Tests Statistiques

Benoit Gaüzère, Stéphane Canu
benoit.gauzere@insa-rouen.fr

INSA Rouen Normandie - ITI

16 mai 2024

Tests Statistiques : pour quoi faire ?

Question :

Vous observez deux groupes d'étudiants : Le premier groupe révise M8, le deuxième non. Au final, le premier groupe a une moyenne égale à a, le deuxième égale à b.
Que concluez vous si :

1. $a = 15$, $b = 8$

Tests Statistiques : pour quoi faire ?

Question :

Vous observez deux groupes d'étudiants : Le premier groupe révise M8, le deuxième non. Au final, le premier groupe a une moyenne égale à a, le deuxième égale à b.
Que concluez vous si :

1. $a = 15$, $b = 8$
2. $a = 13$, $b = 13$

Tests Statistiques : pour quoi faire ?

Question :

Vous observez deux groupes d'étudiants : Le premier groupe révise M8, le deuxième non. Au final, le premier groupe a une moyenne égale à a, le deuxième égale à b.
Que concluez vous si :

1. $a = 15$, $b = 8$
2. $a = 13$, $b = 13$
3. $a = 13,5$, $b = 12,8$

Tests Statistiques : pour quoi faire ?

Statistiques et hasard

- Observations de phénomènes complexes et aléatoires
- Comment avoir une idée si les différences sont dues au hasard ou à une différence réelle ?

Objectifs

- Confronter une hypothèse aux données observées
- Aider à la prise de décision

Introduction aux tests statistiques

Exemple introductif : Test d'une proportion

sur 1000 tirages, j'ai observé :

502 fois pile avec la pièce 1

763 fois pile avec la pièce 2

521 fois pile avec la pièce 3

Deux hypothèses

- $\mathcal{H}_0$ la pièce est normale
- $\mathcal{H}_1$ la pièce est biaisée

Exemple introductif : test d'une proportion I

Hypothèses

$$\begin{cases} \mathcal{H}_0 : \text{la pièce est normale} \\ \mathcal{H}_1 : \text{la pièce est biaisée} \end{cases}$$

L'hypothèse nulle $\mathcal{H}_0$

- $\mathcal{H}_0$ et $\mathcal{H}_1$ ne sont pas interchangeables
- Hypothèse nulle $\mathcal{H}_0$ est l'hypothèse de base, le statu quo.
- Hypothèse alternative $\mathcal{H}_1$ est généralement la négation de $\mathcal{H}_0$
- $\mathcal{H}_0$ sera réfutée si des éléments factuels viennent en contradiction.

Exemple introductif : test d'une proportion II

Modèle : observation et reformulation

Afin de pouvoir décider, nous allons observer l'échantillon $X_1, ..., X_n$ i.i.d. de loi parente $\mathcal{B}(p)$ (une loi de Bernoulli de paramètre p).

$$\begin{cases} \mathcal{H}_0 : \text{la pièce est normale} & \Leftrightarrow \quad p = \frac{1}{2} \\ \mathcal{H}_1 : \text{la pièce est biaisée} & \Leftrightarrow \quad p \neq \frac{1}{2} \end{cases}$$

Exemple introductif : test d'une proportion

Règle de décision

Une règle de décision raisonnable :
Si la fréquence de "pile" est proche de $\frac{1}{2}$ nous allons décider que la pièce est normale.

Problèmes

Même si $p = \frac{1}{2}$, il est toujours possible d'observer n fois pile ...

Comment décider raisonnablement dans ce contexte ?

Construire une règle de décision I

Définition : Règle de décision

Une fonction de l'ensemble des observations vers celui des décisions

Probabilité d'observer l'échantillon

- En supposant que la pièce est normale
- On calcule la probabilité d'obtenir l'échantillon observé
- L'hypothèse nulle $\mathcal{H}_0$ est rejetée si cette probabilité est inférieure à un seuil (0,05 par ex.)

Construire une règle de décision II

Alternative

- On fixe un seuil s directement par rapport à la fréquence observée

$$\text{si } \frac{1}{2} - s < \frac{1}{n}\sum_{i=1}^{n} X_i < \frac{1}{2} + s \quad \text{on décide } \mathcal{H}_0$$

- les seuils sont calculés en se fixant un taux d'erreur

$$\mathbb{P}(\frac{1}{2} - s < \frac{1}{n}\sum_{i=1}^{n} X_i < \frac{1}{2} + s) = 1 - \alpha$$

L'erreur $\alpha = \mathbb{P}(\text{décider } \mathcal{H}_1 \mid \mathcal{H}_0 \text{ est vraie})$

Quelle erreur contrôler ? I

Réalités et décisions

2 possibilités / 2 décisions :

$$\begin{cases} \mathcal{H}_0 : \text{la pièce est normale} \\ \mathcal{H}_1 : \text{la pièce est biaisée} \end{cases} / \begin{cases} \mathcal{D}_0 : \text{on décide que la pièce est normale} \\ \mathcal{D}_1 : \text{on décide que la pièce est biaisée} \end{cases}$$

Deux types d'erreurs possibles

α ce produit N'est PAS efficace - Mais on décide de le fabriquer

$$\boxed{\alpha = \mathbb{P}(\text{décider } \mathcal{D}_1 \mid \mathcal{H}_0 \text{ est vraie})}$$

on décide de ne pas soigner quelqu'un de malade

β ce produit EST efficace - Mais on décide de ne pas le fabriquer

$$\boxed{\beta = \mathbb{P}(\text{décider } \mathcal{D}_0 \mid \mathcal{H}_1 \text{ est vraie})}$$

on décide de soigner quelqu'un qui n'est pas malade

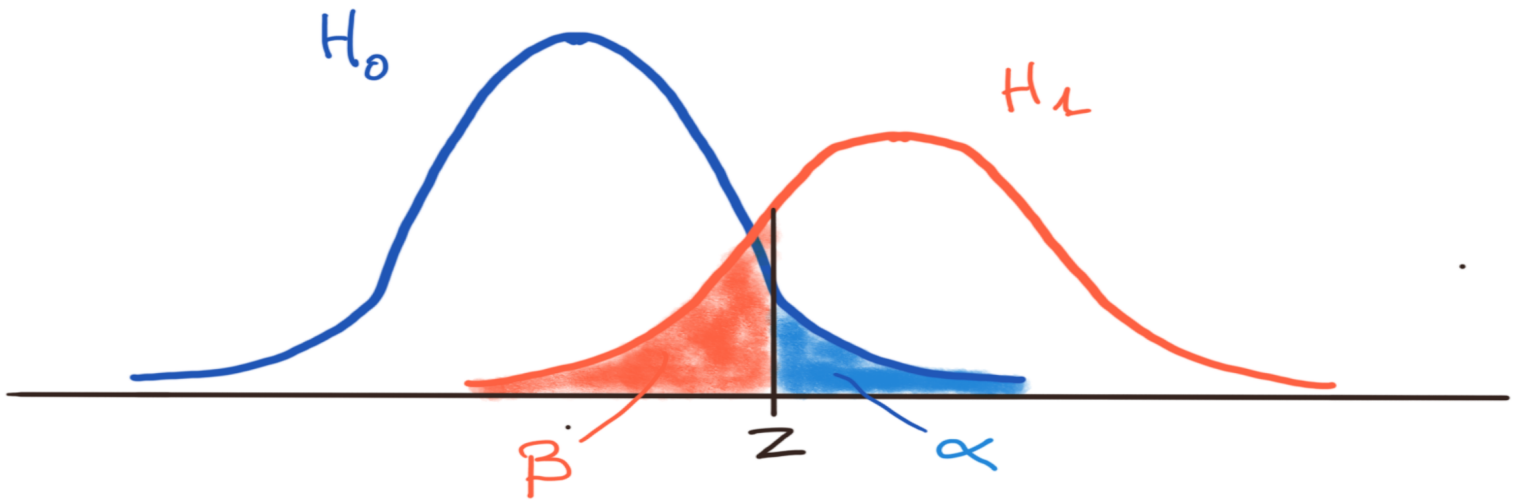
Quelle erreur contrôler ? II

Les types d'erreurs

	$\mathcal{H}_0$	$\mathcal{H}_1$
$\mathcal{D}_0$	0	β
$\mathcal{D}_1$	α	0

- α : Risque de première espèce
- β : Risque de seconde espèce
- $1-\beta$: Puissance du test (Si $\mathcal{H}_1$ est bien spécifiée). Rejeter $\mathcal{H}_0$ alors qu'elle est fausse.

Risques α et β



Remarques sur l'expérience I

Comment fixer les seuils ?

- Minimiser un cout :
  $C_{00}\mathbb{P}(\mathcal{D}_0 \mid \mathcal{H}_0) + C_{01}\alpha + C_{10}\beta + C_{11}\mathbb{P}(\mathcal{D}_1 \mid \mathcal{H}_1)$
  - $\alpha = \mathbb{P}(\mathcal{D}_1 \mid \mathcal{H}_0) = \mathbb{P}(\text{décider } \mathcal{D}_1 \mid \mathcal{H}_0 \text{ est vraie})$
  - C_{01} : le cout associé à la fabrication d'un produit moins efficace
- Contrôler une erreur
  - pour α fixé, on construit une règle de décision qui minimise β
  - le résultat est dit **statistiquement significatif** lorsqu'il est improbable qu'il puisse être obtenu par un simple hasard.

