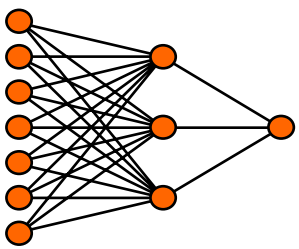
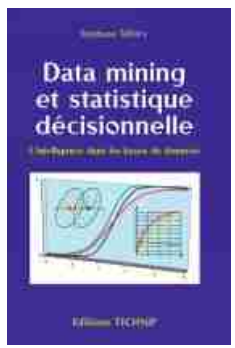
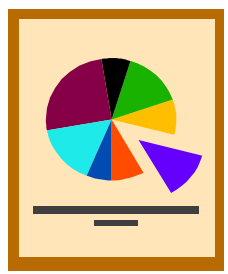
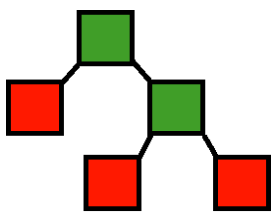
Stéphane Tufféry

# DATA MINING
## & STATISTIQUE DÉCISIONNELLE







# Plan du cours

- Qu'est-ce que le data mining ?
- À quoi sert le data mining ?
- Les 2 grandes familles de techniques
- Le déroulement d'un projet de data mining
- Coûts et gains du data mining
- Facteurs de succès - Erreurs - Consulting
- *L'analyse et la préparation des données*
- Techniques descriptives de data mining
- Techniques prédictives de data mining
- Logiciels de statistique et de data mining
- Informatique décisionnelle et de gestion
- CNIL et limites légales du data mining
- Le web mining
- Le text mining

# La préparation des données :
# Analyse exploratoire

# Les différents formats de données

- Données *continues* (ou *d'échelle*)
  - dont les valeurs forment un sous-ensemble infini de **R** (exemple : salaire)
- Données *discrètes*
  - dont les valeurs forment un sous-ensemble fini ou infini de **N** (exemple : nombre d'enfants)
- Données *catégorielles* (ou *qualitatives*)
  - dont l'ensemble des valeurs est fini — ces valeurs sont numériques ou alphanumériques, mais quand elles sont numériques, ce ne sont que des codes et non des quantités (ex : PCS, n° de département)
- Données *textuelles*
  - lettres de réclamation, rapports, dépêches AFP…

# Précisions sur les formats

- Les données *continues* et *discrètes* sont des quantités :
  - on peut effectuer sur elles des opérations arithmétiques
  - elles sont ordonnées (on peut les comparer par la relation d'ordre <)
- Les données *catégorielles* ne sont pas des quantités
  - mais sont parfois ordonnées : on parle de données catégorielles ***ordinales*** (exemple : « faible, moyen, fort »)
  - données ordinales souvent traitées comme données discrètes
  - les données catégorielles ***nominales*** ne sont pas ordonnées
- Les données *textuelles* contiennent :
  - des abréviations
  - des fautes d'orthographe ou de syntaxe
  - des ambiguïtés (termes dont le sens dépend d'un contexte non facilement détectable automatiquement)

# Méthodes et formats gérés

- La *régression linéaire* traite les variables continues
- L'*analyse discriminante* traite les variables explicatives continues et les variables « cible » nominales
- La *régression logistique* traite les variables explicatives continues, binaires ou nominales, et les variables « cible » nominales ou ordinales
- Les *réseaux de neurones* traitent de préférence les variables continues dans [0,1]
- Certains *arbres de décision* (CHAID) traitent directement les variables discrètes et catégorielles mais discrétisent les variables continues
- D'autres arbres (CART, C4.5, C5.0) peuvent aussi traiter directement les variables continues

> *Toutes les méthodes ne gèrent pas tous les types de données*

# Changement de format

| **type de départ** | **type d'arrivée** | **opération** | **principe** |
|---|---|---|---|
| continu | discret | discrétisation | découpage des valeurs en tranches |
| discret ou qualitatif | continu | ACM | une **A**nalyse des **C**orrespondances **M**ultiples fournit des facteurs continus à partir des données de départ |

# Pourquoi discrétiser ?

- Traiter simultanément des données quantitatives et qualitatives
- Appréhender des liaisons non linéaires (de degré >1) entre les variables continues
  - par une ACM, une régression logistique ou une analyse discriminante DISQUAL (Gilbert Saporta)
- Neutraliser les valeurs extrêmes
  - qui sont dans la 1ère et la dernière tranches
- Gérer les valeurs manquantes
  - rassemblées dans une tranche supplémentaire spécifique
- Gérer les ratios dont le numérateur et le dénominateur peuvent être tous deux > 0 ou < 0
- Renforcer la robustesse d'un modèle (on constate souvent que 2 ou 3 classes permettent d'augmenter l'aire sous la courbe ROC par rapport à 4 ou 5 classes)

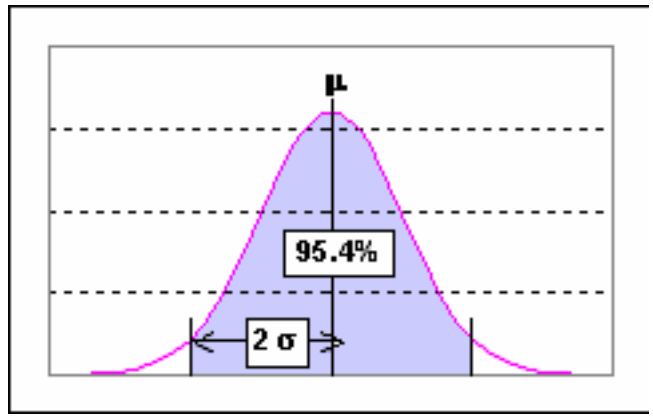
# Comment discrétiser ?

- Il faut garder en tête que :
  - il faut éviter d'avoir de grands écarts entre le nombre de modalités des différentes variables
  - mieux vaut éviter les modalités d'effectif trop petit
  - un nombre convenable de modalités pour une variable discrète ou catégorielle tourne autour de 4 ou 5
  - tenir compte de la variable cible le cas échéant
- pour les raisons que :
  - le poids d'une variable est proportionnel au nb de modalités
  - le poids d'une modalité est inversement proportionnel à son effectif
  - avoir peu de modalités peut fait perdre de l'information
  - avoir beaucoup de modalités implique des petits effectifs et une moindre lisibilité

# Analyse exploratoire des données

- Explorer la distribution des variables
- Vérifier la fiabilité des variables
  - valeurs incohérentes ou manquantes

    $\Rightarrow$ imputation ou suppression
- Détecter les valeurs extrêmes
  - voir si valeurs aberrantes à éliminer

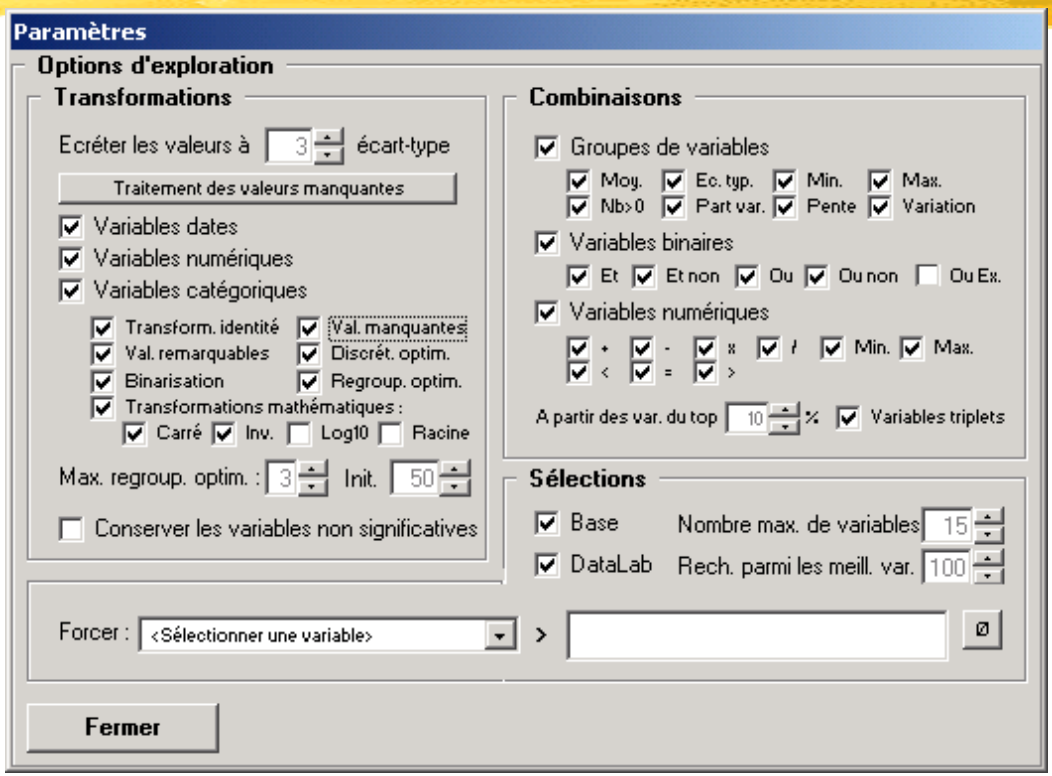



- Variables continues
  - détecter la non-monotonie ou la non-linéarité justifiant la discrétisation
  - tester la normalité des variables (surtout si petits effectifs) et les transformer pour augmenter la normalité
  - facultatif : tester l'homoscédasticité (égalité des matrices de variances-covariances)

# Analyse exploratoire des données

- Variables discrètes
  - regrouper certaines modalités trop nombreuses ou avec des effectifs trop petits (poids trop grand)
- Créer des indicateurs pertinents d'après les données brutes
  - prendre l'avis des spécialistes du secteur étudié
  - date de naissance + date 1er achat $\Rightarrow$ âge du client au moment de son entrée en relation avec l'entreprise
  - l'ensemble des variables « produit $P_i$ acheté (Oui/Non) » permet de déduire le nombre de produits achetés
  - nombre et montant des achats $\Rightarrow$ montant moyen d'un achat
  - dates d'achat $\Rightarrow$ récence et fréquence des achats
  - plafond de la ligne de crédit + part réellement utilisée $\Rightarrow$ taux d'utilisation du crédit

# Analyse exploratoire des données



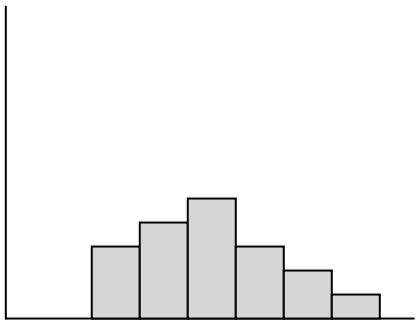

- Détecter les liaisons entre variables
  - entre variables explicatives et à expliquer (bon)
  - entre variables explicatives entre elles (multicolinéarité : mauvais dans certaines méthodes)

# Imputation des valeurs manquantes

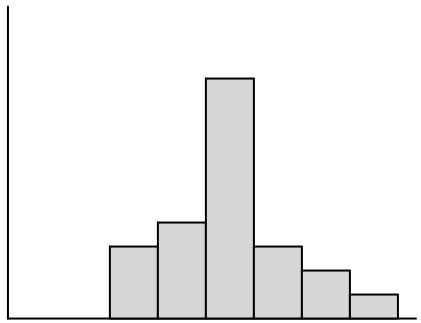
- Solutions autres que l'imputation statistique :
  - suppression des observations (si elles sont peu nombreuses)
  - ne pas utiliser la variable concernée ou la remplacer par une variable proche mais sans valeur manquante
  - traiter la valeur manquante comme une valeur à part entière
  - remplacement des valeurs manquantes par source externe
- Imputation statistique
  - par le mode, la moyenne ou la médiane
  - par une régression ou un arbre de décision
  - imputation simple (minore la variabilité et les intervalles de confiance des paramètres estimés) ou multiple (remplacer chaque valeur manquante par n valeurs, par ex. n = 5, puis faire les analyses sur les n tables et combiner les résultats pour obtenir les paramètres avec leurs écart-types

