiane, Q_3, $[D_9]$, maximum
  $\Rightarrow$ aider à visualiser des valeurs extrêmes.

Exemple

Cas Rola-Cola

- Etude de la variable numérique : Comsommation de boisson au cola
- Statistiques et représentations graphiques
- Feuille rola_cola.Proportion2

Détection des observations atypiques (*Outliers*)

- La longueur de chaque moustache doit être inférieure à $1,5(Q_3 - Q_1)$.

Etude du lien entre deux variables

- 2 variables X et Y
  - X : variable explicative
  - Y : variable à expliquer
- 2 variables quantitatives : régression simple, corrélation simple
- 2 variables qualitatives : Test du khi-deux d'indépendance
- X quantitative, Y qualitative : régression logistique
- X qualitative, Y quantitative : analyse de la variance à un facteur

Deux variables quantitatives : nuage de points

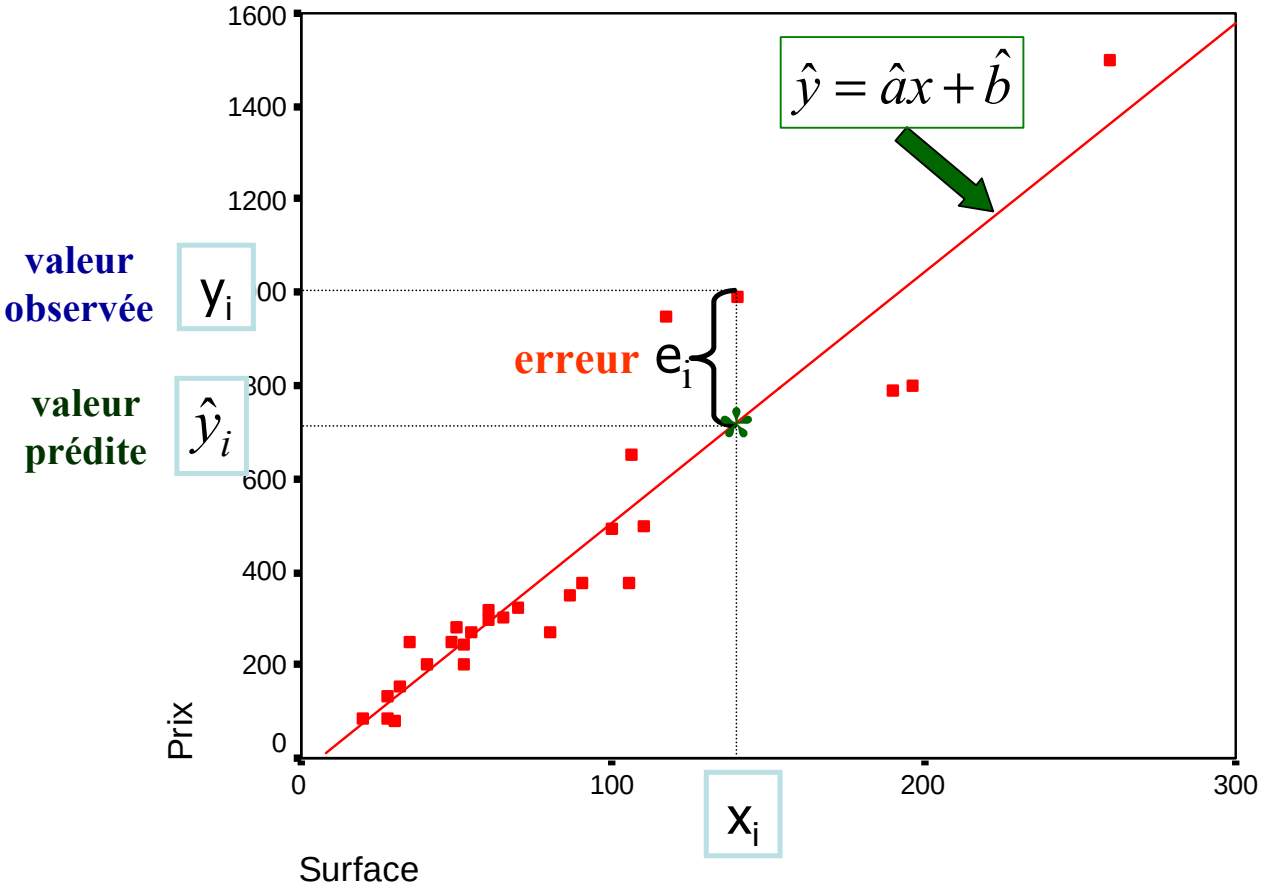
- Diagramme de dispersion
- Coefficient de corrélation
- Eventuellement, si cela a un sens, droite d'ajustement (des moindres carrés)

