Fouille de données

Comprehension des données

Analyse exploratoire des données

Table of Contents

Définition

Le Datamining

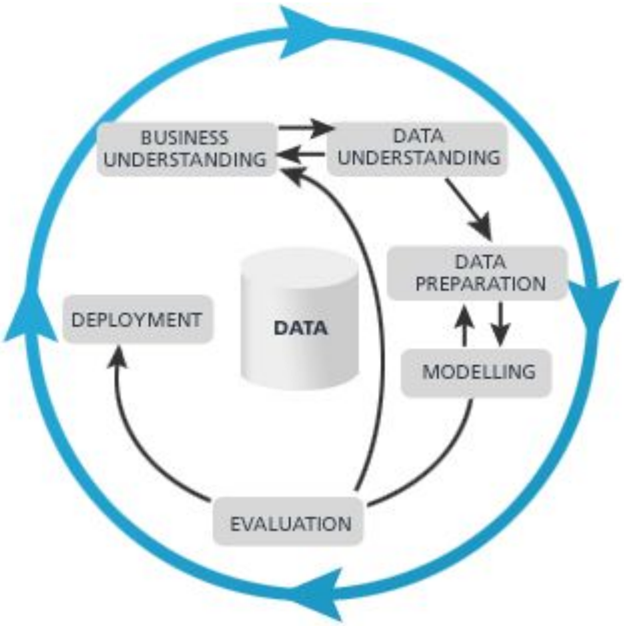
- Le datamining, ou fouille de données, est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de (souvent grandes) bases de données informatiques,
  - de façon automatique ou semi-automatique,
  - en vue de détecter dans ces données des règles, des associations, des tendances **inconnues ou cachées**,
  - des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données .
- Les apports
  - Tirer parti des informations historisées
  - Limiter la subjectivité humaine dans le processus de décision
- Le spectre d'application du datamining
  - gestion de la relation client,
  - e-commerce,
  - détection automatique de la fraude dans la téléphonie mobile ou l'utilisation des cartes bancaires,
  - contrôle qualité, pilotage de la production,
  - enquêtes en sciences humaines, études biologiques, médicales et pharmaceutiques ;
  - études agronomiques et agro-alimentaires,
  - détection de maladie (analyse de l'imagerie médicale)
  - prédiction d'audience TV.

Table of Contents

Le CRISP-DM

Les phases du CRISP-DM

- Le CRISP-DM (Cross Industry Standard Process for Data Mining) est considéré comme la méthodologie la plus utilisée pour le datamining :
  - Il a été conçu par un consortium d'entreprises pour être utilisé avec n'importe quel outil de datamining dans n'importe quel domaine.
  - Le CRISP-DM identifie six phases dans le processus de datamining



Le CRISP-DM

Les phases du CRISP-DM

- Compréhension du métier. Cette phase doit permettre de comprendre les objectifs et les besoins d'un point de vue métier,
    - et ainsi convertir cette connaissance en
    - une définition de problème de datamining,
    - et un plan permettant d'atteindre ces objectifs.
- Compréhension des données
    - La phase de compréhension des données commence avec une collecte des données
    - et se poursuit avec des activités qui ont pour objectif de se familiariser avec les données,
    - d'identifier les problèmes de qualité des données, découvrir les premières connaissances dans les données,
    - ou détecter les sous-ensembles intéressants pour former des hypothèses sur les informations cachées.
- Préparation des données
    - Cette phase couvre toutes les activités permettant de construire l'échantillon final (le jeu de données qui sera fourni à l'outil de modélisation).
    - Les tâches de préparation des données sont souvent exécutées plusieurs fois, et non dans un ordre prescrit.
    - Ces tâches incluent la sélection de tables, d'enregistrements et d'attributs, ainsi que la transformation et le nettoyage des données pour l'outil de modélisation.

Le CRISP-DM

Les phases du CRISP-DM

- Modélisation
  - Dans cette phase, diverses techniques de modélisation sont sélectionnées et appliquées, et leurs paramètres ajustés aux valeurs optimales.
  - Généralement, il existe plusieurs techniques pour le même type de problème de datamining.
  - Certaines techniques ont des exigences spécifiques sur la forme des données. Par conséquent, un pas en arrière vers la préparation des données est souvent nécessaire.
- Évaluation du modèle
  - A cette étape du projet, un (ou des) modèles qui semble avoir une haute qualité, du point de vue de l'analyse de données, a été construit.
  - Avant de procéder au déploiement final du modèle, il est important de bien évaluer le modèle, et de revoir les étapes exécutées pour construire le modèle, pour être certain qu'il \textbf{satisfait correctement les objectifs métiers}.
  - A la fin de cette phase, une décision sur l'utilisation ou non des résultats du processus datamining doit être prise.
- Utilisation du modèle
  - Les connaissances extraites des données doivent encore être organisées et présentées de façon à les rendre utilisables par les destinataires du modèle.
    - Fournir une synthèse descriptive des données
    - Mettre en œuvre un processus de fouille de données répétable.
  - Dans tous les cas, c'est l'utilisateur, et non l'analyste de données, qui va mettre en œuvre la phase de déploiement.
  - Il est néanmoins toujours important que l'utilisateur comprenne d'emblée quelles actions devront être menées afin de véritablement faire usage du modèle