Remarques sur l'expérience II

Savoir rejeter

- pour α et β fixés, on construit une règle qui augmente n si ces valeurs ne sont pas atteintes
- Trois décisions possibles :
  - On décide $\mathcal{D}_0$
  - On décide $\mathcal{D}_1$
  - On décide d'effectuer une nouvelle expérience (on retire une pièce)

La pièce honnête : n tirages

$n = 1$ Il n'est pas raisonnable de décider quoi que ce soit dans ce cas

La pièce honnête : n tirages

$n = 1$ Il n'est pas raisonnable de décider quoi que ce soit dans ce cas

$n = 2$ trois cas possibles : il semble préférable d'augmenter n.

La pièce honnête : n tirages

$n = 1$ Il n'est pas raisonnable de décider quoi que ce soit dans ce cas

$n = 2$ trois cas possibles : il semble préférable d'augmenter n.

$n = 3$ quatre cas possibles : il semble toujours préférable d'augmenter n.

La pièce honnête : n tirages

$n = 1$ Il n'est pas raisonnable de décider quoi que ce soit dans ce cas

$n = 2$ trois cas possibles : il semble préférable d'augmenter n.

$n = 3$ quatre cas possibles : il semble toujours préférable d'augmenter n.

$n = 4$ $\mathbb{P}(\text{quatre piles}) = \frac{1}{16}$.
" si j'observe quatre piles ou quatre face je décide que la pièce est biaisée "
$\Rightarrow$ je vais me tromper une fois sur huit (pour les pièces normales)

La pièce honnête : n tirages

$n = 1$ Il n'est pas raisonnable de décider quoi que ce soit dans ce cas

$n = 2$ trois cas possibles : il semble préférable d'augmenter n.

$n = 3$ quatre cas possibles : il semble toujours préférable d'augmenter n.

$n = 4$ $\mathbb{P}(\text{quatre piles}) = \frac{1}{16}$.
" si j'observe quatre piles ou quatre face je décide que la pièce est biaisée "
$\Rightarrow$ je vais me tromper une fois sur huit (pour les pièces normales)

$n = 10$ " si j'observe dix piles ou dix faces je décide que la pièce est biaisée " $\Rightarrow$ je vais me tromper une fois sur $2^9 = 512$ (pour les pièces normales). Mais que ce passe t'il si la pièce est biaisée ?

La pièce honnête : n tirages

$n = 1$ Il n'est pas raisonnable de décider quoi que ce soit dans ce cas

$n = 2$ trois cas possibles : il semble préférable d'augmenter n.

$n = 3$ quatre cas possibles : il semble toujours préférable d'augmenter n.

$n = 4$ $\mathbb{P}(\text{quatre piles}) = \frac{1}{16}$.
" si j'observe quatre piles ou quatre face je décide que la pièce est biaisée "
$\Rightarrow$ je vais me tromper une fois sur huit (pour les pièces normales)

$n = 10$ " si j'observe dix piles ou dix faces je décide que la pièce est biaisée " $\Rightarrow$ je vais me tromper une fois sur $2^9 = 512$ (pour les pièces normales). Mais que ce passe t'il si la pièce est biaisée ?

$n = 100$ cent-un cas possibles. on décide que la pièce est fausse lorsque le nombre de piles observés est inférieur à k ou supérieur à $n - k$. on fixe k de sorte que $\mathbb{P}(\sum_{i=1}^{n} X_i < k) = \frac{\alpha}{2}$. Mais que ce passe t'il si la pièce est biaisée ?

La pièce honnête : n tirages

$n = 1$ Il n'est pas raisonnable de décider quoi que ce soit dans ce cas

$n = 2$ trois cas possibles : il semble préférable d'augmenter n.

$n = 3$ quatre cas possibles : il semble toujours préférable d'augmenter n.

$n = 4$ $\mathbb{P}(\text{quatre piles}) = \frac{1}{16}$.
" si j'observe quatre piles ou quatre face je décide que la pièce est biaisée "
$\Rightarrow$ je vais me tromper une fois sur huit (pour les pièces normales)

$n = 10$ " si j'observe dix piles ou dix faces je décide que la pièce est biaisée " $\Rightarrow$ je vais me tromper une fois sur $2^9 = 512$ (pour les pièces normales). Mais que ce passe t'il si la pièce est biaisée ?

$n = 100$ cent-un cas possibles. on décide que la pièce est fausse lorsque le nombre de piles observés est inférieur à k ou supérieur à $n - k$. on fixe k de sorte que $\mathbb{P}(\sum_{i=1}^{n} X_i < k) = \frac{\alpha}{2}$. Mais que ce passe t'il si la pièce est biaisée ?

$n = 1000$ si la pièce est biaisée, il faut que le biais soit tout petit pour que je ne le détecte pas. . .

La P-valeur (p-value) I

Définition : p-valeur

la p-valeur de l'échantillon par rapport au test c'est la probabilité de tirer un échantillon encore plus rare que celui observé.

Exemple

Si, sur 10 tirages d'un pièce on obtient 2 piles, alors la p-valeur p du test de " loyauté " de la pièce est :

$$p = \mathbb{P}(N = 0)+\mathbb{P}(N = 1)+\mathbb{P}(N = 2)+\mathbb{P}(N = 8)+\mathbb{P}(N = 9)+\mathbb{P}(N = 10)$$

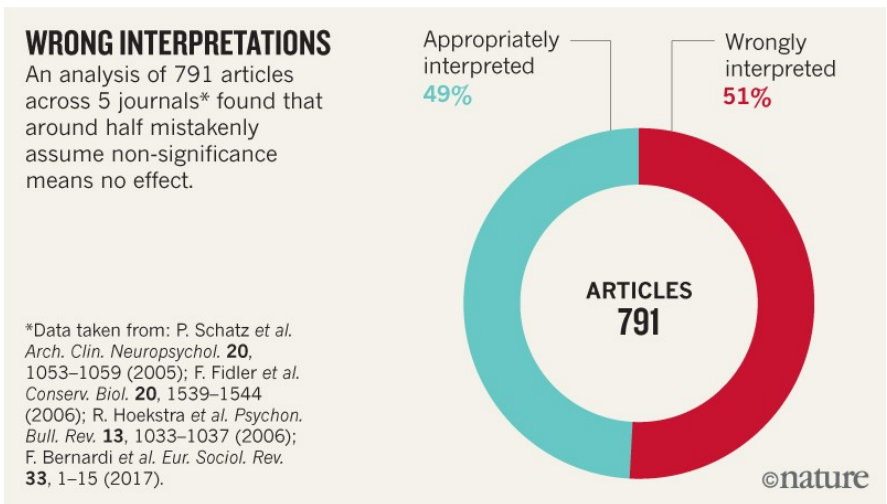
où N est une v.a. distribuée suivant une loi de Bernoulli de paramètre $1/2$.

La P-valeur (p-value) II

Interprétation

- La p-valeur indique dans quelle mesure les observations sont conformes à l'hypothèse H_0
- La p-valeur n'indique pas la probabilité que l'hypothèse H_0 soit vraie.

Attention aux utilisations abusives/mauvaises



https://www.nature.com/articles/d41586-019-00857-9

Démarche d'un test statistique

Schéma Global

1. Choix de H_0 et H_1.
2. Déterminer α et β.
3. Choix de la statistique du test (on y reviendra)
4. Calculer la p-value associée aux données
5. Décider

- Généralement, les conditions du test sont déterminées **avant** la récolte des données

Exemple

Les données

Le niveau de pluie annuel X, établi par des météorologues, suit une loi normale $\mathcal{N}(600, 100^2)$. Une entreprise propose une méthode pour augmenter le niveau de pluie de 50 mm par an, et donc suivre une loi normale $\mathcal{N}(650, 100^2)$. Leur procédé est testé pendant 9 années. On mesura les niveaux de pluie suivants :

Année	1	2	3	4	5	6	7	8	9
Pluie (mm)	510	614	780	512	501	534	603	788	650

Le problème

- On décide d'investir dans la solution si les faits permettent de conclure à la vraisemblance de l'augmentation du niveau de pluie.
- Niveau de risque : 5 %.

Exemple I

Les hypothèses

- $\mathcal{H}_0 : X \sim \mathcal{N}(600, 100^2)$
- $\mathcal{H}_1 : X \sim \mathcal{N}(650, 100^2)$

Les risques

On décide d'accepter de se tromper en croyant que le niveau de pluie a augmenté alors que dans les faits non dans 5 % des cas.

$$\alpha = 0,05$$

β sera subi.