Données

- Y : variable à expliquer numérique (dépendante)
- X : variable explicative numérique ou binaire (indépendante)
- Tableau de données

	X	Y
1	x_1	y_1
⋮	⋮	⋮
i	x_i	y_i
⋮	⋮	⋮
n	x_n	y_n

La droite des moindres carrés



On cherche

$\hat{a}$ et $\hat{b}$

minimisant

$$\sum_{i=1}^{n} e_i^2$$

Coefficient de détermination R^2, coefficient de corrélation $Cor(X, Y)$

- Formule de décomposition :

$$\sum (y_i - \overline{y})^2 = \sum (\hat{y}_i - \overline{y})^2 + \sum e_i^2$$

- Coefficient de détermination :

$$R^2 = \frac{\sum (\hat{y}_i - \overline{y})^2}{\sum (y_i - \overline{y})^2}$$

- Coefficient de corrélation :

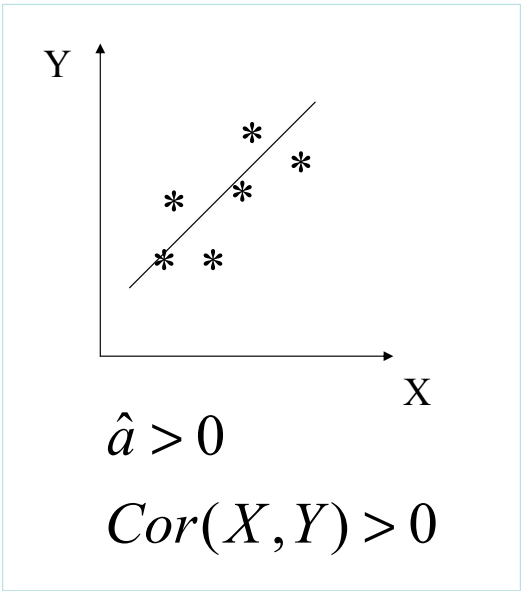
$$Cor(X, Y) = \text{sign}(\hat{a})\sqrt{R^2}$$

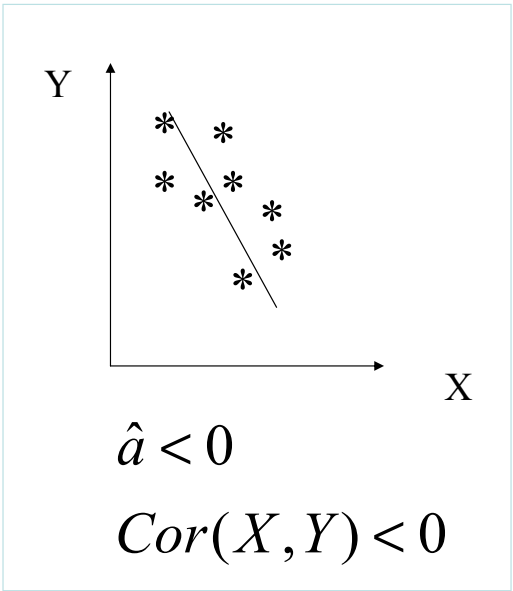
Corrélation entre deux variables : calcul direct de $Cor(X, Y)$