Table of Contents

Compréhension des données

Les objectifs

- collecter les données,
- se familiariser avec les données,
- d'identifier les problèmes de qualité des données,
- de découvrir les premières connaissances dans les données,
- ou de détecter les sous-ensembles intéressants pour former des hypothèses sur les informations cachées.

Compréhension des données

Les tâches génériques

- **Collecter les données initiales**
  - Lister les jeux (ensembles) de données collectés, avec leurs localisations, les méthodes utilisées pour les collecter, et tout problème rencontré.
  - Les solutions proposées pour régler ces problèmes doivent aussi être enregistrées.
  - Cela continuera une base solide pour une réplication éventuelle du projet ou pour l'exécution d'un projet similaire dans le futur.
- **Décrire les données**
  - la description des attributs
  - le format des données,
  - la quantité de données (par exemple, nombre d'enregistrements et d'attributs dans chaque table),
  - Évaluer si les données collectées permettent de satisfaire les besoins identifiés
- **Explorer les données**
  - Cette tâche utilise les requêtes, la visualisation, et les techniques de reporting.
  - Cela inclut les distributions des attributs clés (e.g. la variable de réponse d'une tâche de prédiction).
  - Cela inclut aussi les relations entre paires (ou un nombre limité) de variables, des résultats d'agrégations simples, les propriétés d'une sous-population, et toute autre analyse statistique simple.
  - Ces analyses peuvent contribuer à affiner la description des données et peuvent fournir des entrées
    - à l'étape de vérification de la qualité des données
    - aux étapes de préparation des données (surtout celles liées à la transformation des données).

:

Compréhension des données

Les tâches génériques

- **Vérifier la qualité des données.** Examiner la qualité des données avec des questions comme:
  - Les données sont-elles complètes
    - Tous les cas requis sont-ils couverts?
  - Le jeux de données est-il correct, ou contient-il des erreurs,
    - s'il y a des erreurs, quelle est leur fréquence?
  - Les données contiennent-elles des des valeurs manquantes?
    - Le cas échéant, comment sont-elles représentées,
    - se produisent-elles,
    - et quelle est leur fréquence.
  -
-

:

Table of Contents

Les boîtes à moustaches

Box plots

- Les Box plots sont notre méthode privilégiée pour visualiser rapidement les statistiques récapitulatives d'une variable
- Les boîtes à moustaches comprennent :
  - la médiane (ligne médiane dans la case)
  - Intervalle inter-quartile (IQR, parties supérieure et inférieure des cases - c'est là que se trouvent 50 % de vos données)
  - Limites internes Q1-1,5*IQR et Q3+1,5*IQR
  - moustaches (les lignes noires s'étendant jusqu'aux valeurs les plus basses et les plus élevées qui se situent toujours dans les limites)
  - les valeurs aberrantes sont les données qui se situent en dehors des valeurs minimales et maximales internes