# L'imputation n'est jamais neutre

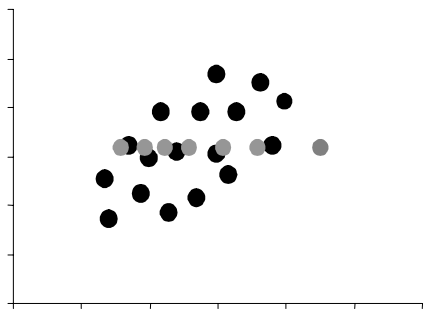
- Surtout si les données ne sont pas manquantes au hasard
- Déformation des variances et des corrélations



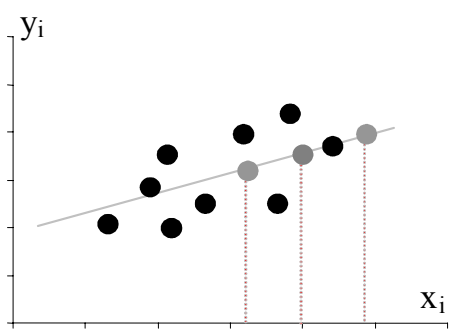
avant imputation

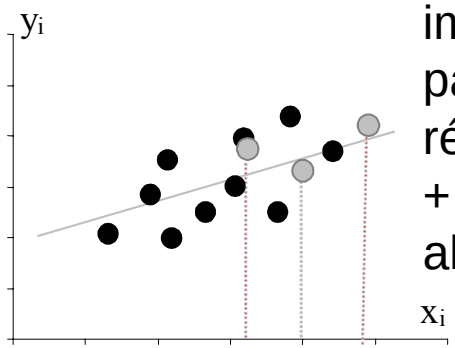


après imputation par la moyenne



imputation par
← moyenne
ou
régression
→

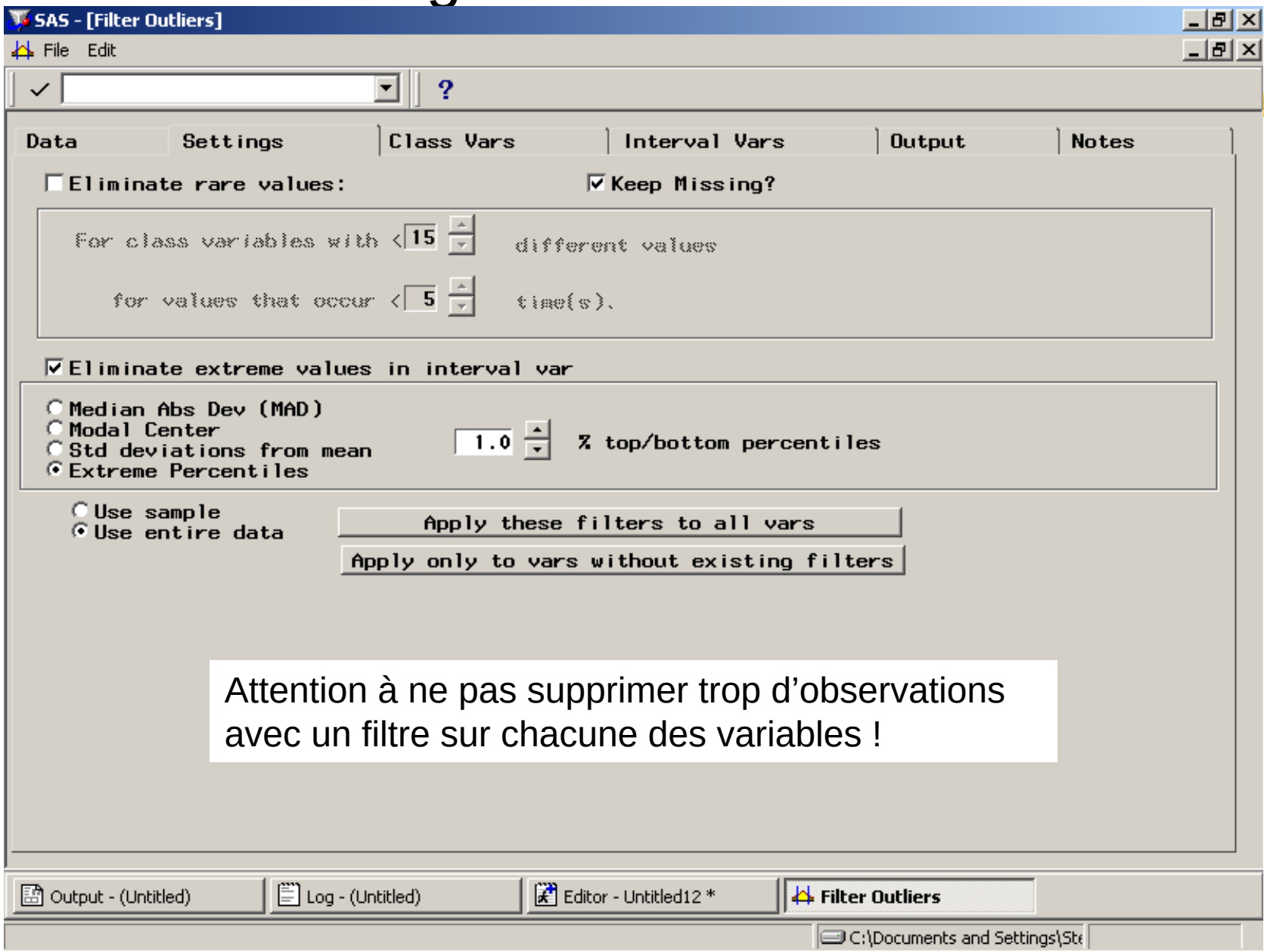



imputation par régression + résidu aléatoire

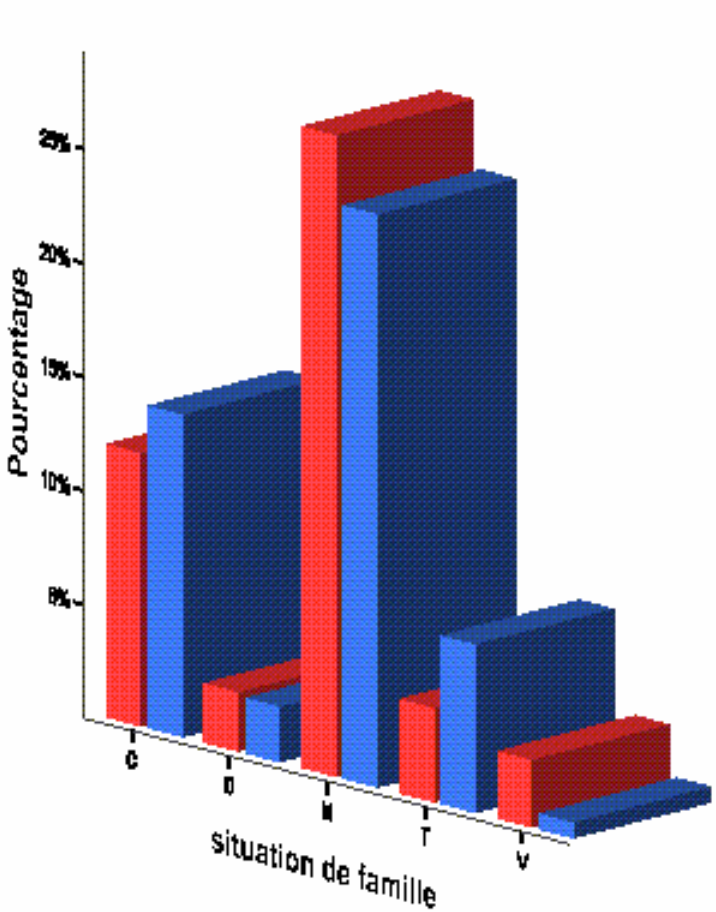
source : J.-P. Nakache – A. Gueguen, RSA 2005

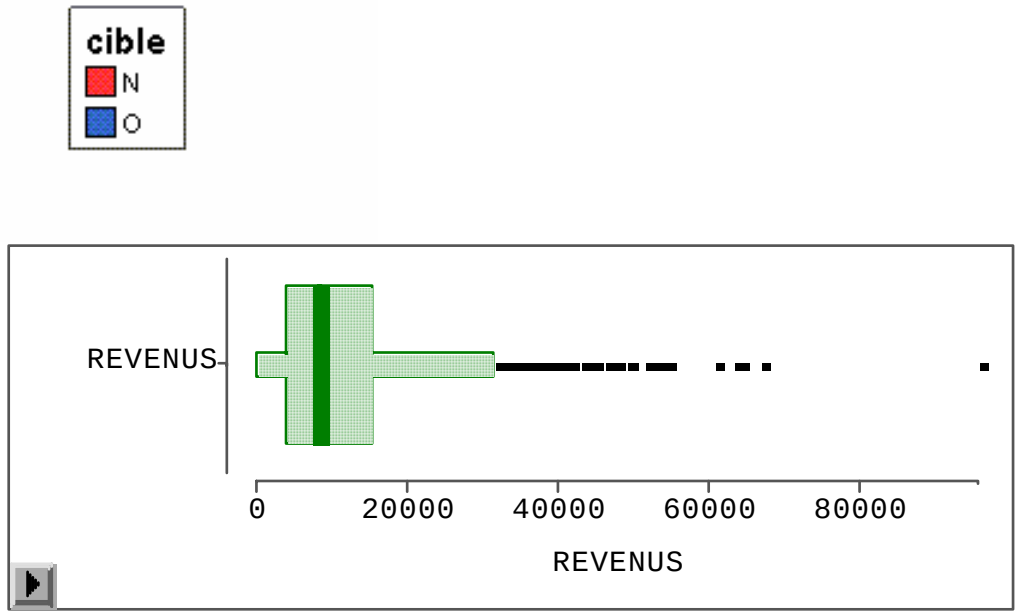
# Filtrage des extrêmes




Attention à ne pas supprimer trop d'observations avec un filtre sur chacune des variables !