- Mesure la force et le sens de la liaison linéaire entre les deux variables numériques

$$Cor(X, Y) = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}}$$

- Toujours compris entre -1 et 1
- $Cor(X, Y) = 0$: X et Y non corrélées





Rappel sur les tests d'hypothèses

- Test d'hypothèses : raisonnement par l'absurde
- Hypothèse nulle H_0 : hypothèse inverse
- Objectif : calculer le degré de confiance en rejettant l'hypothèse nulle.

La corrélation $Cor(X, Y)$ est-elle significative au risque $\alpha = 0.05$?

- Notations
  - $\rho =$ corrélation au niveau de la population
  - $Cor(X, Y) =$ corrélation au niveau de l'échantillon
- Test :
  - $H_0 : \rho = 0$
  - $H_1 : \rho \neq 0$
- Règle de décision : On rejette H_0 au risque $\alpha = 0.05$ de se tromper si

$$|Cor(X, Y)| \geq \frac{2}{\sqrt{n}}$$

(Bonne approximation pour $n > 20$)

La corrélation $Cor(X, Y)$ est-elle significative au risque α ?

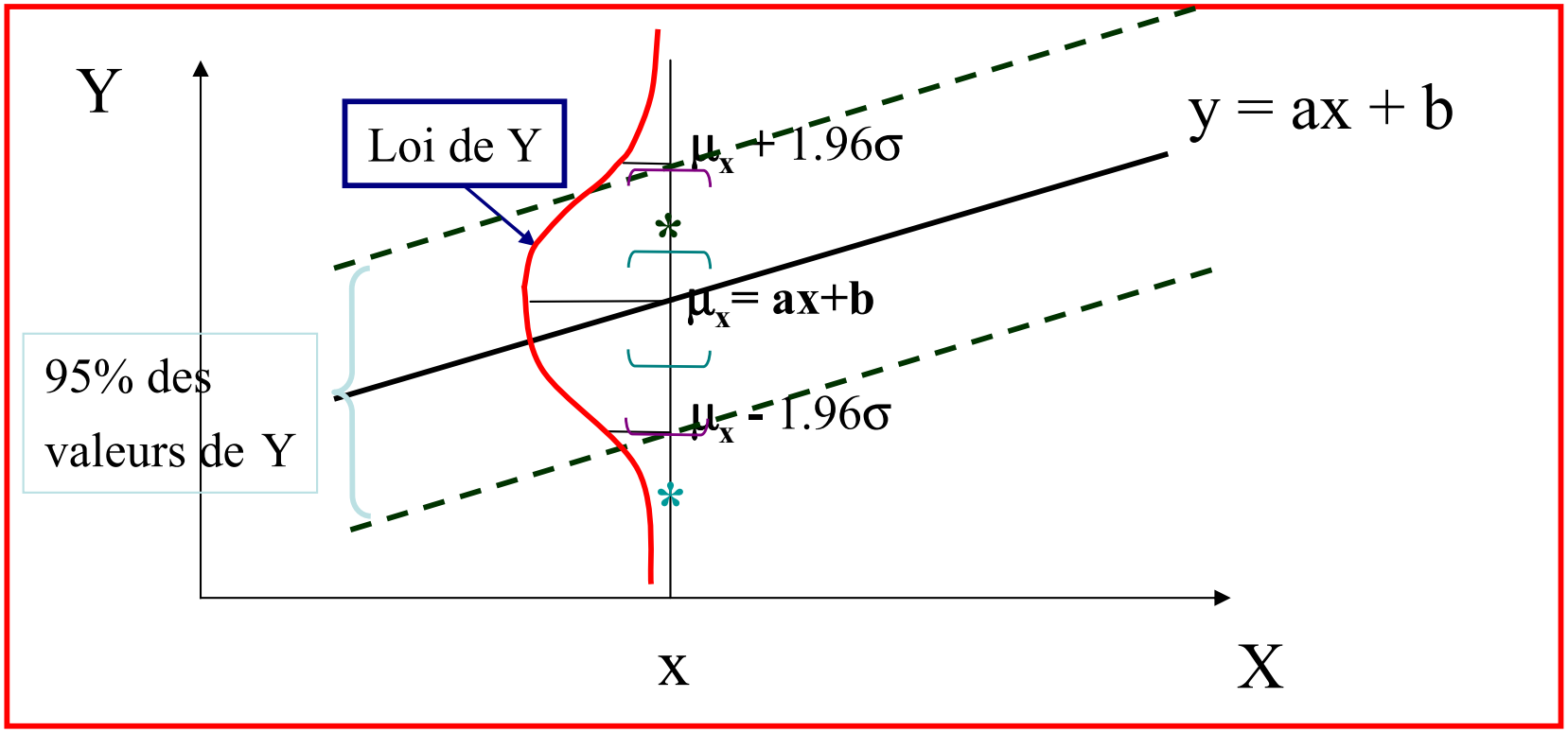
- Notations
  - $\rho =$ corrélation au niveau de la population
  - $Cor(X, Y) =$ corrélation au niveau de l'échantillon
- Test :
  - $H_0 : \rho = 0$
  - $H_1 : \rho \neq 0$
- Règle de décision : On rejette H_0 au risque α de se tromper si