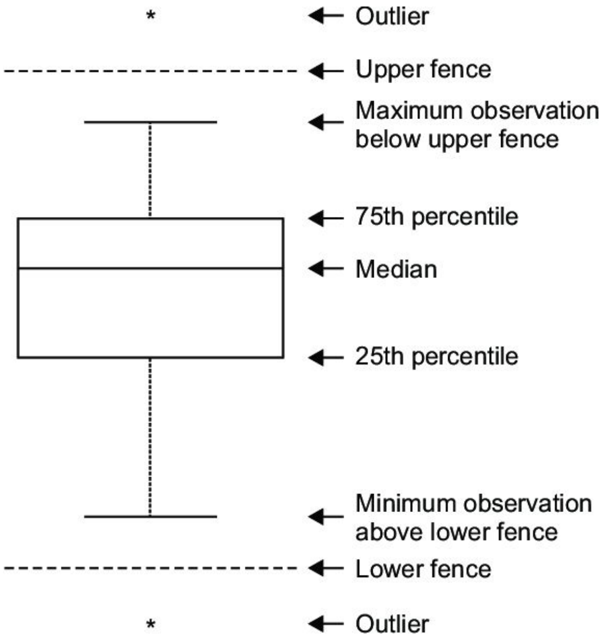


Table of Contents

Test d'hypothèse & significativité

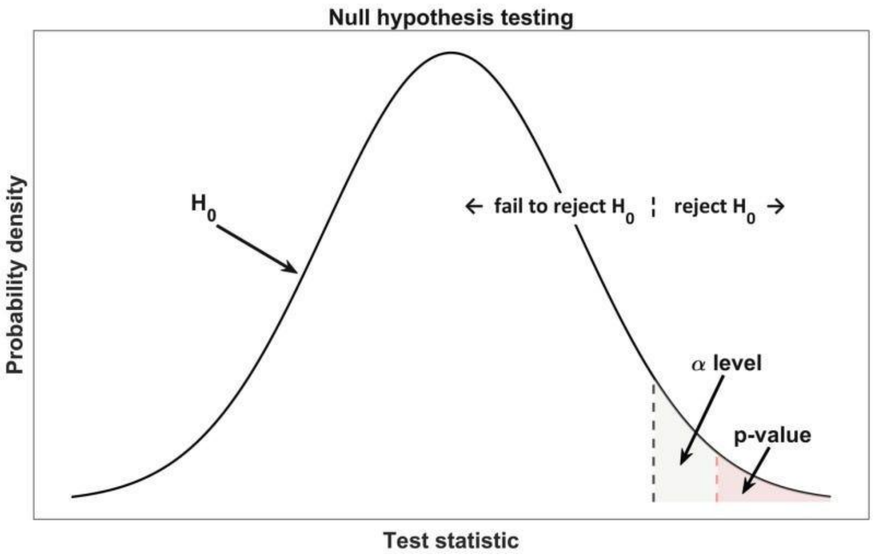
Test d'hypothèse

- Le test statistique cherche à départager deux hypothèses,
  - l'une appelée hypothèse nulle (H0)
  - et l'autre hypothèse alternative (H1)
  - Alors que l'hypothèse nulle est unique, l'hypothèse alternative correspond à une infinité de situations.
- Prise de décision et erreurs
  - mécanisme binaire de la prise de décision,
  - et les deux types d'erreur :
    - rejeter à tort une hypothèse,
    - ou accepter à tort une hypothèse

	on accepte H_0	on rejette H_0
H_0 est vraie	bonne décision	erreur (première espèce)
H_0 est fausse	erreur (seconde espèce)	bonne décision