Exemple II

Statistique de test

- On veut tester la moyenne μ.
- μ est estimée par $\bar{X_n} = \frac{1}{n} = \sum_{i=1}^{9} X_i$
- D'après le TCL, si $\mathcal{H}_0$ est vraie, alors

$$\bar{X_n} \sim \mathcal{N}(600, \frac{100^2}{9})$$

Exemple III

Mise en oeuvre du test

- $\bar{x_n} = 610,2$
- Cette observation est-elle probable sous $\mathcal{H}_0$?
-

$$\mathbb{P}(\text{Rejet } \mathcal{H}_0 | \mathcal{H}_0) < \alpha$$
$$\mathbb{P}(\bar{X_n} > \bar{x_n}) < \alpha$$

- $Z = \frac{\bar{X_n} - 600}{100/3} \sim \mathcal{N}(0,1)$

Exemple IV

Calcul de la p-value

$$\begin{aligned}\mathbb{P}(\bar{X}_n > 610,2) &= \mathbb{P}(Z > \frac{610.2 - 600}{100/3}) \\ &\simeq \mathbb{P}(Z > 0.306) \\ &\simeq 1 - \mathbb{P}(Z < 0.306) \\ &\simeq 1 - 0.62\ldots \\ &\simeq 0.38\end{aligned}$$

Décision

- $0.38 > \alpha$
- Il n'est pas raisonnable de rejeter $\mathcal{H}_0$

Exemple V

Erreur β de seconde espèce

- Quel est le risque d'accepter $\mathcal{H}_0$ alors que $\mathcal{H}_1$ est vraie ?
- Pour $\alpha = 0,05$, la valeur seuil k_α du rejet de $\mathcal{H}_0$ est :

$$\mathbb{P}(\bar{X_n} > k_\alpha) = \alpha$$
$$\mathbb{P}(Z > u_{1-\alpha}) = \alpha$$

avec $u_{1-\alpha}$ le quantile de la loi normale $\mathcal{N}(0,1)$ et $k_\alpha = u_{1-\alpha} * \sigma/\sqrt{(n)} + \mu$.

- Dans notre cas, nous avons $k_\alpha \simeq 655$
- $\beta = \mathbb{P}(\bar{(}X_n) < k_\alpha | \mathcal{H}_1 \simeq 0.56$
- β est trés élevé, donc la probabilité de garder $\mathcal{H}_0$ alors que $\mathcal{H}_1$ est vraie est également très élevée.
- La puissance est de l'ordre de 0.44, il faut augmenter n !

Les différents tests possibles

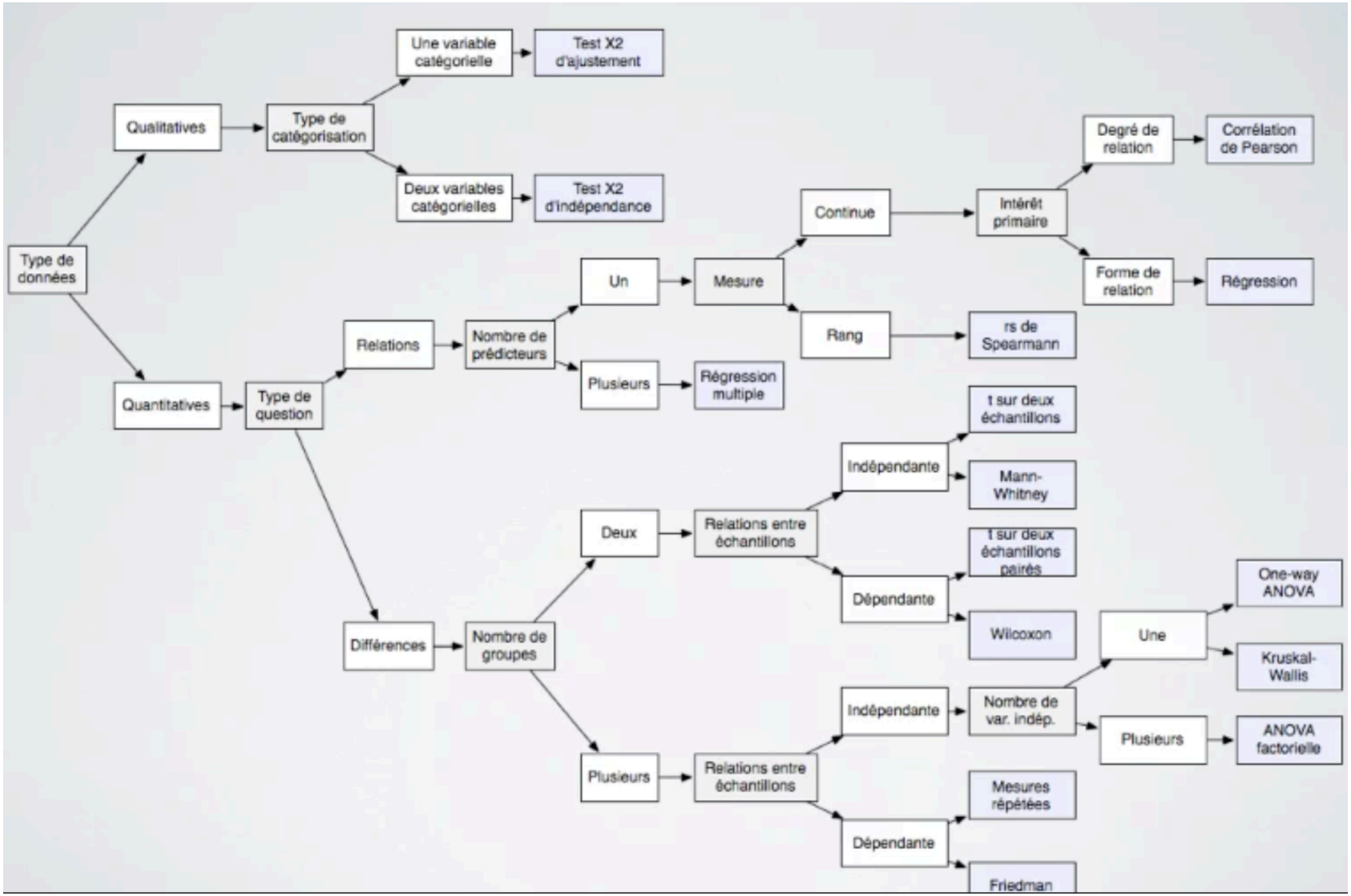
Classification

- Conformité
- Comparaison
- Ajustement
- Indépendance

Paramètres de la loi connus ?

- Tests paramétriques
- Tests non-paramètriques

Une grande famille de tests



https ://lepcam.fr/wp-content/uploads/2016/04/Choix-test-statistique.pdf

Conclusion

Tests statistiques

- Aide à prendre une décision
- Controle (imparfait) de l'erreur
- À utiliser avec prudence

Ce qu'il nous reste à voir

- Mise en oeuvre de tests particuliers
- Quels tests dans quelles conditions
- Test du χ^2
- Test de Student

Comparaisons de variables qualitatives : le test du χ^2

Plan

L'exemple du maïs

	Violet	Jaune
Lisse	135	44
Rugueux	47	14

Peut on estimer raisonnablement que la proportion est de 9/3/3/1 ?

source

L'exemple des 3 jurys

Lors d'un examen national, les résultats des trois jurys ont été les suivants :

	reçu	refusé
Jury 1	50	5
Jury 2	47	14
Jury 3	56	8

Peut on estimer raisonnablement que les trois jurys ont eu le même comportement ou se sont-ils comportés de manière différente ?

Attention, si l'on estime qu'ils se sont comportés de manière différente, il va falloir refaire l'examen, car il aura été injuste.

Une table de contingence, deux questions possibles I

Test d'adéquation

	Violet	Jaune
Lisse	135	44
Rugueux	47	14

- la proportion de grains observée est elle conforme à la distribution théorique ?
- Les variables qualitatives "couleur" et "texture" suivent elles la distribution stipulée a priori ?

Une table de contingence, deux questions possibles II

Test d'indépendance

	reçu	refusé
Jury 1	50	5
Jury 2	47	14
Jury 3	56	8

- les jury sont-ils analogues ?
- Les variables qualitatives "jury" et "reçu/refusé" sont elles indépendantes ?

Test du χ^2

Les deux tests suivent la même méthodologie et ne différent que par des détails

Les trois étapes d'un test : hypothèses, modèle et décision I

La question : les jury sont-ils analogues ?

Hypothèses

- Choix de l'état de référence $\mathcal{H}_0$

$$\begin{cases} \mathcal{H}_0 : \text{les jurys sont identiques} \\ \mathcal{H}_1 : \text{les jurys sont différents} \end{cases}$$

Modèle

Si les jurys sont identiques alors les variables sont indépendantes

$$\mathbb{P}(\text{reçu par le jury 1}) = \mathbb{P}(\text{reçu})\mathbb{P}(\text{passer par le jury 1})$$

Les trois étapes d'un test : hypothèses, modèle et décision II

Décision

- Calcul de la $p - valeur$:

  Si l'on admet l'équivalence des jurys, quelle est la probabilité d'observer un tableau encore plus "différent" ?