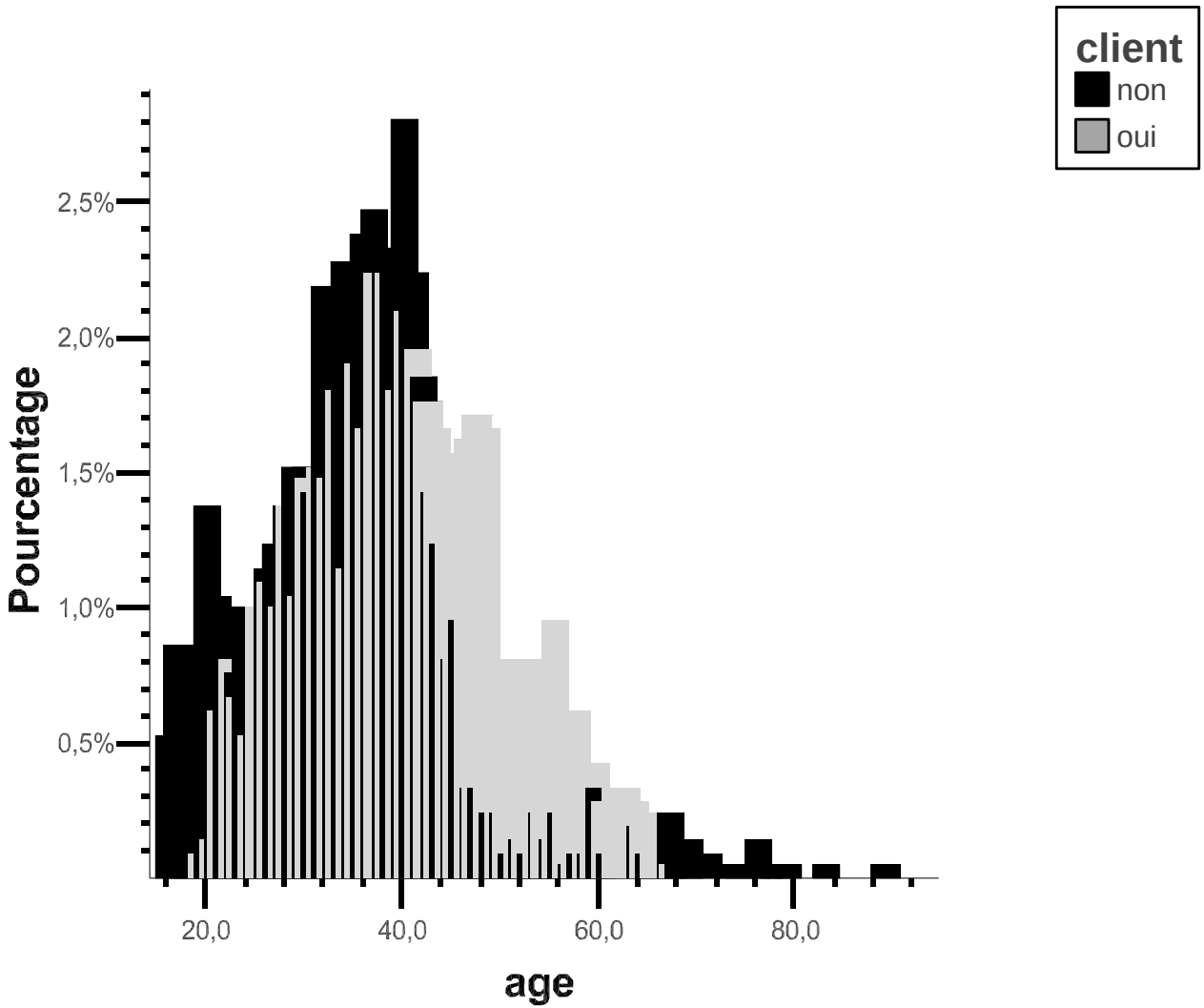
# Analyse exploratoire des données 1/2




Box-Plot (boîte à moustaches)

# Analyse exploratoire des données 2/2



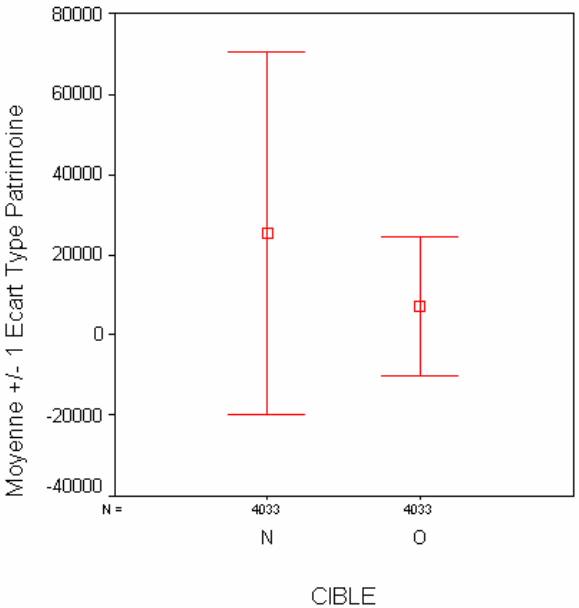
# Caractéristiques de dispersion

- Étendue (souvent peu significative à cause des extrêmes)
- Écart interquartile $q_3 - q_1$
- Variance
  - $H_0$ = égalité des variances d'une variable dans plusieurs groupes : homoscédasticité (contraire : hétéroscédasticité)
  - test de Levene, de Bartlett ou de Fisher
  - proba < 0,05 $\Rightarrow$ hétéroscédasticité
- Écart-type
- Coefficient de variation
  - écart-type / moyenne
  - X dispersée si CV(X) > 25 %
  - grandeur sans unité $\Rightarrow$ utile pour comparer la dispersion des variables

**Test of Homogeneity of Variance**

| | Levene Statistic | df1 | df2 | Sig. |
|---|---|---|---|---|
| 1996 Sales | 6.708 | 2 | 387 | .001 |

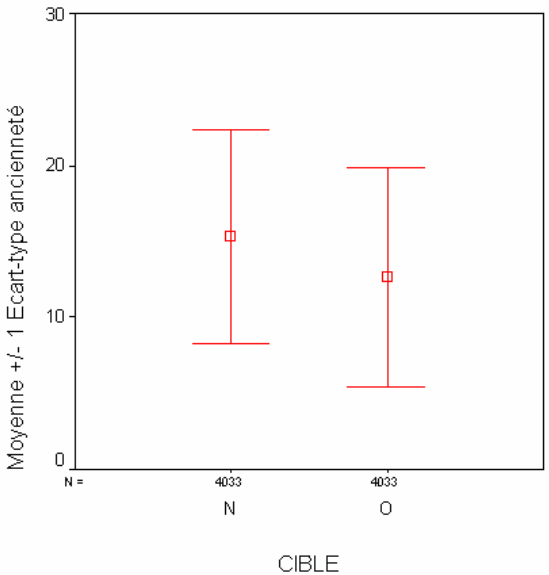
# Homogénéité des variances

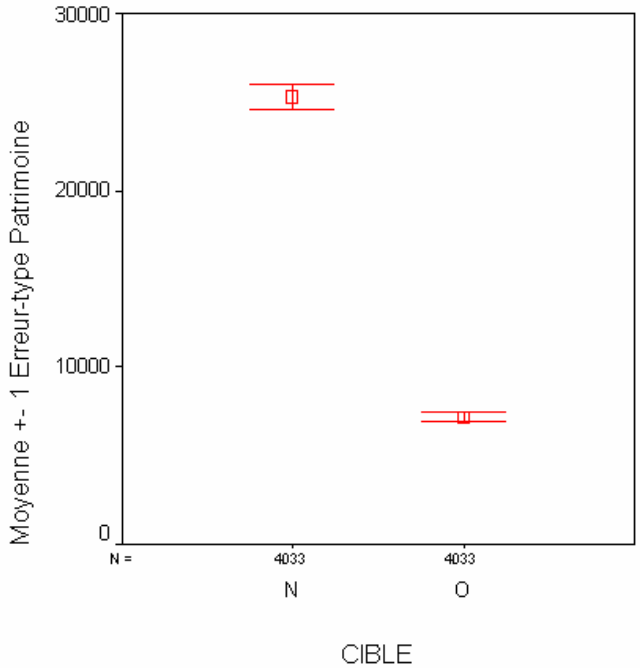



KO ↑

OK ➔

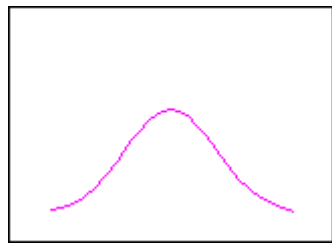
Erreur-type = écart-type de la moyenne = écart-type des observations / racine carrée de l'effectif



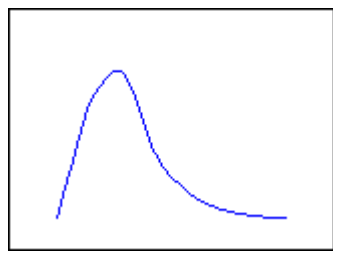



# Caractéristiques de forme 1/2

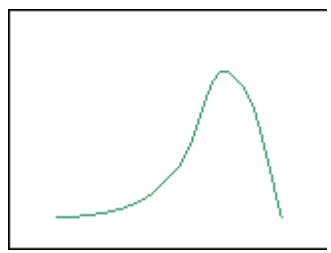
- Coefficient d'asymétrie (« skewness »)
  - = 0 si la série de données est symétrique



  - > 0 si elle est allongée vers la droite



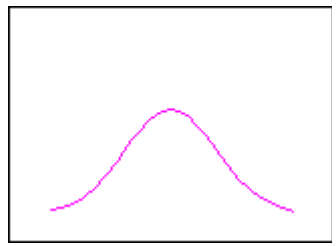
  - < 0 si elle est allongée vers la gauche



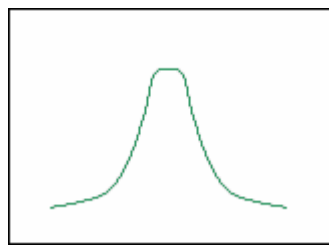
- Asymétrie positive fréquente dans les données économiques

# Caractéristiques de forme 2/2

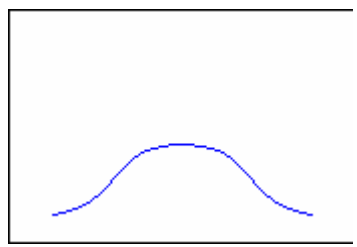
- Coefficient d'aplatissement (« kurtosis »)
  - = 3 si aplatie comme Gauss



  - > 3 si plus concentrée que Gauss



  - < 3 si plus aplatie que Gauss



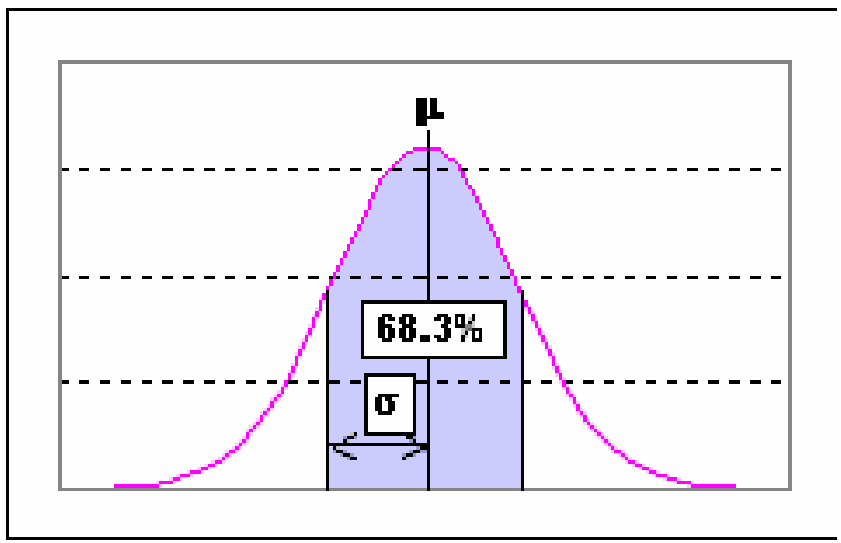
- Kurtosis (loi uniforme sur [0,1]) = 1,8
- **On normalise souvent le kurtosis en soustrayant 3**
  - SAS et SPSS le font

# Propriétés de la loi normale

- Elle est entièrement définie par sa moyenne $\mu$ et son écart-type $\sigma$ et toute loi $N(\mu,\sigma)$ peut se ramener à une loi N(0,1)
- Sa moyenne = sa médiane = son mode
- Loi limite des lois : binomiale, de Poisson et du $\chi^2$
- Loi très fréquente dans les phénomènes biologiques et médicaux
- Les observations sont distribuées symétriquement autour de la moyenne :
  - 68 % des observations se trouvent à une distance $\pm \sigma$ de $\mu$
  - 95 % des observations se trouvent à une distance $\pm 2\sigma$ de $\mu$
  - 99,8 % des observations se trouvent à une distance $\pm 3\sigma$ de $\mu$
- Une transformation permet parfois de rendre normale une variable

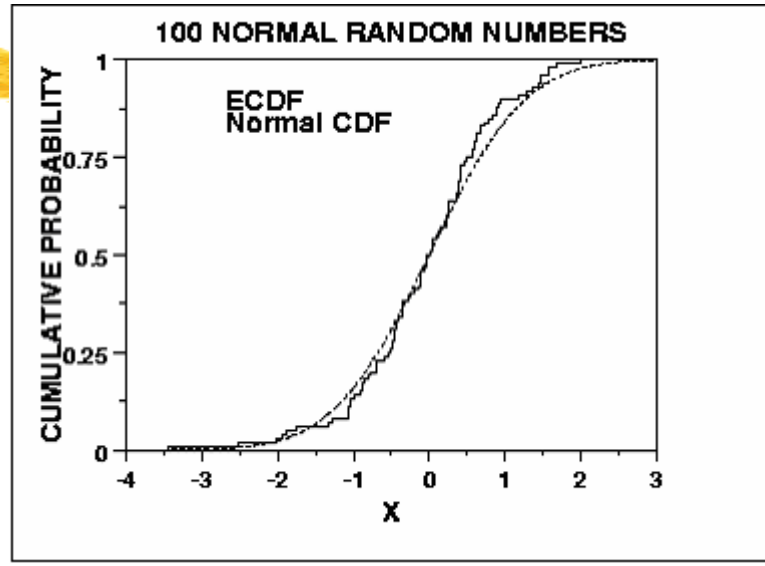
# Situation de la loi normale

- Référence pour de nombreux indices
- Cadre de nombreux tests (t de Student, ANOVA, corrélation de Pearson)
- Hypothèse dans l'analyse discriminante de Fisher, dans la régression linéaire, etc.