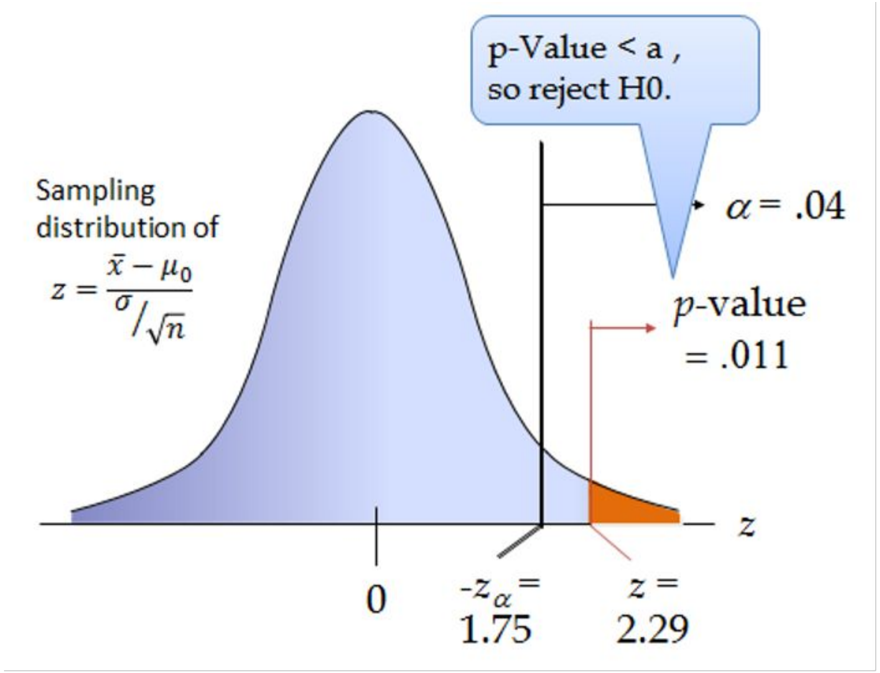
Test d'hypothèse & significativité

Test d'hypothèse

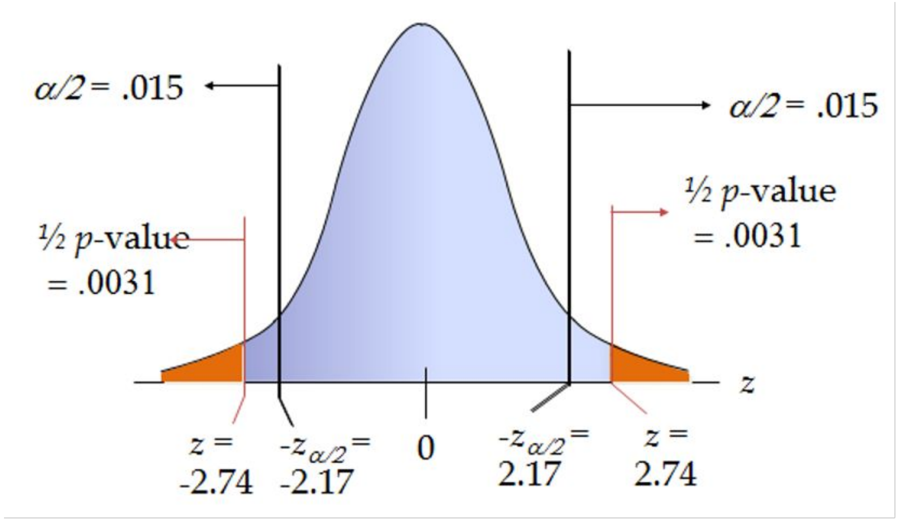


Test d'hypothèse & significativité

Test d'hypothèse

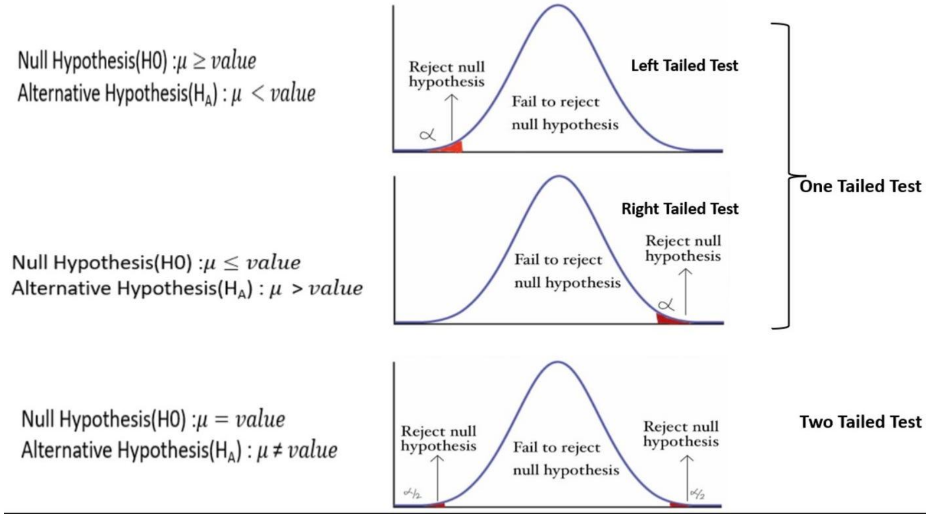


Test d'hypothèse



Test d'hypothèse & significativité

Test d'hypothèse



Test d'hypothèse

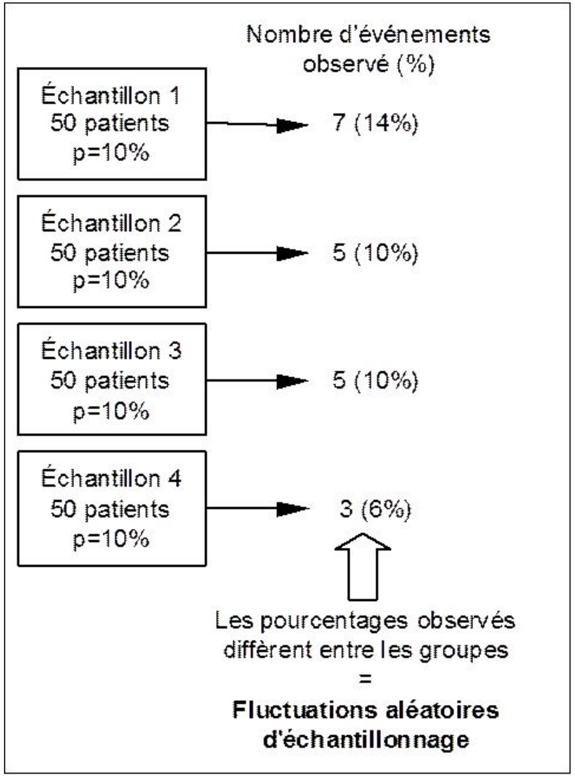
P-value	Decision
Less than 0.05*	**Reject Null** (H_0) Hypothesis Statistical difference between groups
Greater than 0.05*	**Fail to Reject** Null (H_0) Hypothesis No statistical difference between groups, or not enough evidence (data) to find a difference