- Prise de décision :

  Si cette probabilité est faible (typiquement inférieure à $0,05$), on rejette l'hypothèse $\mathcal{H}_0$

Le tableau de référence I

Calcul d'un tableau de référence

- Quelles seraient les valeurs du tableau sous hypothèse d'indépendance
- Calcul d'un tableau **théorique**

Calcul des effectifs marginaux

Tableau observé :

	Reçu	Refusé	
Jury 1	50	5	55
Jury 2	47	14	61
Jury 3	56	8	64
	153	27	180

Le tableau de référence II

Effectifs théoriques

Tableau théorique :

	reçu	refusé	
Jury 1	46,75	8,25	30,56%
Jury 2	51,85	9,15	33,89 %
Jury 3	54,40	9,60	35,56 %
	85 %	15 %	180

$$\frac{N_{\bullet j}}{n} = \widehat{\mathbb{P}}_{\bullet j} \to \frac{153}{180} = 0,85 \qquad \underbrace{0,85 \times 0,30}_{\widehat{\mathbb{P}}_{ij} = \widehat{\mathbb{P}}_{i\bullet}\widehat{\mathbb{P}}_{\bullet j}} \times 180 = \frac{153 \times 55}{180} = 46,75$$

Comment mesurer l'écart entre le tableau observé et le tableau théorique ?

Comment mesurer l'écart les deux tableaux ?

Définition : Distance du χ^2

Soit O un tableau de contingence de I lignes et J colonnes d'effectif total n. Soit T un tableau de probabilité de même dimension. On appelle distance du χ^2 entre les tableaux O et T la quantité :

$$D(O,T) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - T_{ij})^2}{T_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(N_{ij} - n\,\widehat{\mathbb{P}}_{ij}\right)^2}{n\,\widehat{\mathbb{P}}_{ij}}$$

avec $O_{ij} = N_{ij}$ les effectifs observés, $T_{ij} = n\,\widehat{\mathbb{P}}_{ij}$ les effectifs théoriques sous l'hypothèse d'indépendance, n l'effectif total et $\widehat{\mathbb{P}}_{ij}$ la probabilité estimée sous hypothèse d'indépendance.

Pour notre exemple, on obtient $D(O,T) = 4,84$.

Distance du χ^2 est-elle importante ?

$$D(O,T) = 4,84$$

- Est-ce une différence importante ?
- p-valeur $= \mathbb{P}\big(D(O,T) \geq 4,84\big)$

Loi du χ^2

Sous l'hypothèse $\mathcal{H}_0$, D est distribué suivant une loi du χ^2 à $(3-1)(2-1) = 2$ degrés de libertés :

$$p\text{-valeur} = \mathbb{P}(D \geq 4,84) = 0,0887$$

- On estime à $8,9\%$ la probabilité d'observer un tableau encore plus différent que celui que l'on a.
- On conclut que la distance n'est pas très grande et que l'on ne peut pas rejeter l'hypothèse d'indépendance des jurys.

Plan

La loi du χ^2 I

Soit $Y \sim \mathcal{N}(0,1)$ une variable aléatoire normale centrée réduite.
Soit $Y_1, Y_2, \ldots, Y_n$ un échantillon de n réalisations i.i.d. de cette variable aléatoire.

Définition : La loi du χ^2

On appelle loi du χ^2 à n degrés de libertés la loi de la variable aléatoire Z_n :

$$Z_n = \sum_{i=1}^{n} Y_i^2$$

La loi du χ^2 II

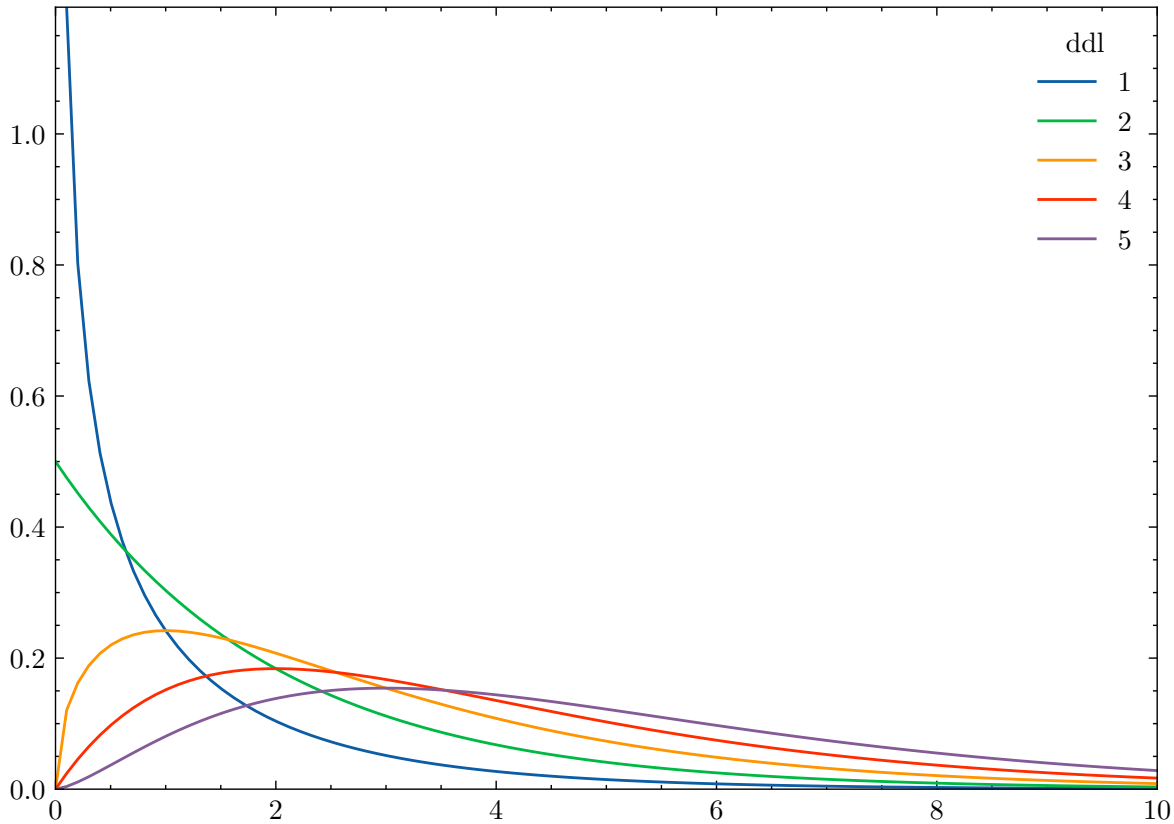


Figure – Exemples de loi du χ^2 pour 1 (bleu), 2 (rouge), 3 (vert), 4 (violet) et 5 (bleu ciel) degrés de liberté

Propriétés et approximation

$$\mathbb{E}(Z_n) = n \qquad \text{Var}(Z_n) = 2n \qquad \text{mode}(Z_n) = n - 2$$

En effet :

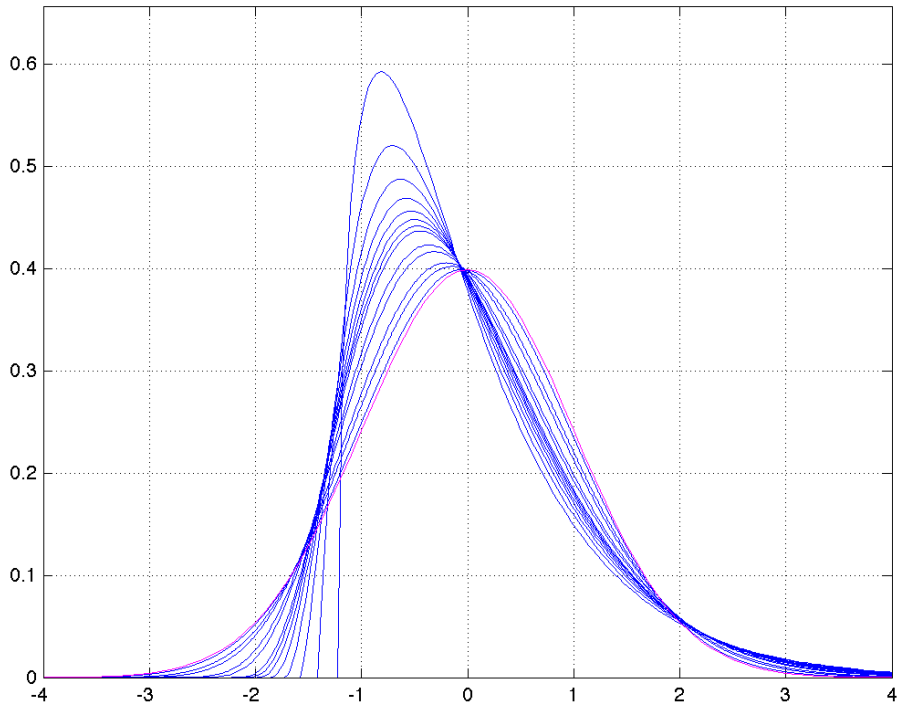
$$\mathbb{E}(Z_n) = \mathbb{E}(\sum_{i=1}^{n} Y_i^2) = n\mathbb{E}(Y^2) = nVar(Y) = n$$

Lorsque le nombre de degrés de libertés est important on peut utiliser une approximation asymptotique de la loi du χ^2. Les plus utilisées sont les approximations de Paul Levy et de Fisher :

Levy : $\frac{Z_n - n}{\sqrt{2n}} \sim \mathcal{N}(0,1)$ Fisher : $\sqrt{2Z_n} - \sqrt{2n-1} \sim \mathcal{N}(0,1)$

Il existe aussi une loi du χ^2 dite décentrée. C'est la somme de carrés de carrés d'une variable gaussiènne non centrée. Nous ne l'utiliserons pas dans ce cours.