$$|Cor(X, Y)| \geq \frac{t_{1-\alpha/2}(n-2)}{\sqrt{t^2_{1-\alpha/2}(n-2) + n - 2}}$$

- Plus petit α conduisant au rejet de H_0.

Modèle de la régression simple

- Modèle : $Y = aX + b + \epsilon$, avec $\epsilon \sim N(0, \sigma)$.



L'écart-type σ représente à peu près le quart de l'épaisseur du nuage.

Estimation de a, b et σ

- Estimation de a et b :
  - $\hat{a}$ = estimation de a ($\hat{a} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$)
  - $\hat{b}$ = estimation de b ($\hat{b} = \overline{y} - \hat{a}.\overline{x}$)
- Estimation de σ :
  - $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 =$ estimation de σ^2
  - $\hat{\sigma} = \sqrt{\hat{\sigma}^2} =$ estimation de σ

X et Y qualitatives

- On s'intéresse à l'indépendance entre les deux variables
- $\Rightarrow$ Test khi-deux (χ^2) de l'indépendance

Tableau de contingence

	1		j		p	
1	k_{11}		k_{1j}		k_{1p}	$k_{1.}$
i	k_{i1}		k_{ij}		k_{ip}	$k_{i.}$
n	k_{n1}		k_{nj}		k_{np}	$k_{n.}$
	$k_{.1}$		$k_{.j}$		$k_{.p}$	$k = \sum k_{ij}$

Tableau de fréquences

$$f_{ij} = \frac{k_{ij}}{k}$$

	1		j		p	
1	f_{11}		f_{1j}		f_{1p}	$f_{1.}$
i	f_{i1}		f_{ij}		f_{ip}	$f_{i.}$
n	f_{n1}		f_{nj}		f_{np}	$f_{n.}$
	$f_{.1}$		$f_{.j}$		$f_{.p}$	f

Lien entre deux variables

- Visualiser les associations entre les modalités des deux variables
- Tester l'indépendance entre les lignes et les colonnes
  - On observe k_{ij} ($k_{i.} = \sum_j k_{ij}$, $k_{.j} = \sum_i k_{ij}$, $k = \sum_{ij} k_{ij}$)
  - Si les variables sont indépendantes, alors $k_{ij}/k_{i.} = k_{.j}/k$ quel que soit i et $k_{ij}/k_{.j} = k_{i.}/k$ quel que soit j
  - Les $k_{ij}/k_{i.}$ sont appelés les profils lignes (il y en a autant que de lignes) et les $k_{ij}/k_{.j}$ les profils colonnes.
  - Sous l'hypothèse d'indépendance, $k_{ij} = k_{i.} * k_{.j}/k$