* Assuming $\alpha = 0.05$

Test d'hypothèse & significativité

Concept de "significativité"

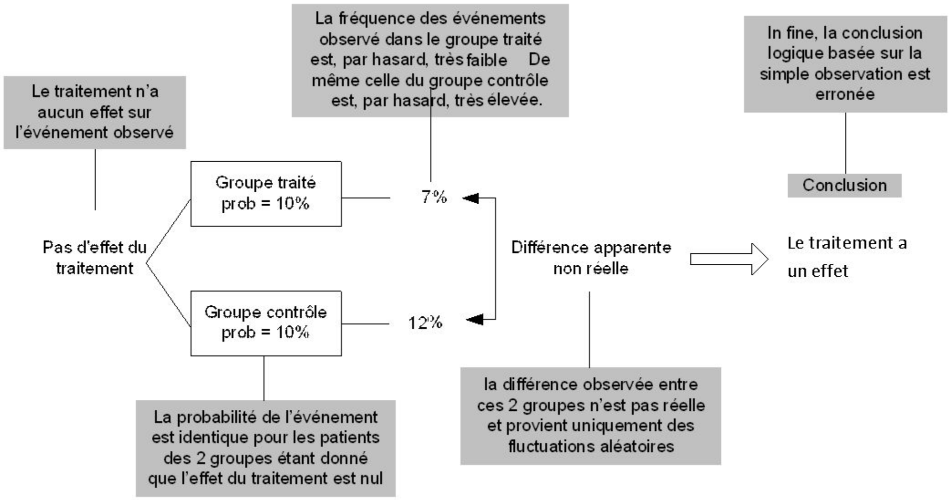
- Les fluctuations aléatoires
  - La survenue d'un événement clinique chez un patient est en partie imprévisible et s'apparente donc à un phénomène aléatoire.
  - Par exemple, la survenue sur une période de 5 ans d'un accident cardio-vasculaire chez un sujet hypertendu est imprévisible.
- Dans un essai thérapeutique,
  - l'hypothèse nulle correspond à l'absence d'effet du traitement étudié.
  - l'hypothèse alternative est l'hypothèse que l'on cherche à « prouver » : l'effet du traitement n'est pas nul.



Test d'hypothèse & significativité

Erreur alpha

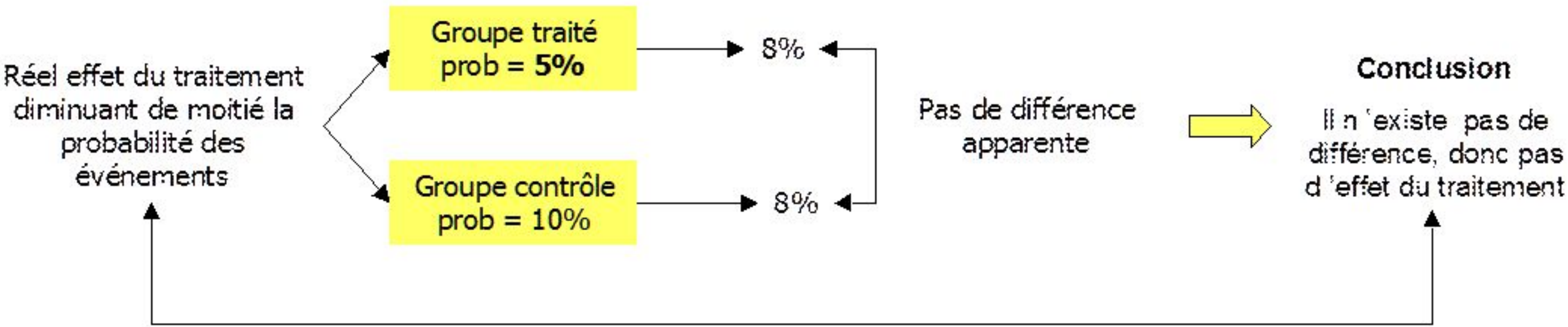
- Erreur alpha est de conclure à l'efficacité d'un traitement qui, en fait, est inefficace.