Vitesse de convergence de la loi du χ_2



$$Z_n \sim \chi^2_n \qquad\qquad W_n = \frac{Z_n - n}{\sqrt{2n}}$$

La convergence est lente. L'approximation normale peut être utilisée pour $n > 30$.

Le cas de la moyenne I

La moyenne empirique commme v.a.

Considérons $X_1, X_2, ..., X_n$ un échantillon de n réalisations i.i.d. d'une variable aléatoire normale $\mathcal{N}(\mu, \sigma^2)$ d'espérance μ et de variance σ^2.
On a alors $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

Le cas de la moyenne II

Espérance et Variance de $\overline{X}$

- Espérance

$$\begin{aligned} \mathbb{E}(\overline{X}) &= \mathbb{E}(\tfrac{1}{n}\sum_{i=1}^{n} X_i) \\ &= \tfrac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) \\ &= \tfrac{1}{n}\sum_{i=1}^{n} \mu \qquad\qquad = \mu \end{aligned}$$

- Variance

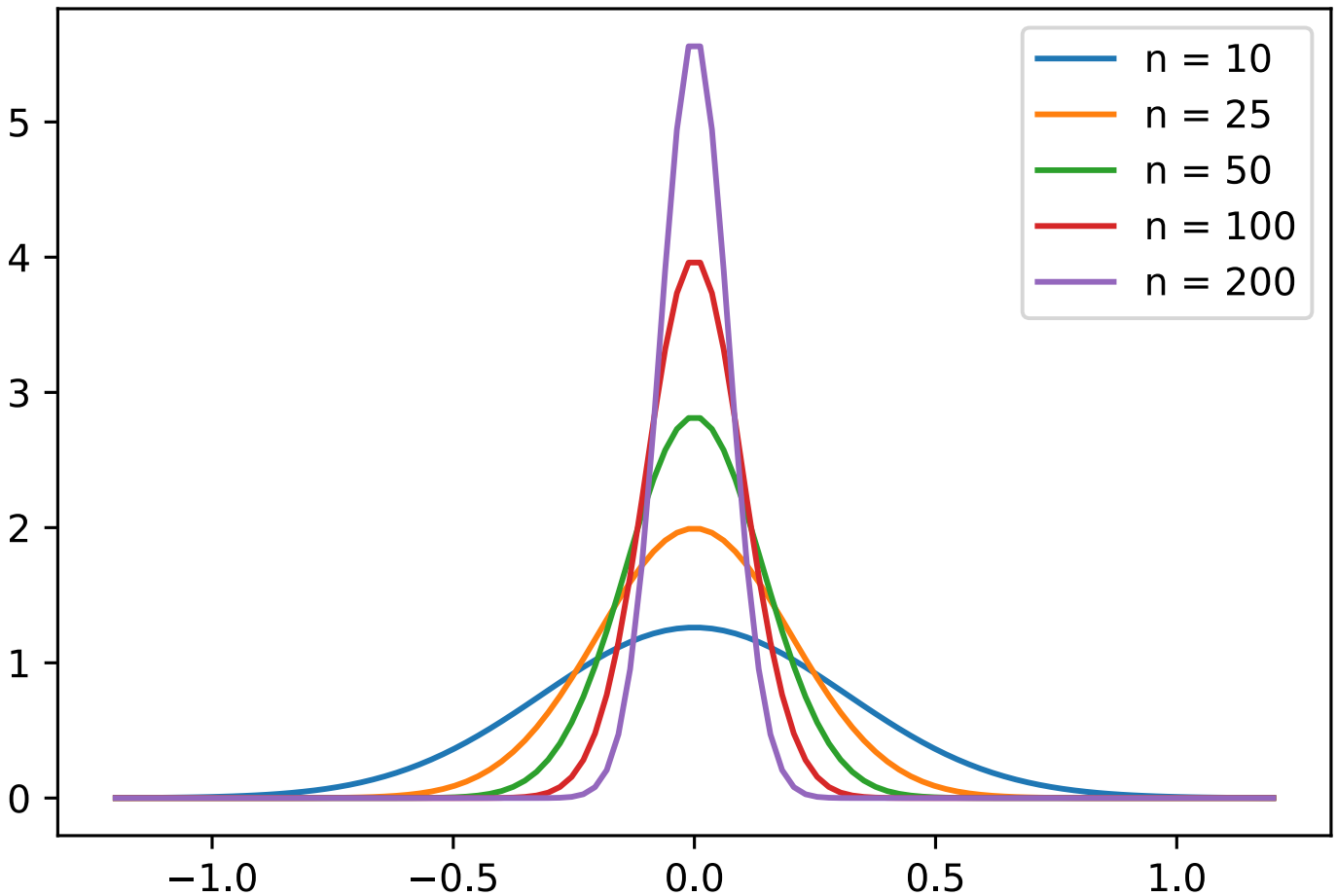
$$\begin{aligned} V(\overline{X}) &= V(\tfrac{1}{n}\sum_{i=1}^{n} X_i) \\ &= \tfrac{1}{n^2}\sum_{i=1}^{n} V(X_i) \\ &= \tfrac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \qquad\qquad = \frac{\sigma^2}{n} \end{aligned}$$

Le cas de la moyenne III

Loi de $\overline{X}$

- $\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
- Version centré/réduit $\sqrt{n}\frac{(\overline{X}-\mu)}{\sigma} \sim \mathcal{N}(0,1)$
- Version χ^2 : $n\frac{(\overline{X}-\mu)^2}{\sigma^2} \sim \chi^2_1$
- La moyenne se concentre autour de l'espérance

Le cas de la moyenne IV



Vers la moyenne I

Loi normale centrée réduite et χ^2

Soit $X_1, X_2, ..., X_n$ un échantillon de n réalisations i.i.d. d'une variable aléatoire normale $\mathcal{N}(\mu, \sigma^2)$ d'espérance μ et de variance σ^2.

Nous avons alors :

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2_n$$

- Étant donné que $Y_i = \frac{X_i - \mu}{\sigma}$ suit une loi normale centrée réduite.

Vers la moyenne II

Remplaçons μ par son estimation

Par contre, il est moins intuitif de montrer que :

$$\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

- Lorsque l'on remplace le paramètre μ par son estimation $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ on perd un degré de liberté.
- On a la décomposition suivante :

$$\underbrace{\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}}_{\chi^2_n} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} + \underbrace{n\frac{(\overline{X} - \mu)^2}{\sigma^2}}_{\chi^2_1}$$

- Ce qui permet de conclure en invoquant le théorème de Cochran sur l'additivité des degrés de liberté.
- Wikipedia

Théorème du χ^2 I

Théorème : Théorème du χ^2 (Pearson)

$$X_i = \frac{N_i - n\,\widehat{\mathbb{P}}_i}{\sqrt{n\,\widehat{\mathbb{P}}_i}} \qquad \sum_{i=1}^{I} X_i^2 \longrightarrow \chi^2_{I-1}$$

$$X_{ij} = \frac{N_{ij} - n\,\widehat{\mathbb{P}}_{ij}}{\sqrt{n\,\widehat{\mathbb{P}}_{ij}}} \qquad \sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}^2 \longrightarrow \chi^2_{(I-1)(J-1)}$$

Théorème du χ^2 II

Eléments de preuve dans le cas d'une variable à $I = 2$ modalités.

dans ce cas on a $n = N_1 + N_2$ et $p_1 + p_2 = 1$ et $N_1 \sim \mathcal{B}(n, p_1)$. Pour des échantillons assez grand on peut accepter une estimation gaussienne

$$\frac{N_1 - np_1}{\sqrt{np_1(1-p_1)}} \sim \mathcal{N}(0,1) \qquad \Longrightarrow \qquad \frac{(N_1 - np_1)^2}{np_1(1-p_1)} \sim \chi^2_1$$

et

$$\begin{aligned}
\tfrac{(N_1-np_1)^2}{np_1(1-p_1)} &= \tfrac{(N_1-np_1)^2}{np_1(1-p_1)}(1-p_1+p_1) \\
&= \tfrac{(N_1-np_1)^2}{np_1} + \tfrac{(N_1-np_1)^2}{n(1-p_1)} \\
&= \tfrac{(N_1-np_1)^2}{np_1} + \tfrac{(N_1-np_1-n+n)^2}{n(1-p_1)} \\
&= \tfrac{(N_1-np_1)^2}{np_1} + \tfrac{(N_2-np_2)^2}{n(1-p_1)} \qquad = \sum_{i=1}^{2} \frac{(N_i - np_i)^2}{np_i}
\end{aligned}$$