- Non normalité moins gênante si les effectifs sont grands

# Tests de normalité




- Test de Kolmogorov-Smirnov
  - mesure l'écart maximum (en valeur absolue) entre la fonction de répartition de la variable testée et celle d'une variable normale
  - hypothèse nulle $H_0$ : les données suivent une distribution donnée (ici une distribution normale, mais le test de K-S s'applique + généralement à d'autre distributions continues)
  - test plus sensible au centre de la distribution

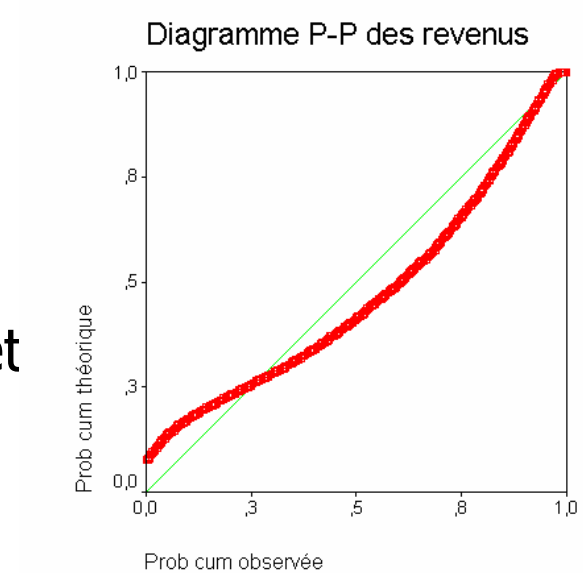
Tests for Normality

| Test Statistic | Value | p-value |
|---|---|---|
| Shapiro-Wilk | 0.985964 | 0.0000 |
| Kolmogorov-Smirnov | 0.047426 | <.0100 |
| Cramer-von Mises | 0.329094 | <.0050 |
| Anderson-Darling | 2.656450 | <.0050 |

Invalidation de $H_0$ : {distribution normale} ⇒ la distribution n'est pas normale

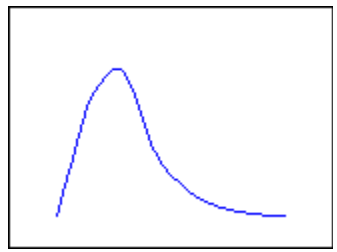
# Tests de normalité

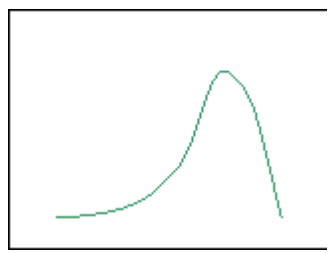
- Test d'Anderson-Darling
  - modifie Kolmogorov-Smirnov en donnant plus de poids aux queues de distribution
- Test de Lilliefors
  - perfectionne K-S quand on ne connaît pas la moyenne et la variance de la var.
  - car dans ce cas le test de K-S est conservateur si calculé avec la moyenne et la variance estimées sur l'échantillon
- Test de Shapiro-Wilk
  - mesure l'alignement sur la droite ci-contre correspondant à une distribution normale
  - le meilleur test sur de petits échantillons
  - présent dans la proc UNIVARIATE de SAS



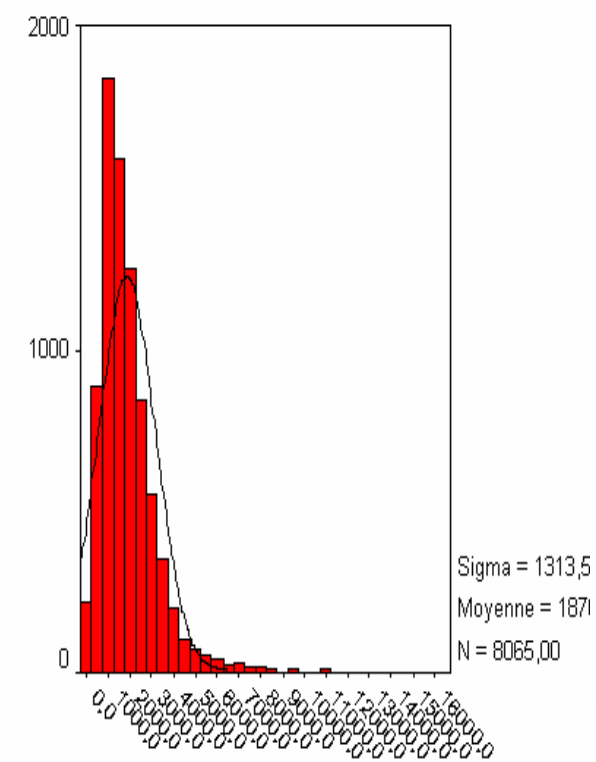
# Normalisation : transformations

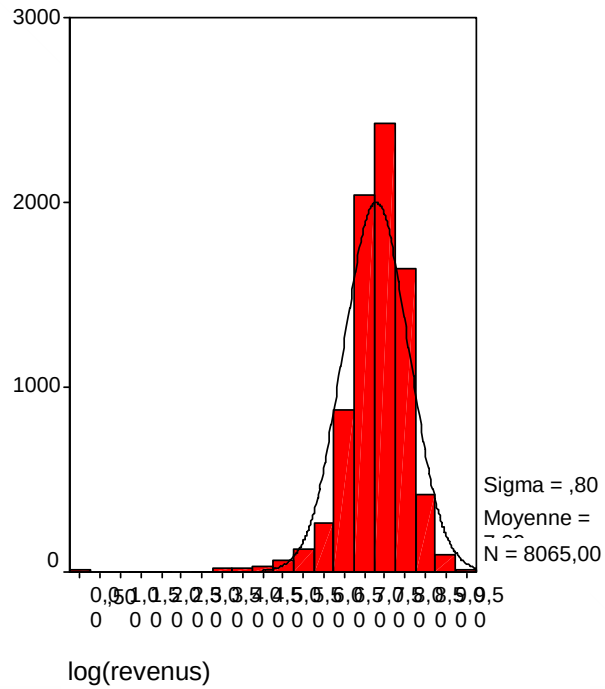
- Log (V)
  - transformation la plus courante pour corriger un coefficient d'asymétrie > 0
  - Si V ≥ 0, on prend Log (1 + V)
- Racine carrée (V) si coefficient d'asymétrie > 0
- -1/V ou $-1/V^2$ si coefficient d'asymétrie > 0
- $V^2$ ou $V^3$ si coefficient d'asymétrie < 0
- Arc sinus (racine carrée de V/100)
  - si V est un pourcentage compris entre 0 et 100
- Certains logiciels déterminent automatiquement la transformation la plus adaptée
  - en utilisant l'algorithme de Box et Cox ou la loi de Taylor

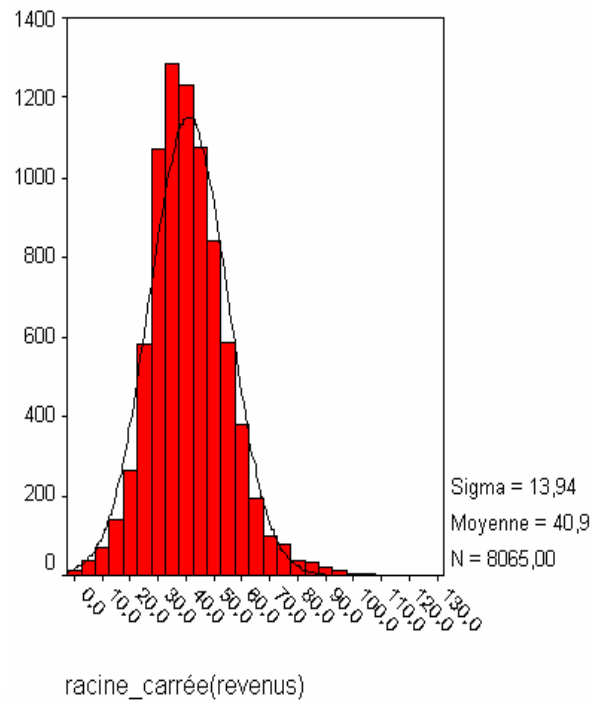




# Normalisation : exemple des revenus




**Revenus :**

**Asymétrie = 2,38**

**Aplatissement = 11,72**

**Log(1+revenus) :**

**Asymétrie = - 2,03**

**Aplatissement = 12,03**

**Racine(revenus) :**

**Asymétrie = 0,64**

**Aplatissement = 1,76**

La racine carrée normalise ici mieux que le logarithme

# Tableau de contingence

- Un **tableau de contingence** est le croisement de 2 variables catégorielles A et B : le coefficient $x_{ij}$ du tableau = nb d'individus x tels que $A(x) = a_i$ et $B(x) = b_j$
- Le test du $\chi^2$ permet de détecter une dépendance entre les deux variables
- La contribution au $\chi^2$ de chaque cellule du tableau de contingence montre les liaisons entre modalités des 2 variables : soit sur-effectif, soit sous-effectif, soit équilibre
- S'il y a de nombreuses modalités, il est fastidieux de parcourir toutes les cellules
- S'il y a plus de 2 variables à croiser, c'est encore + ardu, voire impossible $\Rightarrow$ se tourner vers l'analyse des correspondances multiples

# Pièges des tableaux de contingence et paradoxe de Simpson 1/2

| **Tous clients** | | | | |
|---|---|---|---|---|
| | *sans achat* | *avec achats* | *TOTAL* | *taux d'achat* |
| *courriel* | 2 850 | 58 | 2 908 | 1,99% |
| *téléphone* | 2 185 | 78 | 2 263 | 3,45% |
| *TOTAL* | 5 035 | 136 | 5 171 | 2,63% |
| | | | | |
| **Hommes** | | | | |
| | *sans achat* | *avec achats* | *TOTAL* | *taux d'achat* |
| *courriel* | 2 341 | 24 | 2 365 | 1,01% |
| *téléphone* | 515 | 3 | 518 | 0,58% |
| *TOTAL* | 2 856 | 27 | 2 883 | 0,94% |
| | | | | |
| **Femmes** | | | | |
| | *sans achat* | *avec achats* | *TOTAL* | *taux d'achat* |
| *courriel* | 509 | 34 | 543 | 6,26% |
| *téléphone* | 1 670 | 75 | 1 745 | 4,30% |
| *TOTAL* | 2 179 | 109 | 2 288 | 4,76% |