Test d'hypothèse & significativité

Erreur beta

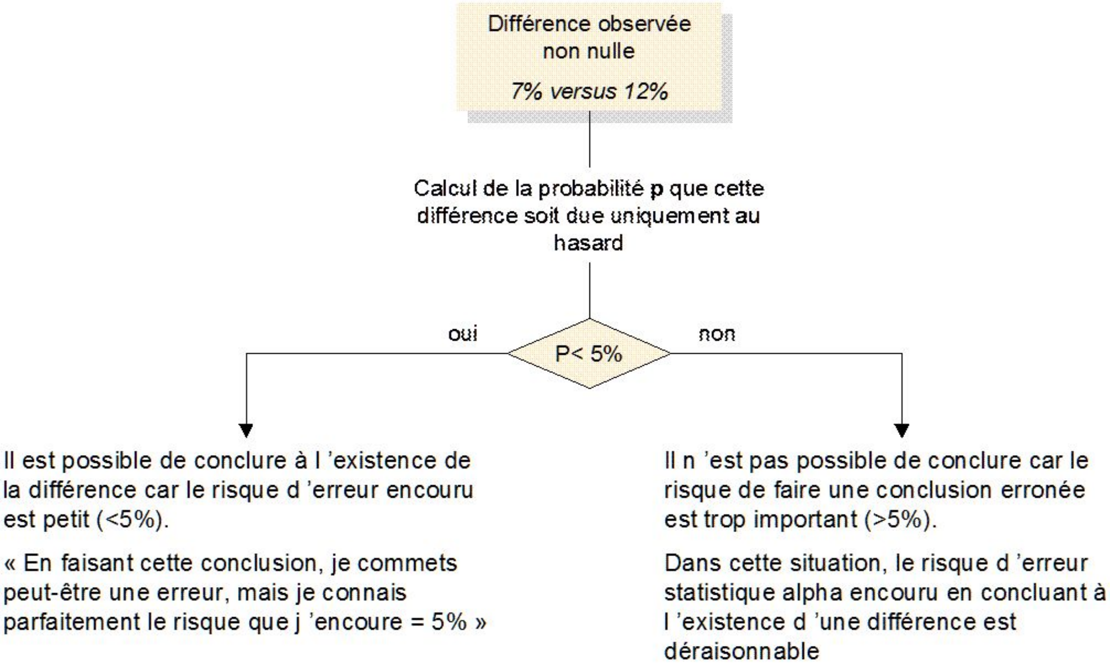
- Une erreur statistique bêta fait courir le risque de ne pas mettre en évidence l'efficacité d'un traitement.



Test d'hypothèse & significativité

Seuil de risque

- Seuil de risque de conclusion erronée acceptable (seuil de risque alpha = 5 %)
- Lorsque $p \leq 5$ % la différence est dite « statistiquement significative ».
- fonction p-value dans R



Test ANOVA

Données de l’exemple

Source of Variability	SS	df	MS	F
Between the Treatment Groups	$\sum n_i(\bar{X}_{i.} - \bar{X}_{..})^2$	$k - 1$	$\frac{\text{SS between}}{\text{df between}}$	$\frac{\text{MS between}}{\text{MS within}}$
Within the Treatment Groups	$\sum\sum(X_{ij} - \bar{X}_{i.})^2$	$N - k$	$\frac{\text{SS within}}{\text{df within}}$	
Total	$\sum\sum(X_{ij} - \bar{X}_{..})^2$	$N - 1$		

Note that:
SS total = SS between + SS within
df total = df between + df within

n_i is the number of people in treatment level i
$\bar{X}_{i.}$ is the mean for treatment level i
$\bar{X}_{..}$ is the grand mean for the entire study
X_{ij} is the score for the j^{th} person in the i^{th} treatment level
k is the number of treatment levels
N is the total number of people in the entire experiment
SS stands for summed-squared deviations of the expected value from the observed value
MS stands for mean-squared deviations of the expected value from the observed value
$F_{obtained}$ is compared to the $F_{critical}$(df between, df within) found in the F table.

Table of Contents

Données de l’exemple

1	**2**	**3**
1	2	2
2	4	3
5	2	4

Test ANOVA

etape 1

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu_3 \\ H_a &: au\ moins\ un\ diff\acute{e}rence\ entre\ les\ moyennes \end{aligned} \tag{1}$$

$$\alpha = 0.5$$

Test ANOVA

etapes 2.a et 2.c

$$
\begin{aligned}
df_{between} &= k - 1 = 3 - 1 = 2 \\
df_{within} &= N - k = 9 - 3 = 6 \\
df_{total} &= 8 \\
F_{crit} &= 5.14
\end{aligned}
\tag{2}
$$

Test ANOVA

etape 2.b

Table F The *F* Distribution

$\alpha = .05$

df_D \ df_N	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14

Test ANOVA

etape 3.a

$$
\begin{aligned}
\bar{X}_1 &= 2.67 \\
\bar{X}_2 &= 2.67 \\
\bar{X}_3 &= 3.00 \\
\bar{\bar{X}} &= \frac{S_{totale}}{N} = \frac{25}{9} = 2.78
\end{aligned}
\tag{3}
$$