□

Mise en œuvre du test du χ^2

1. On construit un tableau de contingence O des observations (2 variables qualitatives de respectivement I et J modalités)
2. On calcule les marginales $p_i = \frac{1}{n}\sum_{j=1}^{J} O_{ij}$
3. On calcule pour chaque case du tableau des effectifs théoriques $T_{ij} = np_ip_j$ (en supposant l'indépendance)
4. On calcule la distance du χ^2

$$D(O,T) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$

5. On calcule le nombre de degrés de liberté du χ^2 : $d = (I-1)(J-1)$
6. On regarde dans les tables d'une variable aléatoire Z distribuée suivant une loi χ^2 à d degrés de liberté la p-valeur de $D(O,T)$

$$\text{p-valeur} = \mathbb{P}\big(Z \geq D(O,T)\big)$$

7. On décide qu'on ne peut pas conclure à la dépendance si la p-valeur est supérieure à α

Conditions d'utilisation du test du χ^2

- Observations tirées au hasard
- Observations indépendantes
- n suffisamment grand
- des Effectifs > 5 pour chaque élément du tableau

Conclusion

- La question face à une table de contingence :
  - étant donné une distribution de probabilités ?
    - $\rightarrow$ l'exemple du maïs
  - ces deux variables qualitatives sont elles indépendantes ?
    - $\rightarrow$ l'exemple des jurys

- La réponse :
  - poser les hypothèses
  - calculer la distance du χ^2 entre la table observé et la table théorique
  - proposer une décision à partir de la p-valeur

- Un test adoubé par la pratique...
  - réputé robuste

- Attention à la décision
  - surtout dans le cas de tests multiples

Repéres bibliographiques

- Sur wikipedia
- Cours MIT
- Un autre cours

Qualitatif vs Quantitatif : Le Test de student

L'exemple de l'effet d'un médicament I

Patient	Groupe	Pression sanguine
t1	traitement	88
t2	traitement	83
t3	traitement	82
t4	traitement	101
t5	traitement	99
t6	traitement	85
t7	traitement	87
t8	traitement	89
t9	traitement	88
p10	placebo	88
p11	placebo	82
p12	placebo	101
p13	placebo	106
p14	placebo	96
p15	placebo	92
p16	placebo	112
p17	placebo	97
	qualitative	quantitative

L'exemple de l'effet d'un médicament II

Question

Le traitement fait-il diminuer *significativement* la pression sanguine ?

Les hypothèses

$$\begin{cases} \mathcal{H}_0 : \text{le traitement est inefficace} \\ \mathcal{H}_1 : \text{le traitement la fait baisser} \end{cases}$$

Réponse

- Comparer les deux échantillons à travers la différence de leurs moyennes

$$\overline{x}_t - \overline{x}_p = 90,2 - 96,7 = -6,5$$

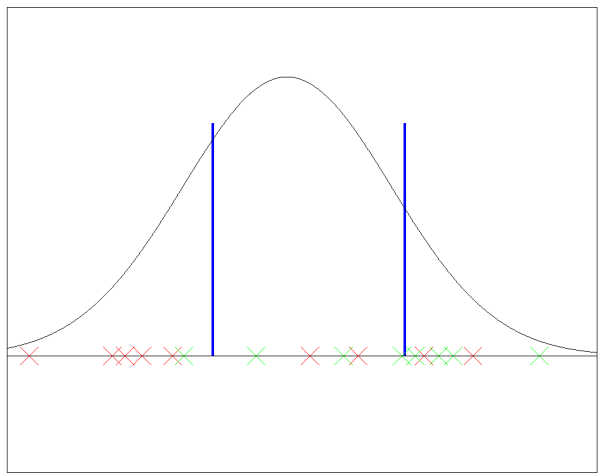
- Cette valeur de -6,5 peut elle s'expliquer par un hasard raisonnable ?

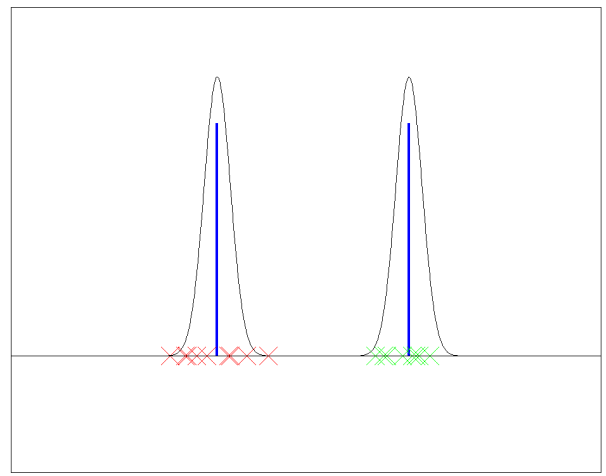
Un hasard raisonnable I

$\overline{x}_t - \overline{x}_p = -6,5$ peut elle s'expliquer par un hasard raisonnable ?

$$\overline{x}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} x_{ti} \qquad \overline{x}_p = \frac{1}{n_p} \sum_{i=1}^{n_p} x_{pi}$$

Influence de la variance





Prendre en compte la variance : le modèle I

Cas de variance connue et similaire

Trois hypothèses :

1. L'hypothèse gaussienne :
   - Mesure des patients avec traitement : $X_t \sim \mathcal{N}\left(\mu_t, \sigma^2\right)$
   - Mesure des patients sous placebo : $X_p \sim \mathcal{N}\left(\mu_p, \sigma^2\right)$
2. Même variance connue : $\sigma_t^2 = \sigma_p^2 = \sigma^2 = 60$

Les hypothèses du test

$$\begin{cases} \mathcal{H}_0 : \text{inefficace} & \mu_t = \mu_p \\ \mathcal{H}_1 : \text{la pression baisse} & \mu_t < \mu_p \end{cases}$$

Prendre en compte la variance : le modèle II

Les moyennes comme variable aléatoire

- Moyenne avec traitement : $\overline{X}_t \sim \mathcal{N}\left(\mu_t, \frac{\sigma^2}{n_t}\right)$
- Moyenne sous placebo : $\overline{X}_p \sim \mathcal{N}\left(\mu_p, \frac{\sigma^2}{n_p}\right)$

$$\begin{aligned} \mathbb{E}(\overline{X}) &= \mathbb{E}(\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbb{E}(X_i) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \mu \qquad = \mu \end{aligned} \qquad \begin{aligned} V(\overline{X}) &= V(\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i) \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^n V(X_i) \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^n \sigma^2 \qquad = \frac{\sigma^2}{n} \end{aligned}$$

Différence des moyennes

$$\overline{X}_t - \overline{X}_p \sim \mathcal{N}\left(\mu_t - \mu_p, \sigma^2(\frac{1}{n_t} + \frac{1}{n_p})\right)$$

Test avec variance connue I

Modèle

$$\overline{X}_t - \overline{X}_p \sim \mathcal{N}\left(\mu_t - \mu_p, \sigma^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)\right)$$

Hypothèses

$$\begin{cases} \mathcal{H}_0 : \text{le traitement n'a pas d'effet} & \mu_t - \mu_p = 0 \\ \mathcal{H}_1 : \text{le traitement est efficace} & \mu_t - \mu_p < 0 \end{cases}$$

- sous $\mathcal{H}_0$

$$U = \frac{\overline{X}_t - \overline{X}_p}{\sqrt{\sigma^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}} \sim \mathcal{N}(0,1)$$

Test avec variance connue II

Statistique u

$$u = \frac{90,2 - 96,7}{\sqrt{60\left(\frac{1}{9} + \frac{1}{8}\right)}} = -1.73$$

Décision

- $$p = \mathbb{P}(U \leq -1.7343) = 0,041$$
- $p < \alpha \Rightarrow \mathcal{H}_0$ est rejetée, il est raisonnable d'admettre que le traitement a un effet.

Récapitulons : le test de comparaison des moyennes

1. **La question** : les deux groupes sont ils des réalisations de la même loi ?
2. **Le modèle** : gaussien
3. **Les hypothèses** : même variance σ^2 connue
4. Calcul de

$$u = \frac{\overline{x}_t - \overline{x}_p}{\sqrt{\sigma^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}}$$

$\overline{x}_t$ moyenne avec traitement
$\overline{x}_p$ moyenne sans traitement
n_t nombre de cas avec traitement
n_p nombre de cas sans traitement

5. Calcul de la p-valeur $U \sim \mathcal{N}(0, 1)$

$$\texttt{p-valeur} = \mathbb{P}(U \leq u)$$

6. On décide qu'on ne peut pas conclure à l'efficacité du traitement si la `p-valeur` est supérieur à α

Cas de la variance inconnue I

Méthode

- On remplace σ^2 par son estimateur $\widehat{\sigma}^2$
- La nouvelle v.a. suit une loi et Student à $n_t + n_p - 2$ degrés de liberté

$$T_{n_t+n_p-2} = \frac{\overline{X}_t - \overline{X}_p}{\sqrt{\widehat{\sigma}^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}} \sim \mathcal{T}_{n_t+n_p-2}$$

avec $\widehat{\sigma}^2 = \frac{1}{n_t+n_p-2}\left(\sum_{i=1}^{n_t}(X_{ti} - \overline{X}_t)^2 + \sum_{i=1}^{n_p}(X_{pi} - \overline{X}_p)^2\right)$.