# Pièges des tableaux de contingence et paradoxe de Simpson 2/2

| **Tous clients** | | | | |
|---|---|---|---|---|
| | *sans achat* | *avec achats* | *TOTAL* | *taux d'achat* |
| *courriel* | 1 400 | 100 | 1 500 | 6,67% |
| *téléphone* | 1 375 | 125 | 1 500 | 8,33% |
| *TOTAL* | 2 775 | 225 | 3 000 | 7,50% |
| | | | | |
| **Hommes** | | | | |
| | *sans achat* | *avec achats* | *TOTAL* | *taux d'achat* |
| *courriel* | 950 | 50 | 1 000 | 5,00% |
| *téléphone* | 475 | 25 | 500 | 5,00% |
| *TOTAL* | 1 425 | 75 | 1 500 | 5,00% |
| | | | | |
| **Femmes** | | | | |
| | *sans achat* | *avec achats* | *TOTAL* | *taux d'achat* |
| *courriel* | 450 | 50 | 500 | 10,00% |
| *téléphone* | 900 | 100 | 1 000 | 10,00% |
| *TOTAL* | 1 350 | 150 | 1 500 | 10,00% |

# Paradoxe de Simpson : explication

- Dans le dernier exemple :
  - les hommes ne répondent pas mieux au téléphone qu'au courriel
  - de même pour les femmes
  - et pourtant, le téléphone semble avoir globalement un meilleur taux d'achat
- Explication :
  - un individu pris au hasard ne répond pas mieux au téléphone
  - mais les femmes achètent plus et on a privilégié le téléphone pour les contacter (liaison positive entre les variables « sexe » et « canal de distribution »)

# La préparation des données :
# Tests statistiques

# Principe général d'un test d'hypothèse

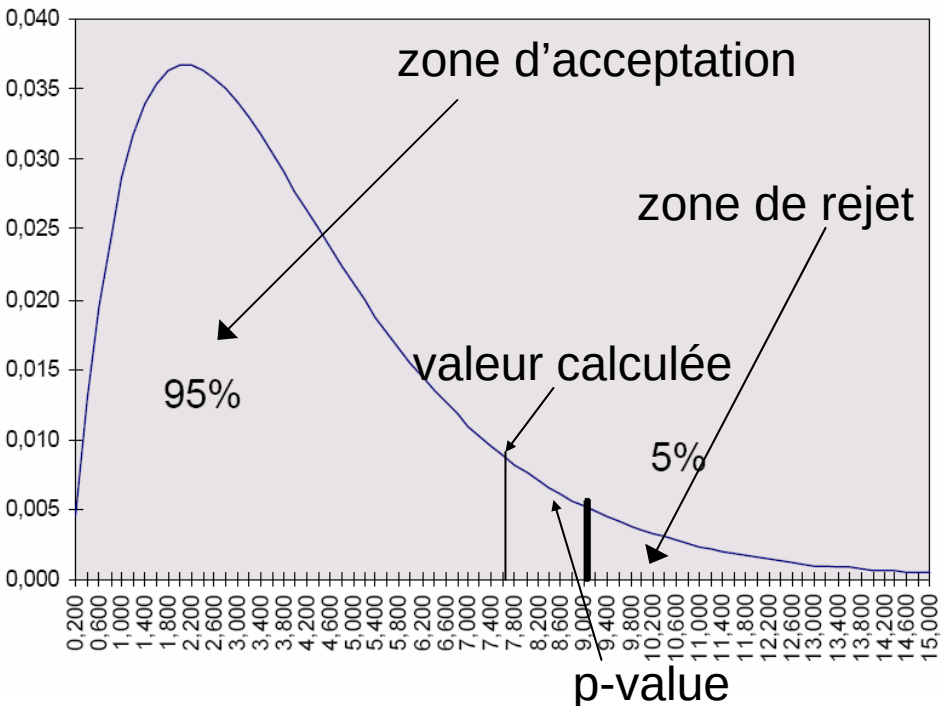
- Quand on veut démontrer l'hypothèse $H_1$ que :
  - une moyenne mesurée dans un échantillon est significativement différente de la moyenne dans la population
    - significativement = ne résulte pas uniquement du hasard
  - des moyennes mesurées dans 2 échantillons sont significativement différentes
  - une variable ne suit pas une loi théorique donnée
  - deux variables sont significativement différentes
  - un échantillon n'est pas homogène mais est composé de plusieurs sous-populations
- … on soumet l'hypothèse contraire $H_0$ à un test T qui doit être satisfait si $H_0$ est vraie
- … puis on montre que T n'est pas satisfait $\Rightarrow$ $H_0$ est faux
- Vocab. : $H_0$ : hypothèse nulle – $H_1$ : hypothèse alternative

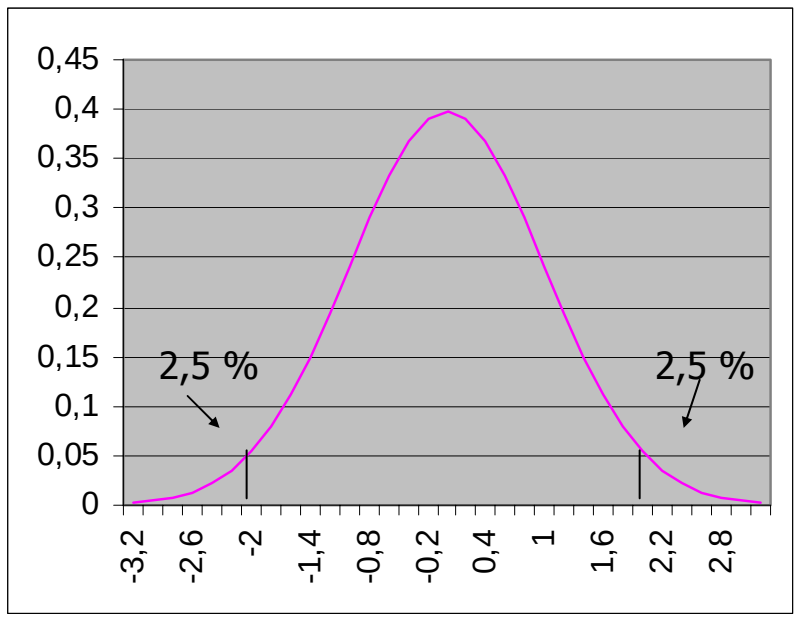
# Exemples

- Égalité de moyennes dans 2 échantillons : test de Student
- Égalité de moyennes dans k > 2 échantillons : analyse de la variance
- Égalité de 2 variances : test de Fisher-Snedecor
- Égalité de 2 distributions : test de Kolmogorov-Smirnov
- Indépendance de 2 variables qualitatives : test du $\chi^2$
  - ce test est non-paramétrique
  - mais non exact -> le test exact correspondant est le test de Fisher (ne pas confondre avec le test de Fisher-Snedecor)
  - voir plus loin ces notions de « paramétrique » et « exact »

# Déroulement du test d'hypothèse 1/2

- À l'hypothèse nulle $H_0$ est associée une statistique, fonction des observations, qui suit une loi théorique connue si $H_0$ est vraie
  - exemple : si l'hypothèse nulle est ($H_0$ : $\mu = \mu_0$), alors $\frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$ suit une loi normale réduite (*n* grand)



# Déroulement du test d'hypothèse 2/2

- Dans la distribution de cette loi, on choisit une zone de rejet (unilatérale ou bilatérale), caractérisée par une probabilité $\alpha$ d'être dans cette zone
  - on choisit souvent $\alpha = 0,05$ (= 5 %)
  - le complémentaire est la zone d'acceptation (si $\alpha = 0,05$, il s'agit de la région autour de la moyenne où se trouvent 95 % des valeurs de la statistique)
- On mesure la valeur de la statistique sur l'échantillon et on compare cette valeur aux valeurs théoriques de la loi
- Si cette valeur mesurée tombe dans la zone de rejet, on rejette $H_0$
  - sinon, on ne la rejette pas

# Niveau de signification (p-value)

- Niveau de signification = degré de signification = p-value = probabilité d'obtenir une statistique de test aussi extrême ($\geq$ ou $\leq$) que la valeur mesurée sur l'échantillon si $H_0$ est vraie
- Utilisation de la p-value :
  - p-value $\geq \alpha \Rightarrow$ ne pas rejeter $H_0$
  - p-value $< \alpha \Rightarrow$ rejeter $H_0$ (on considère qu'il est trop peu probable d'avoir une si faible p-value si $H_0$ est vraie, pour admettre que $H_0$ est vraie)
- Intérêt de la p-value :
  - elle a un sens absolu, qui ne dépend pas de la loi de probabilité et du nombre de degrés de liberté

# Risques d'erreur

- Les deux erreurs possibles dans un test sont :
  - le rejet d'une $H_0$ vraie (risque de 1ère espèce, ou de type I)
    - probabilité de cette erreur = $\alpha$
  - le non rejet d'une $H_0$ fausse (risque de 2de espèce, ou de type II)
    - probabilité de cette erreur = $\beta$

| | | REALITE | REALITE |
|---|---|---|---|
| | | **$H_0$ vraie** | **$H_0$ fausse** |
| DECISION | **$H_0$ non rejetée** | décision correcte $(1 - \alpha)$ | risque $\beta$ (type II) |
| DECISION | **$H_0$ rejetée** | risque $\alpha$ (type I) | décision correcte $(1 - \beta)$ |

- On ne peut réduire simultanément $\alpha$ et $\beta$

# Puissance d'un test 1/2

- Puissance d'un test : 1 – risque $\beta$
- Probabilité de rejeter $H_0$ si celle-ci est fausse
  - décision correcte
- Le risque $\beta$ et la puissance 1 – $\beta$ dépendent de :
  - la vraie valeur du paramètre de la population (plus elle est éloignée de la valeur testée, plus le risque $\beta$ baisse)
  - l'écart-type $\sigma$ de la population ($\sigma \nearrow \Rightarrow \beta \searrow$)
  - le risque $\alpha$ choisi ($\alpha \nearrow \Rightarrow \beta \searrow$)
  - la taille n de l'échantillon ($n \nearrow \Rightarrow \beta \searrow$)

- Synonymes : test libéral = test puissant
- Antonymes : test conservateur $\neq$ test puissant

# Puissance d'un test 2/2

- La puissance d'un test augmente avec la taille de l'échantillon
  - plus les observations sont nombreuses, plus on a d'éléments permettant de rejeter $H_0$ si elle est fausse
- Attention, avec des tests puissants, on rejette facilement $H_0$ dès que le nb d'observations augmente
  - ex : le test du $\chi^2$, le test de Hosmer-Lemeshow
  - ex : les tests de normalité
- Remarque : les tests d'hypothèse s'appliquent bien à des hypothèses $H_0$ contraignantes (ex : $\mu = \mu_0$) car elles conduisent à des tests T précis $\Rightarrow$ les tests permettent de prouver qu'un échantillon est hétérogène ou n'a pas été constitué par un tirage au hasard, mais non l'inverse

# Tests asymptotiques et exacts

- Test asymptotique
  - approximation valable quand les effectifs sont assez grands et les tableaux de données assez denses
  - ex : test du $\chi^2$ (si effectifs théoriques $\geq 5$)
- Test exact
  - utilisable sur des données éparses
  - calcul direct de probabilité
  - … prenant en compte tous les cas de figure possibles
  - calcul pouvant être coûteux en temps machine
    - variante : approximation par la méthode de Monte-Carlo
  - ex : test de Fisher

# Tests paramétriques et non-paramétriques

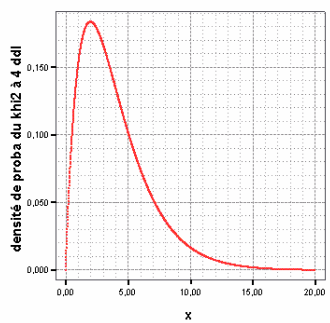
- Tests paramétriques
  - supposent que les variables suivent une loi particulière (normalité, homoscédasticité)
  - parfois plus puissants que des tests non-paramétriques, mais rarement beaucoup plus
  - ex : test de Student, ANOVA
- Tests non-paramétriques
  - ne supposent pas que les variables suivent une loi particulière
  - se fondent souvent sur les rangs des valeurs des variables plutôt que sur les valeurs elles-mêmes
  - peu sensibles aux valeurs aberrantes
  - à privilégier avec de petits effectifs (< 10)
  - par définition, les tests d'adéquation à une loi (ex : tests de normalité) sont non-paramétriques
  - ex : test de Wilcoxon, test de Kruskal-Wallis