Test ANOVA

etape 3.b

$$
\begin{aligned}
SS_{total} = \Sigma(X_{ij} - \bar{\bar{X}})^2 = & (1 - 2.78)^2 + (2 - 2.78)^2 + (5 - 2.78)^2 \\
& (2 - 2.78)^2 + (4 - 2.78)^2 + (2 - 2.78)^2 \\
& (2 - 2.78)^2 + (3 - 2.78)^2 + (4 - 2.78)^2 \\
& SS_{total} = 13.56
\end{aligned} \tag{4}
$$

Test ANOVA

etape 3.c

$$
\begin{aligned}
SS_{within} = \Sigma\Sigma(X_{ij} - \bar{X}_i)^2 = & (1 - 2.67)^2 + (2 - 2.67)^2 + (5 - 2.67)^2 \\
& (2 - 2.67)^2 + (4 - 2.67)^2 + (2 - 2.67)^2 \\
& (2 - 3)^2 + (3 - 3)^2 + (4 - 3)^2 \\
& SS_{within} = 13.33
\end{aligned} \tag{5}
$$

Test ANOVA

etape 3.d

$$
\begin{aligned}
SS_{between} &= SS_{total} - SS_{within} \\
SS_{between} &= 13.56 - 13.33 = 0.23
\end{aligned}
\tag{6}
$$

Test ANOVA

etape 4

$$MS_{between} = \frac{SS_{between}}{df_{between}} = \frac{0.23}{2} = 0.12 \tag{7}$$

$$MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{13.34}{6} = 2.22 \tag{8}$$

Test ANOVA

etape 5

$$\begin{aligned} &F = \frac{MS_{between}}{MS_{within}} = \frac{0.12}{2.22} = 0.05 \\ &F < F_{crit} \\ &Ne\ pas\ rejeter\ H_0 \\ &\surd\ H_0\ :\ \mu_1 = \mu_2 = \mu_3 \end{aligned} \tag{9}$$

P-value

- Si la valeur p d'une ANOVA est inférieure à . 05, nous rejetons alors l’hypothèse nulle selon laquelle la moyenne de chaque groupe est égale.
- Dans ce scénario, nous pouvons ensuite effectuer le test de Tukey pour déterminer exactement quels groupes diffèrent les uns des autres.

Table of Contents

Test TUKEY

Formule

$$q_{A,B} = \frac{|meanA - meanB|}{\sqrt{\frac{\text{within-group SSE}}{n}}}$$

- n : number of samples per group

Table of Contents

Test Khi-carré

- Le test du χ^2 ne peut s'utiliser que lorsque le statisticien veut étudier la relation ou pas entre deux caractères qualitatifs.
- Si la valeur de χ^2 dans un tableau de contingence est supérieur à la valeur critique correspondant à une probabilité donnée (hypothèse nulle), l'hypothèse nulle (le variables sont indifférentes) est rejetée.
- Pour les tables de contingence on calcule la statistique χ^2 comme
- nij est l'effectif observé
- Dans les tableaux de contingence le nombre de degrés de liberté est égale au produit de nombre de colonne moins un et nombre de lignes moins un

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

Où n* es l'effectif théorique c'est-à-dire l'effectif que l'on aurait eu si les variables étaient indépendantes :

$$n^*_{ij} = \frac{n_{i.} n_{.j}}{n}$$

L'analyse exploratoire des données

Test Khi-carré

- La table du KHI-DEUX nous indique la valeur seuil pour un risque de première espèce α et un ddl
- On va comparer cette valeur seuil fournie par la table avec la valeur obtenue par notre calcul.
- Si la valeur calculée est supérieure au seuil fourni par la table, la conclusion à tirer est que les deux variables ne sont pas indépendantes, mais étroitement liées.

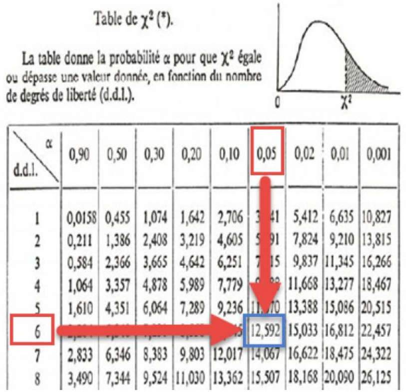


Table de χ^2 (*).