Cas de la variance inconnue II

Application

- Calcul de la statistique

$$t = \frac{90,2 - 96,7}{\sqrt{63,4\left(\frac{1}{9} + \frac{1}{8}\right)}} = -1.68$$

- Calcul de la p-value

  $\texttt{p-valeur} = \mathbb{P}(T_{n_t+n_p-2} \leq -1.68) = 0,056$

- Version python

```
from scipy.stats import t
t.cdf(-1.68,8+9-2)
```

La loi de Student : définition

- Soit $N \sim \mathcal{N}(0,1)$ une v. a. normale centrée réduite.
- Soit X_n la v. a. distribuée suivant une loi du χ^2 à n ddl
  - Par exemple $X_n = \sum_{i=1}^{n} N_i^2$
- supposons que N et X_n soient linéairement indépendantes

Définition : La loi de student

On appelle loi de student à n degrés de libertés la loi de la variable aléatoire T_n

$$T_n = \frac{N}{\sqrt{\frac{X_n}{n}}} \qquad \text{avec} \qquad N \sim \mathcal{N}(0,1) \qquad X_n \sim \chi^2_n$$

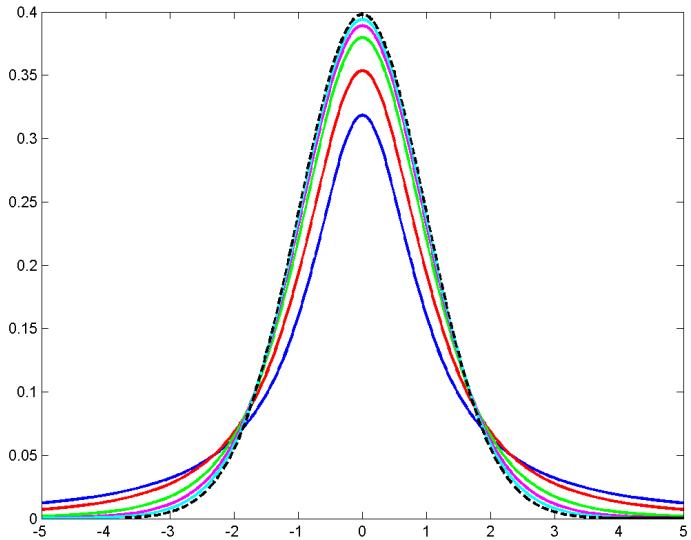
Propriétés et approximation

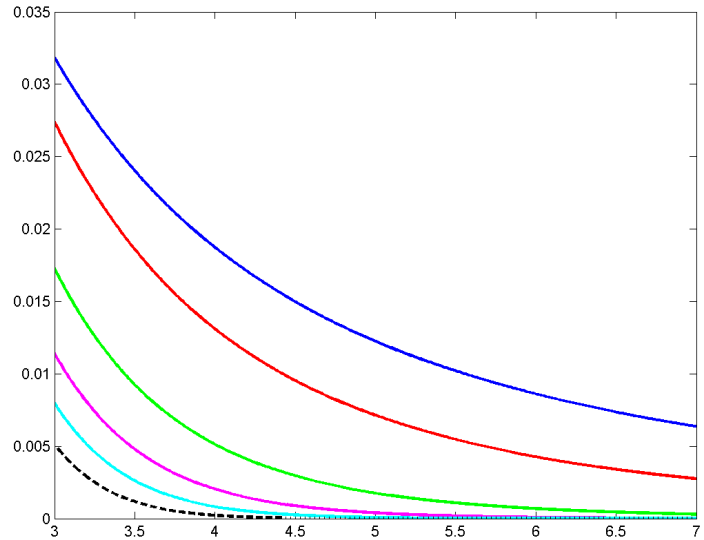
- Publiée pour la première fois en 1908 par William Sealy Gosset qui travaillait chez Guinness (la brasserie de Dublin). Pour des raisons commerciales, il a du utiliser le pseudonyme de Student, qui restera attaché à cette loi.
- Tend vers une loi normale pour $n > 30$

$$T_n \xrightarrow[n \to +\infty]{} \mathcal{N}(0,1)$$

$N \sim \mathcal{N}(0,1)$	$\mathbb{P}(N > 2) = 0,023$	`norm.sf(2,0,1)`
$T \sim \mathcal{T}_1$	$\mathbb{P}(T > 2) = 0,148$	`t.sf(2,1)`
$T \sim \mathcal{T}_2$	$\mathbb{P}(T > 2) = 0,092$	`t.sf(2,2)`
$T \sim \mathcal{T}_{10}$	$\mathbb{P}(T > 2) = 0,038$	`t.sf(2,10)`

La loi de Student : $T_n = \frac{N}{\sqrt{\frac{X_n}{n}}}$





ddl : 1 (bleu), 2 (rouge), 5 (vert), 10 (violet) et 20 (cyan). En noir, une loi normale .

Résumé

$$U \sim \mathcal{N}(0, \sigma^2) \qquad N = \frac{U}{\sigma} \sim \mathcal{N}(0,1) \qquad T = \frac{N}{\widehat{\sigma}} = \frac{N}{\sqrt{\frac{N_1^2 + N_2^2}{2}}} \sim \mathcal{T}_2$$

Application à une moyenne d'un échantillon gaussien I

La moyenne centrée réduite comme v.a.

- Soit $X \sim \mathcal{N}(\mu, \sigma^2)$, avec σ^2 inconnue.
- $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$
- Version centrée réduite :

$$Y = \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0, 1)$$

- On a également :

$$Z_{n-1} = \sum_{i=1}^{n} \left(\frac{(X_i - \overline{X})}{\sigma} \right)^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

Application à une moyenne d'un échantillon gaussien II

$$T_{n-1} = \frac{\overline{X} - \mu}{\frac{\widehat{\sigma}}{\sqrt{n}}} = \frac{\overline{X} - \mu}{\sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}}} = \frac{\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\sigma^2}}{n-1}}} = \frac{Y}{\sqrt{\frac{Z_{n-1}}{n-1}}}$$

avec $\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.
T_{n-1} suit donc une loi de Student à $n-1$ dégrés de libertés.

Ce résultat peut être utilisé pour estimer un encadrement de l'espérance μ donné α :

$$\overline{X} - t_{1-\alpha/2}\frac{S_{n-1}}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{1-\alpha/2}\frac{S_{n-1}}{\sqrt{n}}$$

Le test de Student (t-test) : deux échantillons Gaussiens

Soit $X \sim \mathcal{N}(\mu_x, \sigma^2)$ et $Y \sim \mathcal{N}(\mu_y, \sigma^2)$ deux lois de même variance inconnue et leurs échantillons respectifs $X_1, ..., X_{n_x}$ et $Y_1, ..., Y_{n_y}$.
Les variables $\overline{X} = \frac{1}{n}\sum_{i=1}^{n_x} X_i$ et $S_x^2 = \sum_{i=1}^{n_x}(X_i - \overline{X})^2$ sont caractérisées par les lois :

$$\overline{X} \sim \mathcal{N}\Big(\mu_x, \frac{\sigma^2}{n_x}\Big) \quad \overline{Y} \sim \mathcal{N}\big(\mu_y, \frac{\sigma^2}{n_y}\big)$$

$$\frac{S_x^2}{\sigma^2} \sim \chi^2_{n_x-1} \; ; \quad \frac{S_y^2}{\sigma^2} \sim \chi^2_{n_y-1}$$

Et nous avons donc

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_x - \mu_y, (\frac{1}{n_x} + \frac{1}{n_y})\sigma^2\right) \text{ et } \frac{S_x^2}{\sigma^2} + \frac{S_y^2}{\sigma^2} \sim \chi^2_{n_x+n_y-2}$$

Le test de Student (t-test)

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_x - \mu_y, (\frac{1}{n_x} + \frac{1}{n_y})\sigma^2\right) ; \frac{S_x^2}{\sigma^2} + \frac{S_y^2}{\sigma^2} \sim \chi^2_{n_x+n_y-2}$$

On définit alors la variable de Student :

$$T_{n_x+n_y-2} = \sqrt{n_x + n_y - 2}\, \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right) S_{xy}^2}}$$

avec $S_{xy}^2 = S_x^2 + S_y^2 = \sum_{i=1}^{n_x}(X_i - \overline{X})^2 + \sum_{i=1}^{n_y}(Y_i - \overline{Y})^2$

Si l'on fait l'hypothèse que $\mu_x = \mu_y$

$$T = \sqrt{n_x + n_y - 2}\, \frac{\overline{X} - \overline{Y}}{\sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right) S_{xy}^2}}$$

suit une loi de Student à $n_x + n_y - 2$ degrés de liberté.