# Exemple du test du $\chi^2$

- Test utilisé pour s'assurer que la distribution d'une variable suit une loi probabiliste donnée, en comparant la distribution observée (d'effectifs $\{O_i\}_i$) et la distribution théorique (d'effectifs $\{T_i\}_i$).
- Hypothèse $H_0$ : {effectifs observés = effectifs théoriques}
- Si $H_0$ est vraie, alors chaque élément $(O_i - T_i)/ T_i$ tend vers une loi normale centrée réduite (d'après le théorème central-limite) lorsque l'effectif théorique $T_i$ est assez grand (traditionnellement $n \geq 5$)
- Donc, si $H_0$ est vraie et si $T_i \geq 5$ pour tout i, la quantité $\Sigma_i(O_i - T_i)^2/ T_i$ suit une loi qui est une somme de p carrés de lois N(0,1) indépendantes : une telle loi est dite « du $\chi^2$ » à p degrés de liberté
- Si $T_i < 5$ pour au moins un i, préférer un test exact

# Le test du $\chi^2$ appliqué à la liaison entre variables catégorielles

- Le test du $\chi^2$ est souvent utilisé pour tester l'indépendance de deux variables catégorielles X et Y
- Si X et Y sont indépendantes ($H_0$), alors, pour tous i et j :
  - nombre d'individus tels que {X=i et Y=j} =
    nb d'individus tq {X=i} x nb d'individus tq {Y=j} x 1/N
  - où N est le nombre total d'individus
- En notant $O_{ij}$ le terme à gauche de l'égalité ci-dessus, et $T_{ij}$ le terme de droite, le test d'indépendance de X et Y est le test du $\chi^2$ appliqué à la statistique $\Sigma_{ij} (O_{ij} - T_{ij})^2 / T_{ij}$
- Nb de d° de liberté p = (nb lignes – 1) x (nb colonnes – 1)



# Le test exact de Fisher

- Pour la loi du $\chi^2$ à p degrés de liberté, on a : variance = 2p, moyenne = p et mode = p – 2, et cette loi tend vers N(p,$\sqrt{2p}$) si p > 30

- Remarque : pas d'hypothèse sur la loi théorique suivie $\Rightarrow$ le test du $\chi^2$ est non-paramétrique
- En revanche, ce n'est qu'asymptotiquement que $\Sigma_i(O_i - T_i)^2/ T_i$ suit une loi du $\chi^2$ : si les effectifs théoriques sont faibles (< 5), il faut faire un calcul exact de probabilité
- Le test exact remplaçant le $\chi^2$ pour un tableau de contingence n x n est appelé test de Fisher et fait appel à la loi hypergéométrique

# Utilisation de la loi hypergéométrique

- Soient A et B deux variables à 2 modalités
- Elles ont un tableau de contingence 2x2 dont les effectifs seront notés a, b, c, d
- Si A et B sont indépendantes, la probabilité d'avoir un tableau (a,b,c,d) de marges fixées a+c, b+d, a+b, c+d, est donnée par la loi hypergéométrique

$$P(a,b,c,d) := \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!(a+b+c+d)!}$$

- Plus le tableau (a,b,c,d) s'éloigne de l'indépendance, plus la probabilité P(a,b,c,d) est petite
- La p-value du test exact de Fisher = probabilité de la table observée (a,b,c,d) + probabilité de chaque table présentant un plus grand écart à l'indépendance que (a,b,c,d)

# Loi hypergéométrique : exemple

- Soit le tableau suivant de marges (10,9,9,10) :

| | |
|---|---|
| 2 | 7 |
| 8 | 2 |

- La p-value du test unilatéral vaut :
  - P(2,8,7,2) + P(1,9,8,1) + P(0,10,9,0) = 0,01754 + 0,00097 + 0,00001 = 0,01852
- La p-value du test bilatéral vaut :
  - P(2,8,7,2) + P(1,9,8,1) + P(0,10,9,0) + P(8,2,1,8) + P(9,1,0,9) = 0,01852 + 0,00438 + 0,00011 = 0,02301

| | Valeur | ddl | Signification asymptotique (bilatérale) | Signification exacte (bilatérale) | Signification exacte (unilatérale) |
|---|---|---|---|---|---|
| Khi-deux de Pearson | 6,343[b] | 1 | ,01179 | | |
| Correction pour la continuité[a] | 4,237 | 1 | ,03955 | | |
| Test exact de Fisher | | | | ,02301 | ,01852 |
| Nombre d'observations valides | 19 | | | | |

a. Calculé uniquement pour un tableau 2x2

b. 3 cellules (75,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 4,26.

# Loi hypergéométrique : exemple (suite)

- À noter que sur cet exemple, au seuil $\alpha$ de 2 %, le test du $\chi^2$ rejette l'hypothèse $H_0$, tandis que le test exact de Fisher ne rejette pas l'hypothèse d'indépendance
  - on retrouve ici la puissance du test du $\chi^2$ qui lui permet d'être (trop ?) sensible à des écarts faibles à l'indépendance
  - on retrouve le même phénomène avec d'autres tests comme Wilcoxon, dont la p-value exacte est > p-value asymptotique
- Avantage du test exact : existe en version unilatérale, contrairement au test du $\chi^2$
- Correction de continuité de Yates
  - utile pour de petits effectifs, où le test du $\chi^2$ corrigé se rapproche du test exact de Fisher
  - se rapproche du $\chi^2$ asymptotique quand les effectifs croissent

# Loi hypergéométrique : cadre général

- On procède à un tirage sans remise (= simultané) de n boules dans une urne contenant $n_1$ boules gagnantes et $n_2$ boules perdantes
- La loi hypergéométrique de paramètres $(n,n_1,n_2)$ est la loi que suit la variable aléatoire « nb de boules gagnantes tirées »
- Lien avec la proba P(a,b,c,d) du tableau de contingence à marges fixées

| boules gagnantes | boules perdantes | |
|---|---|---|
| a | b | a+b=n |
| c | d | c+d |
| a+c = $n_1$ | b+d = $n_2$ | $n_1$ + $n_2$ |

- Proba de tirer « a » boules gagnantes = proba d'avoir la configuration (a,b,c,d)

# $\chi^2$ : Attention aux effectifs 1/2

| | Classe 1 | Classe 2 | Ensemble |
|---|---|---|---|
| Effectifs observés : | | | |
| A | 55 | 45 | 100 |
| B | 20 | 30 | 50 |
| Total | 75 | 75 | 150 |
| Effectifs attendus si la variable est indépendante de la classe : | | | |
| A | 50 | 50 | 100 |
| B | 25 | 25 | 50 |
| Total | 75 | 75 | 150 |
| Probabilité du $\chi^2$ = 0,08326454 | | | |

- Dans la population de 150 individus, il y a 66,66% d'individus vérifiant A
- Dans la classe 1, il y a 73,33% d'individus vérifiant A

> Le test du $\chi^2$ indique que les écarts entre effectifs observés et attendus ne sont pas significatifs (proba > 0,05). Ici $\chi^2 = 3$.

# $\chi^2$ : Attention aux effectifs 2/2

| | Classe 1 | Classe 2 | Ensemble |
|---|---|---|---|
| Effectifs observés : | | | |
| A | 550 | 450 | 1000 |
| B | 200 | 300 | 500 |
| Total | 750 | 750 | 1500 |
| Effectifs attendus si la variable est indépendante de la classe : | | | |
| A | 500 | 500 | 1000 |
| B | 250 | 250 | 500 |
| Total | 750 | 750 | 1500 |
| Probabilité du $\chi^2$ = $4{,}3205.10^{-8}$ | | | |

- Dans la population de 1500 individus, il y a 66,66% d'individus vérifiant A
- Dans la classe 1, il y a 73,33% d'individus vérifiant A

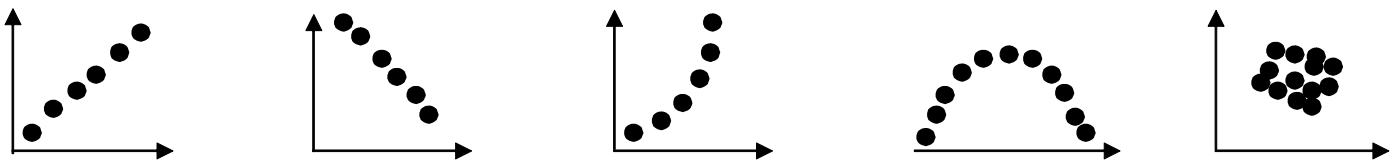
> Ici $\chi^2 = 30$. Test du $\chi^2$ : les écarts sont significatifs (proba < 0,05)

> Quand la taille de la population augmente, le moindre écart devient significatif aux seuils usuels

# V de Cramer

- V de Cramer = $\sqrt{\frac{\chi^2}{\chi^2_{max}}}$
  - mesure directement l'intensité de la liaison de 2 variables catégorielles, sans avoir recours à une table du $\chi^2$
  - en intégrant l'effectif et le nombre de degrés de liberté, par l'intermédiaire de $\chi^2_{max}$
  - $\chi^2_{max}$ = effectif x [min (nb lignes, nb colonnes) – 1]
  - V compris entre 0 (liaison nulle) et 1 (liaison parfaite)

# Liaison entre 2 variables continues : Coeff. de corrélation linéaire (Pearson)



| | A | B | C | D | E |
|---|---|---|---|---|---|
| | corrélation positive | corrélation négative | corrélation positive | pas de corrélation, mais dépendance | indépendance |
| liaison : | monotone linéaire croissante | monotone linéaire décroissante | monotone non linéaire croissante | non monotone | |

La **liaison est nulle** si le coefficient de corrélation = 0 (nuage de points circulaire ou parallèle à un des 2 axes)

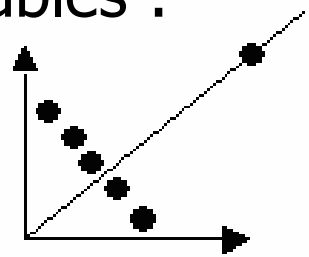
La **liaison est parfaite** si le coefficient de corrélation = +1 ou -1 (nuage de points rectiligne)

La **liaison est forte** si le coefficient de corrélation > +0,8 ou < -0,8 (nuage de points elliptique et allongé)

Mais une **liaison non linéaire** (par ex : quadratique) et surtout **non monotone** n'est pas mesurable par le coefficient de corrélation

# Coefficients de Pearson et Spearman

- Rho de Spearman plus général car calculé sur les rangs des valeurs et non les valeurs elles-mêmes
  - c'est un test non paramétrique (contrairement à Pearson)
- Préférer le rho de Spearman si les variables :
  - ne suivent pas une loi normale
  - ont des valeurs extrêmes
  - ne sont pas continues mais ordinales



corrélation négative
apparemment positive

  - ou pour détecter des liaisons monotones non linéaires
- Comparer r de Pearson et $\rho$ de Spearman :
  - $r > \rho$ => présence de valeurs extrêmes
  - $\rho > r$ => liaison non linéaire non détectée par Pearson
    - exemple : x = 1, 2, 3… et y = $e^1$, $e^2$, $e^3$…