La table donne la probabilité α pour que χ^2 égale ou dépasse une valeur donnée, en fonction du nombre de degrés de liberté (d.d.l.).

d.d.l. \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0158	0,455	1,074	1,642	2,706	[illegible]	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	[illegible]	7,824	9,210	13,815
3	0,584	2,366	3,665	4,642	6,251	[illegible]	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	[illegible]	11,668	13,277	18,467
5	1,610	4,351	6,064	7,289	9,236	[illegible]	13,388	15,086	20,515
6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,490	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125

Test Khi-carré

- H0 : Il n'y a pas de relation entre le sexe et les problèmes à l'école
- H1 : Il y a une relation entre le sexe et les ennuis à l'école
- nous fixerons notre niveau alpha à 0,05

	Got in Trouble	Did Not Get in Trouble	Total
Boys	46	71	117
Girls	37	83	120
Total	83	154	237

Test Khi-carré

- Pour obtenir le nombre attendu pour la cellule supérieure droite,
  - nous multiplierons le total de la ligne (117) par le total de la colonne (83)
  - et diviser par le nombre total d'observations (237). (83 × 117)/237 = 40,97.
- Si les deux variables étaient indépendantes, nous nous attendons à ce que 40,97 garçons aient des ennuis.

	Got in Trouble	Did Not Get in Trouble	Total
Boys	46 (40.97)	71 (76.02)	117
Girls	37 (42.03)	83 (77.97)	120
Total	83	154	237

Test Khi-carré

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$$x^2 = \frac{(46-40.97)^2}{40.97} + \frac{(37-42.03)^2}{42.03} + \frac{(71-76.03)^2}{76.03} + \frac{(83-77.97)^2}{77.97}$$

$$x^2 = 1.87$$

Test Khi-carré

- le nombre de degrés de liberté est égal au nombre de colonnes du tableau moins un multiplié par le nombre de lignes du tableau moins un, soit (r-1)(c-1).
  - Dans notre cas, nous avons (2-1)(2-1), soit un degré de liberté.
- la statistique critique pour un niveau alpha de 0,05 et un degré de liberté est de 3,841, ce qui est supérieur à notre statistique obtenue de 1,87.
- la statistique critique étant supérieure à notre statistique obtenue, nous ne pouvons pas rejeter notre hypothèse nulle.

Table of Contents

t-Test in Linear Regression

- **Linear regression** is used to quantify the relationship between a predictor variable and a response variable
- We want to know if there is a statistically significant relationship between the predictor variable and the response variable.
- We use the following null and alternative hypothesis for this t-test:
  - $\mathbf{H_0}$: $\beta_1 = 0$ (the slope is equal to zero)
  - $\mathbf{H_A}$: $\beta_1 \neq 0$ (the slope is not equal to zero)
- We then calculate the test statistic as follows:
  - $\boldsymbol{t = b / SE_b}$
  - ***b***: coefficient estimate
  - $\mathbf{SE}_{\boldsymbol{b}}$: standard error of the coefficient estimate
- We then calculate the p-value that corresponds to *t* :
  - **Use the t distribution table** (i.e. a table that shows the critical values of the t distribution)
  - Calculate the P-Value of a T-Score in R using pt() function : **pt(q, df, lower.tail = TRUE)**
- If the p-value that corresponds to *t* is less than some threshold (e.g. α = .05) then
  - we reject the null hypothesis
  - and conclude that there is a statistically significant relationship between the predictor variable and the response variable.

The t-Value for Multiple Linear Regression

- the sum of the squared residuals

$$\hat{\sigma}^2 = \frac{\sum_i \hat{u}_i^2}{n-(k+1)} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-(k+1)}$$

$$\widehat{\mathbf{u}}'\widehat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\widehat{\beta})'(\mathbf{y} - \mathbf{X}\widehat{\beta})$$

The t-Value for Multiple Linear Regression

- Consider testing a hypothesis about a single regression coefficient βj
- Consider t-value about βj for the multiple regression
- How do we compute the Standard Errors
- Take the variance-covariance matrix of βˆ and square root the diagonal element corresponding to j.

$$H_0 : \beta_j = c$$

$$T = \frac{\hat{\beta}_j - c}{\hat{SE}(\hat{\beta}_j)}$$

$$\hat{SE}(\hat{\beta}_j) = \sqrt{\widehat{V}(\hat{\beta}_j)} = \sqrt{\widehat{V}(\hat{\beta})_{(j,j)}} = \sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{(j,j)}}$$

t-Test in Linear Regression : example

- We want to analyze the relationship between hours studied and exam score received for 40 of his students.
- He performs simple linear regression using hours studied as the predictor variable and exam score received as the response variable.
- We then calculate the test statistic as follows:
  - $t = b / SE_b$
  - $t = 1.117 / 1.025$
  - $t = 1.089$
- The p-value that corresponds to $t = 1.089$ with df = n-2 = 40 – 2 = 38 is **0.283**
- Since this p-value is not less than .05, we fail to reject the null hypothesis.
- This means that hours studied *does not* have a statistically significant relationship with final exam score.

	Coefficients	Standard Error	t Stat	p-value
Intercept	66.99	6.211	10.785	<0.000
Hours	1.117	1.025	1.089	0.283

Références bibliographiques

Références

- Practical Regression and Anova using R, Julian J. Faraway July 2002