Le test de Student (t-test) I

Les deux échantillons

- $X_{t1}, ..., X_{ti}, ..., X_{tn_t}$
- $X_{p1}, ..., X_{pi}, ..., X_{pn_p}$

Les deux hypothèses

1. l'hypothèse Gaussienne :
   - $X_{ti} \sim \mathcal{N}(\mu_t, \sigma^2)$
   - $X_{pi} \sim \mathcal{N}(\mu_p, \sigma^2)$
2. Même variance inconnue : $\sigma_t^2 = \sigma_p^2 = \sigma^2$

La question

Les deux échantillons que nous observons sont-ils des réalisations d'une même variable aléatoire ?

Le test de Student (t-test) II

Les hypothèses

$$\begin{cases} \mathcal{H}_0 : \text{échantillons de même loi} & \mu_t = \mu_p \\ \mathcal{H}_1 : \text{de lois différentes} & \mu_t > \mu_p \end{cases}$$

La statistique

$$T = \frac{\overline{X}_t - \overline{X}_p}{\sqrt{\widehat{\sigma}^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}} \sim \mathcal{T}_{n_t+n_p-2}$$

avec $\widehat{\sigma}^2 = \frac{1}{n_t+n_p-2}\left(\sum_{i=1}^{n_t}(X_{ti} - \overline{X}_t)^2 + \sum_{i=1}^{n_p}(X_{pi} - \overline{X}_p)^2\right)$

Mise en œuvre du test de student

1. Calcul de
$$\overline{x}_t = \frac{1}{n_t}\sum_{i=1}^{n_t} x_{ti} \quad \text{moyenne avec traitement}$$
$$\overline{x}_p = \frac{1}{n_p}\sum_{i=1}^{n_p} x_{pi} \quad \text{moyenne sans traitement}$$
2. Calcul de $\widehat{\sigma}^2 = \frac{1}{n_t+n_p-2}\Big(\sum_{i=1}^{n_t}(x_{ti}-\overline{x}_t)^2 + \sum_{i=1}^{n_p}(x_{pi}-\overline{x}_p)^2\Big)$
3. Calcul de
$$t = \frac{\overline{x}_t - \overline{x}_p}{\sqrt{\widehat{\sigma}^2\big(\frac{1}{n_t}+\frac{1}{n_p}\big)}}$$
n_t nombre de cas avec traitement
n_p nombre de cas sans traitement
4. Calcul du nombre de degrés de liberté $d = n_t + n_p - 2$
5. Calcul de la p-valeur $T \sim \mathcal{T}_d$ (ou lecture sur les tables)
$$pval = \mathbb{P}(T \geqslant t)$$
6. on décide qu'on ne peut pas conclure à l'efficacité du traitement si la $pval \geqslant \alpha$

Exemple de mise en œuvre du test de student I

Les données

groupe avec traitement (t)	30.02	29.99	30.11	29.97	30.01	29.99
groupe sans traitement (p)	29.89	29.93	29.72	29.98	30.02	29.98

Question : le traitement augmente-t-il la mesure ?

1. $\overline{x}_t = 30.015, \quad \overline{x}_p = 29.92 \qquad \overline{x}_t - \overline{x}_p = 0.095$
2. $\widehat{\sigma}^2 = \frac{1}{10}\left(\sum_{i=1}^{6}(x_{ti} - 30.015)^2 + \sum_{i=1}^{6}(x_{pi} - 29.92)^2\right) \approx 0.0071$
3. $$t = \frac{\overline{x}_t - \overline{x}_p}{\sqrt{\widehat{\sigma}^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}} \approx \frac{0.095}{\sqrt{0.0071\left(\frac{1}{6} + \frac{1}{6}\right)}} = 1.959$$
4. calcul du nombre de degrés de liberté $d = n_t + n_p - 2 = 10$
5. calcul de la p-valeur $T \sim \mathcal{T}_d$ (ou lecture sur les tables)

$$pval = \mathbb{P}(T \geq 1.959) = \texttt{t.sf(1.959,10)} = 0.0393$$

6. on décide qu'on peut conclure à l'efficacité du traitement car la p-valeur est inférieure à 0,05.

L'exemple de la relation entre oxygène dissout et pression I

Patient	O_2	Pression sanguine
p1	0,31	88
p2	0,30	83
p3	0,29	82
p4	0,35	101
p5	0,33	99
p6	0,31	85
p7	0,30	87
p8	0,34	89
p9	0,32	88
p10	0,28	88
p11	0,30	82
p12	0,33	101
p13	0,31	106
p14	0,32	96
p15	0,30	92
p16	0,35	112
p17	0,31	97
	quantitative	quantitative

Question : Il y a t'il une relation entre ces deux variables ?

L'exemple de la relation entre oxygène dissout et pression II

Les hypothèses

$$\begin{cases} \mathcal{H}_0 : \text{indépendance} \\ \mathcal{H}_1 : \text{dépendance} \end{cases}$$

Stratégie de réponse

- Tester la pente a de la droite de régression

$$pression = aO_2 + b + \varepsilon$$

- les hypothèses : $\begin{cases} \mathcal{H}_0 : a = 0 \\ \mathcal{H}_1 : a \neq 0 \end{cases}$
- La régression donne $\hat{a} = 0,12$

Cette valeur peut elle s'expliquer par un hasard raisonnable ?

Un hasard raisonnable...

1. Supposons qu'il y a indépendance $\Leftrightarrow a = 0$
2. Générons plein $(m = 1000, 1000000, +\infty)$ d'échantillons

$$(x_i, y_{ij} = ax_i + b + \varepsilon_{ij}), \quad i = 1, n \;\; j = 1, m$$

3. Pour chacun de ces échantillon calculons $\hat{a}_j$
4. Regardons la probabilité $\mathbb{P}(|\hat{a}| > 0, 12)$
5. Si cette probabilité est trop petite, il n'est pas "raisonnable" de considérer que l'hypothèse d'indépendance est exacte.

Comparaisons de deux variables quantitatives et régression

$$y_i = ax_i + b + \varepsilon_i$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$
indépendance des ε_i

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

$$\hat{a} \sim \mathcal{N}\left(a, \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)$$

$$\frac{\hat{a} - a}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \sim \mathcal{N}(0, 1)$$

$$\hat{\varepsilon}_i = y_i - (\hat{a}x_i + \hat{b})$$

$$\frac{\varepsilon_i}{\sigma} \sim \mathcal{N}(0, 1) \quad \Rightarrow \quad \frac{1}{\sigma^2} \sum_{i=1}^n \varepsilon_i^2 \sim \chi^2_n$$

$$\frac{1}{\sigma^2} \sum_{i=1}^n \hat{\varepsilon}_i^2 \sim \chi^2_{n-2}$$

Pente de la droite de régression et loi de student

$$\frac{\hat{a}-a}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i-\overline{x})^2}}} \sim \mathcal{N}(0,1) \qquad \frac{1}{\sigma^2}\sum_{i=1}^n \hat{\varepsilon}_i^2 \sim \chi^2_{n-2}$$

or $\frac{\mathcal{N}}{\sqrt{\frac{\chi^2_n}{n}}} \sim \mathcal{T}_n^2$ suit une loi de student à n degrés de libertés

$$\frac{\frac{\hat{a}-a}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i-\overline{x})^2}}}}{\sqrt{\frac{1}{\sigma^2(n-2)}\sum_{i=1}^n \hat{\varepsilon}_i^2}} \sim \mathcal{T}_{n-2} \qquad \Longrightarrow \qquad \frac{\hat{a}-a}{\sqrt{\frac{\widehat{\sigma}^2}{S_x^2}}} \sim \mathcal{T}_{n-2}$$

avec $\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \big(y_i - (\hat{a}x_i + \hat{b})\big)^2$ et $S_x^2 = \sum_{i=1}^n (x_i - \overline{x})^2$

Mise en œuvre du test sur la pente de la régression

1. Les hypothèses : $\begin{cases} \mathcal{H}_0 : \text{indépendance} & a = 0 \\ \mathcal{H}_1 : \text{dépendance} & a \neq 0 \end{cases}$
2. Calcul de
$$\hat{a} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
3. Calcul de $\begin{cases} \widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - (\hat{a}x_i + \hat{b})\right)^2 \\ S_x^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 \end{cases}$
4. Calcul de la statistique t
$$t = \frac{\hat{a}}{\sqrt{\frac{\widehat{\sigma}^2}{S_x^2}}}$$
5. Calcul du nombre de degrés de liberté $d = n - 2$
6. Calcul de la p-valeur $T \sim \mathcal{T}_d$ (test bilatéral)
$$pval = 2\mathbb{P}(T \geq |t|)$$
7. on décide qu'on ne peut pas conclure à l'efficacité du traitement si la p-valeur est supérieure à α

Conclusion

Plan

Conclusion

- La question
  - cette variable quantitative est elle indépendantes de cette variable qualitative ?
  - comparaison de deux échantillons quantitatifs

- Hyopthèses à vérifier
  - Distribution normale (par exemple un test du χ^2 adapté)
  - Égalité de variances (test de Fisher)

  sinon il faut faire un autre test comme celui de Wilcoxon ou de Mann et Whitney

- Plusieurs variations du test de student...
  - Un échantillon (test d'une valeur de l'espérance)
  - Deux échantillons appariés
  - Test de la pente de la régression simple

Repéres bibliographiques

- La page wikipedia
- Un autre cours en ligne
- WikiStat