# Multicolinéarité (corrélation multiple)

- Certaines techniques (ADL, régression logistique) sont sensibles à la colinéarité des variables explicatives
- En théorie, il ne suffit pas de vérifier les variables 2 à 2
- Tolérance d'une variable = proportion de la variance non expliquée par les autres variables - doit être > 0,1
  - VIF (variable inflation factor) = 1 / tolérance
- Indices de conditionnement de la matrice des corrélations
  - multicolinéarité modérée (forte) si des indices $\eta_k > 10\ (30)$

| | Valeur propre | Indice de conditionnement | Proportions de la variance | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | (cste) | var 1 | var 2 | var 3 | var 4 | var 5 | var 6 |
| 1 | 3,268 | 1,000 | ,01 | ,00 | ,03 | ,02 | ,01 | ,01 | ,02 |
| 2 | 1,022 | 1,788 | ,00 | ,56 | ,01 | ,02 | ,00 | ,33 | ,00 |
| 3 | ,976 | 1,830 | ,00 | ,42 | ,00 | ,10 | ,00 | ,42 | ,01 |
| 4 | ,811 | 2,008 | ,00 | ,02 | ,07 | ,81 | ,00 | ,14 | ,00 |
| 5 | ,636 | 2,266 | ,01 | ,00 | ,78 | ,04 | ,02 | ,09 | ,00 |
| 6 | ,221 | 3,842 | ,01 | ,00 | ,11 | ,01 | ,20 | ,00 | ,73 |
| 7 | ,065 | 7,099 | ,97 | ,00 | ,00 | ,00 | ,76 | ,00 | ,24 |

# Généralisation à des groupes de variables 1/2

- Analyse de corrélation canonique (linéaire)
  - non plus entre n ($\geq 2$) ensembles composés chacun de 1 variable continue
  - mais entre n ($\geq 2$) ensembles $\{U_i\}$, $\{V_j\}$… de plusieurs variables continues ou binaires
  - on cherche les combinaisons linéaires (variables canoniques) maximisant la corrélation entre $\Sigma_i \lambda_i U_i$, $\Sigma_j \mu_j V_j$…
  - proc CANCORR de SAS (si n = 2) – proc OVERALS de SPSS
  - origine : Harold Hotelling (1936)
  - généralise la régression multiple
    - n = 2 et l'un des ensembles de variables est composé d'une seule variable

# Généralisation à des groupes de variables 2/2

- Analyse de corrélation canonique (linéaire)
  - généralise aussi l'analyse discriminante linéaire
    - n = 2 et l'un des ensembles de variables est composé des indicatrices de la partition à discriminer
  - généralise aussi l'analyse factorielle des correspondances
    - n = 2 et chaque ensemble de variables est composé des indicatrices d'une variable catégorielle

- Analyse de corrélation canonique (non linéaire)
  - entre n ($\geq$ 2) ensembles de variables quelconques
  - permet la détection d'effets non linéaires
  - proc OVERALS de SPSS

# Liaison entre 1 variable continue et 1 variable catégorielle

| **lois suivies** | **2 échantillons** | **3 échantillons et plus (***)** |
|---|---|---|
| normalité – homoscédasticité (*) | test T de Student | ANOVA |
| normalité – hétéroscédasticité | test T de Welch | Welch - ANOVA |
| non normalité – hétéroscédasticité (**) | Wilcoxon – Mann – Whitney | Kruskal – Wallis |
| non normalité – hétéroscédasticité (**) | test de la médiane | test de la médiane |
| non normalité – hétéroscédasticité (**) | | test de Jonckheere-Terpstra (échantillons ordonnés) |

moins puissant

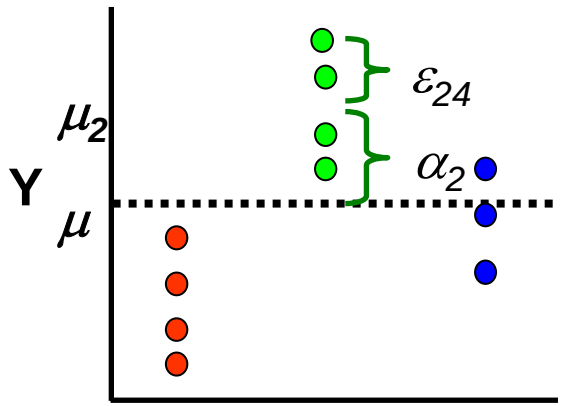
(*) Ces tests supportent mieux la non-normalité que l'hétéroscédasticité.
(**) Ces tests travaillant sur les rangs et non sur les valeurs elles-mêmes, ils sont plus robustes et s'appliquent également à des variables ordinales
(***) ne pas comparer toutes les paires par des tests T $\Rightarrow$ on détecte à tort des différences significatives (au seuil de 95 % : dans 27 % des cas pour 4 moyennes égales)

# Test ANOVA à 1 facteur

- Test d'égalité de la moyenne d'une variable continue Y dans k ($\geq 2$) groupes (définis par les modalités d'une variable nominale)
  - si plusieurs variables continues dépendantes $\Rightarrow$ MANOVA
  - si m variables nominales indépendantes $\Rightarrow$ ANOVA à m facteurs
- Généralise le test de Student quand k > 2
- Ne teste que l'égalité de toutes les moyennes, sans dire le cas échéant lesquelles diffèrent
- Exemples :
  - comparer les productivités de plusieurs usines
  - comparer les rendements de plusieurs champs
  - comparer les effets de plusieurs engrais

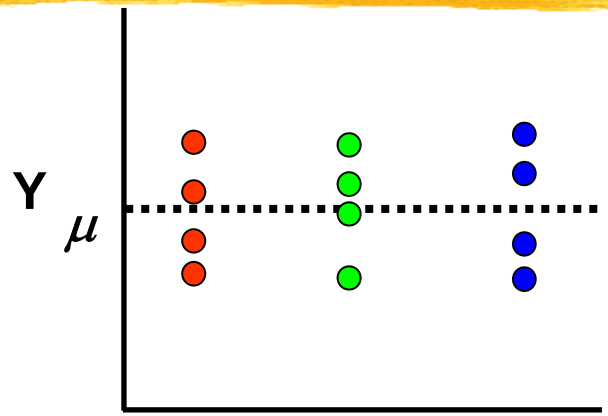
# ANOVA à 1 facteur : modèle général

- $\mathbf{Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}}$
- $Y_{ij}$ = valeur de l'obs. j dans le groupe i
- $\mu$ = moyenne générale de Y
- $\alpha_i$ = moyenne de Y dans le groupe i – $\mu$
- $\varepsilon_{ij}$ = valeur résiduelle
  - distribution normale dans tous les groupes (hypothèse la moins importante pour la qualité du test)
  - moyenne = 0
  - variance égale dans tous les groupes (homoscédasticité)
  - indépendance $\forall$ i, j
    - une observation ne doit pas dépendre des autres du groupe
    - les observations d'un groupe ne doivent pas dépendre de celles des autres groupes (cas d'un même individu présent plusieurs fois – cas de la comparaison de traitements)



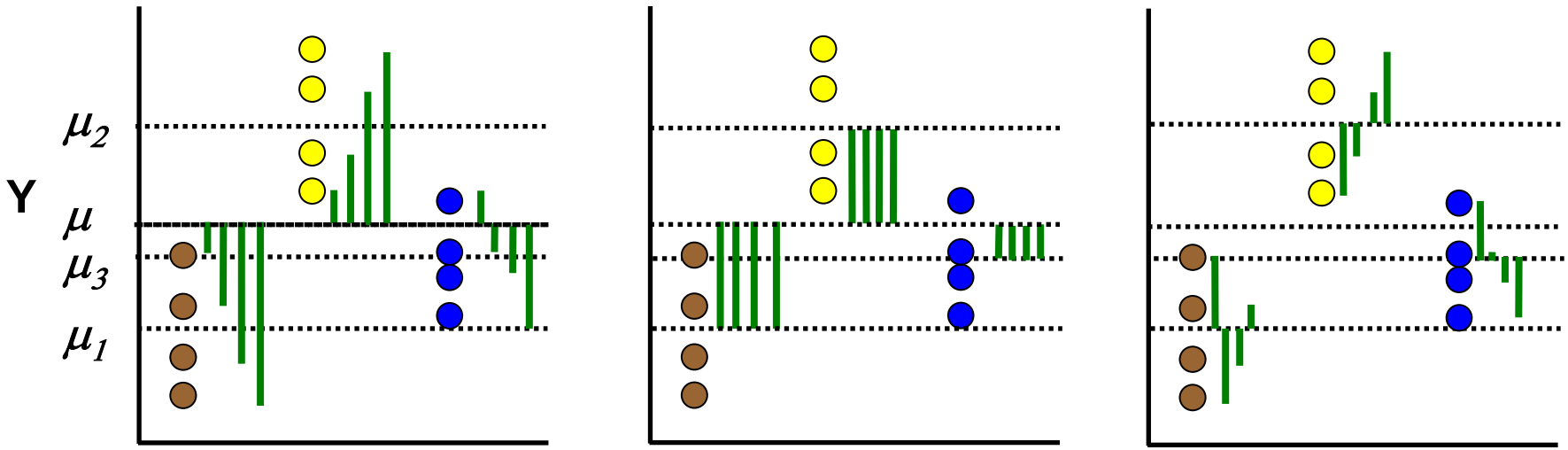
# Hypothèses de l'ANOVA

- $H_0$ : $\mu_1 = \mu_2 = ... = \mu_k$
  - les moyennes sont toutes égales
  - $\alpha_1 = \alpha_2 = ... = \alpha_k = 0$




- $H_1$ : les moyennes ne sont pas toutes égales
  - au moins une moyenne est différente
  - ne signifie pas : $\mu_1 \neq \mu_2 \neq ... \neq \mu_k$
  - pour déterminer quelles moyennes diffèrent significativement :
    - test de Bonferroni
    - test de Scheffé (plus puissant)

# Répartition de la somme des carrés




**SC Totale**

**SC Modèle (interclasse)**

**SC Erreur (intraclasse)**

- **Groupe 1**
- **Groupe 2**
- **Groupe 3**

# Tableau ANOVA et statistique F

| Source de variation | Somme des carrés (SC) | Degrés de liberté (*dl*) | Carré moyen (CM) | *F* |
|---|---|---|---|---|
| **Totale** | $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y})^2$ | ***n - 1*** | ***SC/dl*** | |
| **Inter-classe** | $\sum_{i=1}^{k} n_i(\overline{Y}_i-\overline{Y})^2$ | ***k - 1*** | ***SC/dl*** | $\frac{CM_{interclasse}}{CM_{intraclasse}}$ |
| **Intra-classe** | $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2$ | ***n - k*** | ***SC/dl*** | |

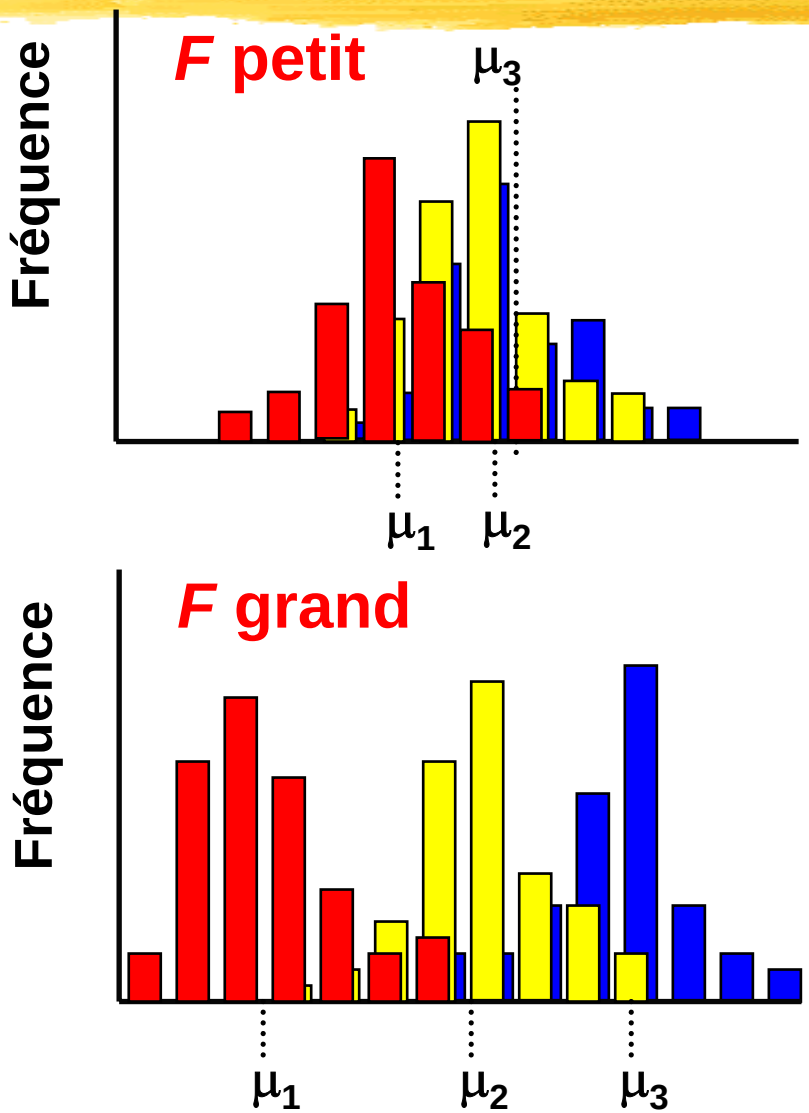
$CM_{inter}/CM_{intra}$ = F à comparer au F d'une loi de Fisher de ddl (k-1,n-k)