Comment étudier l'indépendance

- Examen des profils lignes ou colonnes
- Etude des d_{ij} = rapport observé/théorique $= k_{ij}/(k_{i.} * k_{.j}/k)$
- Statistique du χ^2 :

$$\chi^2 = \sum_{i,j} \frac{(k_{ij} - (k_{i.}k_{.j}/k))^2)}{k_{i.}k_{.j}/k}$$

A comparer à une valeur tabulée dans la table du khi-deux à $(n-1)(p-1)$ degrés de liberté.

Exemple

- Fichier Excel/Open Office Calc alcool.xls

X qualitative et Y quantitative

- Analyse de la variance (il faut que les écart-types soient les mêmes dans chaque groupe) - ANOVA
- De façon intuitive, si la variabilité entre groupes > la variabilité au sein d'un même groupe, on aura tendance à conclure que Y dépend des groupes. Si Y varie autant au sein d'un groupe qu'entre groupes, alors on aura tendance à conclure que X ne semble pas expliquer cette variabilité.
- L'ANOVA va permettre de fixer la limite (en fonction d'un risque α) à partir de laquelle on considère l'effet des groupes comme significatif.

ANOVA

- X définit k échantillons, dans chaque échantillon : n_i - effectif, $\bar{y}_i$ - moyenne, s_i - écart-type
- Global $n = \sum_{i=1}^{k} n_i$; moyenne générale $\bar{y} = \sum_{i=1}^{k} n_i \bar{y}_i / n$
- Y_i : variable Y sur la population i, chaque Y_i suit une loi normale $N(\mu_i, \sigma)$
- Somme des carrés intra-groupe : $\text{ssw} = \sum_{i=1}^{k} (n_i - 1) s_i^2 / (n - k)$
- Somme des carrés inter-groupes : $\text{ssb} = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2 / (k - 1)$
- Rapport de corrélation : $\eta^2 = \text{ssb}/\text{ssw} = F$
- F-Test : $F \geq F_{1-\alpha}(k - 1, n - k)$
- Exemple : fichier MS Excel/Open Office Calc iris.xls

X quantitative et Y qualitative

Régression logistique

- Valeurs de la variable à prédire Y sont binaires (0 ou 1)
- Au lieu de prédire la valeur de Y, on prédit $P(Y = 0|X)$ ou $P(Y = 1|X)$.
- Les probabilités décrivent une sigmoïde (courbe en forme de S) entre 0 et 1
- $$P(Y = 1|x_1, x_2, \cdots, x_k) = \frac{e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}}$$
- β_i à estimer par des programmes (utilisant des méthodes comme MLE - Maximum Likelihood Estimate ou Newton-Raphson)
- $\beta_i = 0$: pas d'effet sur la chance de succès, $\beta_i > 0$: augmente la chance, $\beta_i < 0$: décroît la chance

Résumé - Objectifs

- Préparation et exploration des données
- Nettoyage des données
  - Valeurs extrêmes : transformation, élimination ?
  - Valeurs manquantes : élimination, remplacement (valeur moyenne, régression) ?
- Etape très importante (conditionne la fiabilité de la suite).
- Ce cours : cas d'une ou deux variables
- Cas de plus de 2 variables : cours suivant.

Logiciels de Fouille de données

- Gratuits : Tanagra, Weka, R, Python, etc.
- Payants : SAS, SPSS, S-Plus, etc.

Travail à faire

- Travail en groupe
- Exploration de votre jeu de données avec Tanagra ou un autre outil (Python, R, Weka, ...)