$\eta^2 = SC_{interclasse} / SC_{totale}$ = proportion de la variance expliquée

# Principe du test ANOVA

- On appelle « analyse de la variance » ce qui est en fait un test d'égalité de la moyenne, en raison de la façon de réaliser ce test, qui consiste à décomposer la variance de la variable continue Y en 2 parties :
  - ce qui peut être attribué aux différences entre groupes (variance inter-classe)
  - ce qui peut être attribué aux variations aléatoires (variance intra-classe, appelée « erreur »)
- Si $CM_{inter}/CM_{intra}$ = est grand, c.a.d. si les variations aléatoires sont faibles par rapport à l'effet des différences entre classes, on peut rejeter $H_0$
- Cela se produit quand $CM_{inter}/CM_{intra}$ dépasse la valeur critique de la loi de Fisher au niveau $\alpha$ avec k–1 et n–k degrés de liberté

# Illustration du test ANOVA



# Statistique de Mann-Whitney

- Utilisée pour k = 2 groupes, d'effectifs $n_1$ et $n_2$
  - quand les hypothèses de normalité et d'égalité des variances ne sont pas satisfaites
- Soit $R_i$ = somme des rangs des observations du groupe i
- La statistique du test comparée à une valeur théorique est:

$$U = \min\left\{ n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 , n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \right\}$$

- Avec les observations des 2 groupes $G_1$ et $G_2$ :

  $G_1$ : 3 5 6 10 14

  $G_2$ : 8 12 16 18

- On obtient les rangs

  $G_1$ : 1 2 3 5 7  $G_2$ : 4 6 8 9

- D'où $R_1$ = 18, $R_2$ = 27, U = min(20+15–18,20+10–27) = 3

$U_1$ = nb de fois où une valeur du groupe 1 précède une valeur du groupe 2

# Test non-paramétrique de Wilcoxon-Mann-Whitney

- Statistique de la somme des rangs de Wilcoxon = $R_i$
  - où i est soit le 1er, soit le plus petit (dans SAS) groupe
- Les groupes sont d'autant plus significativement différents :
  - que le U de Mann-Whitney est petit
  - que le S de Wilcoxon est très grand ou très petit
- À chacune de ces statistiques est associé un test dont l'hypothèse nulle est que les rangs du groupe 1 ne diffèrent pas des rangs du groupe 2
  - les tests sont équivalents $\Rightarrow$ test de Wilcoxon-Mann-Whitney
- On peut :
  - comparer U et S à des valeurs lues en table
  - ou, si $n_1$ et $n_2 > 8$, utiliser la convergence sous $H_0$ vers une loi normale $N(\mu,\sigma)$ et calculer $Z = (U-\mu)/\sigma$ et $|Z|$

# Test non-paramétrique de Kruskal-Wallis

- Utilisé pour k ≥ 2 groupes
  - quand les hypothèses de normalité et d'égalité des variances ne sont pas satisfaites
- Soient N = nb d'observations, $n_i$ l'effectif du groupe i et $R_i$ la somme des rangs des observations du groupe i
- La statistique du test est :

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

- Correctif à apporter en cas d'égalités de rangs
- Si les effectifs sont grands ou si k > 6, H tend vers $\chi^2$ à k–1 d° de liberté
  - sinon, regarder valeurs critiques dans une table

# Tests non-paramétriques sur SAS

- La PROC NPAR1WAY de SAS permet d'effectuer le test de Kruskal-Wallis et de Wilcoxon-Mann-Whitney (si k = 2)
  - `PROC NPAR1WAY WILCOXON data=table correct=no;`
  - `class a; /* variable de groupe */`
  - `var x; /* variable quantitative */`
  - `exact; /* test exact facultatif */`
  - `run;`
- Autres options que WILCOXON :
  - ANOVA : anova classique
  - EDF : tests de Kolmogorov-Smirnov et Cramer-von Mises, et, si 2 niveaux de classification seulement, statistique de Kuiper
  - MEDIAN : test de la médiane

# Résultats des tests avec option WILCOXON

| Wilcoxon Two-Sample Test | |
|---|---|
| **Statistic (S)** | 27.0000 |
| | |
| **Normal Approximation** | |
| **Z** | 1.7146 |
| **One-Sided Pr > Z** | 0.0432 |
| **Two-Sided Pr > \|Z\|** | 0.0864 |
| | |
| **t Approximation** | |
| **One-Sided Pr > Z** | 0.0624 |
| **Two-Sided Pr > \|Z\|** | 0.1248 |
| | |
| **Exact Test** | |
| **One-Sided Pr >= S** | 0.0556 |
| **Two-Sided Pr >= \|S - Mean\|** | 0.1111 |

| Wilcoxon Scores (Rank Sums) for Variable x Classified by Variable a | | | | | |
|---|---|---|---|---|---|
| **a** | **N** | **Sum of Scores** | **Expected Under H0** | **Std Dev Under H0** | **Mean Score** |
| **1** | 5 | 18.0 | 25.0 | 4.082483 | 3.600 |
| **2** | 4 | 27.0 | 20.0 | 4.082483 | 6.750 |

| Kruskal-Wallis Test | |
|---|---|
| **Chi-Square** | 2.9400 |
| **DF** | 1 |
| **Pr > Chi-Square** | 0.0864 |

$n_1$, $n_2$, $R_1$ et $R_2$ permettent de calculer U de Mann-Whitney

test non significatif : pas de différences entre les 2 groupes

# La préparation des données :
# Échantillonnage

# L'échantillonnage des données 1/2

- Étape incontournable de plusieurs techniques
  - notamment la prédiction et le classement, dont la plupart des algorithmes mettent en œuvre un *échantillon d'apprentissage* et un *échantillon de test*
  - panel de consommateurs
- Néanmoins il est parfois déconseillé d'effectuer toute une étude sur un échantillon seulement :
  - recherche de typologie de fraudes ou de segments étroits à forte valeur ajoutée
- Dans tous les cas, l'échantillonnage est une opération délicate, qui nécessite une bonne connaissance de la population étudiée

# L'échantillonnage des données 2/2

- Possible à condition :
  - de réussir à contrôler la représentativité de l'échantillon, dont les observations peuvent être extrapolées à l'ensemble de la population
  - d'avoir un nombre minimum d'individus dans l'échantillon (la précision ne croît que comme la racine carrée de l'effectif de l'échantillon)
  - de ne pas rechercher de phénomènes trop rares

- Types d'échantillonnage aléatoire :
  - simple
  - systématique
  - stratifié
  - par grappes

# Exemple d'échantillonnage

- Échantillon de clients, numérotés *aaaffffnn*
  - *aaa* = n° agence (de 1 à 999)
  - *ffff* = n° de foyer dans l'agence (de 1 à 9999)
  - *nn* = rang du client dans le foyer (1=H, 2=F, autres = enfants)
- Échantillonnage simple : tirage aléatoire du n° de client
- Échantillonnage systématique : 1er n° de client tiré aléatoirement, puis n°+k, n°+2k, etc. (NB : si k = 100 !)
- Échantillonnage stratifié : répartition des clients en tranches d'âge, puis n° de client tiré au sort
- Échantillonnage par grappes : tirage aléatoire de l'initiale du nom de famille, puis recensement (attention si l'initiale = « D » ou « L » !)

# Taille d'échantillon (pour un taux)

Quand un événement se produit dans une population avec une probabilité p (exemple : 80 % des clients sont satisfaits $\Rightarrow$ p = 0,8), cette probabilité peut être estimée à partir d'un échantillon de taille n de cette population. Cette probabilité p est estimée par la fréquence f = k/n de survenance de l'événement dans l'échantillon. Comme la variable k suit une loi binomiale B(n,p) de moyenne $\mu = n.p$ et de variance $\sigma^2 = n.p.(1-p)$, la fréquence f suit une loi binomiale de moyenne = p et de variance = p.(1-p)/n. On sait que lorsque n est grand, la loi binomiale tend vers une loi normale de paramètres $(\mu,\sigma)$. Sachant que 95 % des valeurs d'une loi normale $(\mu,\sigma)$ se trouvent dans l'intervalle $[\mu - 1{,}96\sigma , \mu + 1{,}96\sigma]$, la fréquence f a une probabilité de 95 % de se trouver dans l'intervalle de confiance :

$$[p - 1{,}96 \sqrt{\frac{p(1-p)}{n}} , p + 1{,}96 \sqrt{\frac{p(1-p)}{n}} ]$$

Donc l'intervalle

$$[f - 1{,}96 \sqrt{\frac{f(1-f)}{n}} , f + 1{,}96 \sqrt{\frac{f(1-f)}{n}} ]$$

a une probabilité proche de 95 % de contenir la vraie valeur de p.

On dit que l'intervalle ci-dessus est l'intervalle de confiance au seuil de risque de 5 % (le plus fréquemment utilisé). Au seuil de risque de 1 %, il faudrait remplacer la constante 1,96 ci-dessus par 2,5758.

# Taille d'échantillon (pour une moyenne)

Si l'on veut estimer la moyenne m d'une variable X dans la population entière, à partir des valeurs mesurées $x_1$, …, $x_n$ dans un échantillon de n individus, voici comment il faut procéder.

On calcule la moyenne $\mu = \frac{1}{n} \sum_i x_i$, puis l'écart-type *d'échantillon* σ selon la formule :

$$\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{n-1}}$$

Ensuite, si n ≤ 30, on va lire un paramètre $t_\alpha$ dans la table de la distribution de Student à n-1 degrés de liberté, en se fixant un seuil de risque α (généralement α = 0,05, c'est-à-dire 5 %). Si le test est bilatéral, il y a 2 zones de rejet, chacune avec une probabilité α/2, soit α au total. Dans ce cas, on remplace $t_\alpha$ par $t_{\alpha/2}$, et on regarde donc généralement $t_{0,025}$.

Si n > 30, c'est même plus simple, la loi de Student est approchée par la loi normale centrée réduite, et on va chercher $t_\alpha$ dans cette table ; en particulier, $t_{0,025} = 1,96$.

Enfin, on peut conclure qu'au seuil de risque α, la moyenne m est dans l'intervalle de confiance :

$$[\mu - t_\alpha \frac{\sigma}{\sqrt{n}} \ , \mu + t_\alpha \frac{\sigma}{\sqrt{n}} ].$$

# Retour au data mining

- Les considérations précédentes peuvent être utilisées pour interpréter une classification

- Pour chaque segment…
  - … et chaque variable continue : on compare sa moyenne dans le segment à sa moyenne générale
  - … et chaque variable catégorielle : on compare la proportion de chaque modalité dans le segment à sa proportion dans la population entière.

- On peut ainsi caractériser chaque segment par les variables qui le singularisent le + de la population entière