

Statistique  
Compétence 2  
Ressource R4.04  
Traitement numérique des données

David Xu

IUT de Sceaux GEA 2

09 janvier 2024

Dans l'épisode précédent

# À quoi sert la statistique ?

- Analyse précise des données (intervalles de confiance).
- Comparaison de données (tests statistiques).

# Faire un intervalle de confiance

- Calculer une moyenne estimée  $\hat{m}$ .
- Calculer un écart-type estimé  $\hat{\sigma}$ .
- Lire dans les tables de la loi gaussienne et de la loi de Student.
- Construire un intervalle de confiance dans 3 cas :
  - Loi inconnue et  $n \geq 30$ .
  - Pourcentage (loi de Bernoulli) et  $n \geq 30$ .
  - Loi normale et  $n \leq 30$ .

# Tests statistiques

- Test de la moyenne.
- Test du  $\chi^2$ .

# Un exemple

Une usine produit des tablettes de chocolats qui doivent être de 100g en moyenne.

Les 100 dernières tablettes produites pèsent :

105	102	96	103	101
⋮	⋮	⋮	⋮	⋮
101	98	100	98	104

- Question : la machine est-elle dérégulée ?

# Test de la moyenne

Pour répondre scientifiquement, on fait un test de la moyenne.

- La valeur de référence :  $m_0 = 100$  (on veut des tablettes de 100g).
- Les poids des tablettes sont indépendants et de même loi.
- Le poids des tablettes suit une loi (inconnue) de moyenne  $m$  (inconnue) et d'écart-type  $\sigma$  (inconnu).
- La question que l'on se pose est :  $m = m_0$  ou  $m \neq m_0$  ?

On peut calculer  $\hat{m}$  et  $\hat{\sigma}$  comme dans l'épisode précédent.

# Flashback

- Moyenne estimée :

$$\hat{m} = \frac{X_1 + \dots + X_n}{n}$$

- Écart-type estimé :

$$\hat{\sigma} = \sqrt{\frac{(X_1 - \hat{m})^2 + \dots + (X_n - \hat{m})^2}{n - 1}}$$



## Idées ?

Idée d'une méthode pour décider si la machine est dérégulée ?

Exemple :  $\hat{m} = 100.8$ ,  $\hat{\sigma} = 2.3$  et  $n = 100$ . Au risque 5%, comment décider si la machine est dérégulée ?

WOOCLAP

## Idées

- 1 Idée 1 : si  $\hat{m} = m_0$  la machine est bien réglée ; si  $\hat{m} \neq m_0$  la machine est dérégulée. **Trop restrictif.**
- 2 Idée 2 : on se donne un intervalle  $\hat{m} = [95; 105]$ , la machine est bien réglée ; sinon non. **Comment choisir cet intervalle ?**
- 3 Idée 3 : on construit l'intervalle de confiance de  $m$  (inconnue) et si  $m_0$  est dedans, la machine est bien réglée. **Ça marche ! Mais ce ne sera pas généralisable à tous les types de tests.**
- 4 Idée 4 : le calcul d'une statistique et d'une zone de rejet.

# Vocabulaire

Lors d'un test statistique, on test une hypothèse  $H_0$  contre une autre hypothèse  $H_1$ .

Hypothèse $H_0$	Hypothèse $H_1$
Hypothèse nulle	Hypothèse alternative
Hypothèse non coûteuse	Hypothèse coûteuse
Si $H_0$ est "vraie", on ne fait rien	Si $H_1$ est vraie, il faut agir
Fonctionnement normal	Changement de situation
$H_0 : m = m_0$	$H_1 : m \neq m_0$ (ou $H_1 : m < m_0$ ou $H_1 : m > m_0$ )

# Déroulement d'un test statistique

On souhaite tester l'hypothèse  $H_0$  contre l'hypothèse  $H_1$ .

- 1 On se donne un risque d'erreur  $\alpha$  petit (10%, 5%, 1%, 0.01%).
- 2 On suppose que  $H_0$  est vraie.
- 3 À partir de nos observations  $X_1, \dots, X_n$ , on calcule une statistique  $\hat{z}$  dont la loi est connue sous l'hypothèse  $H_0$ .
- 4 Déterminer une zone de rejet  $Z_r$  à partir de
  - la loi de la statistique,
  - de  $H_1$  et
  - de  $\alpha$ .
- 5 On compare la valeur  $\hat{z}$  à  $Z_r$  :
  - Si  $\hat{z} \notin Z_r$ , alors **on ne rejette pas  $H_0$** .
  - Sinon, **on accepte  $H_1$** .

# Les erreurs

Lors d'un test, il peut y avoir 2 types d'erreurs :

On accepte $H_1$ , mais $H_0$ est vrai	On ne rejette pas $H_0$ , mais $H_0$ est fausse
Chiant à court terme, pas chiant à long terme	Invisible à court terme, dramatique à long terme
Erreur de 1ère espèce	Erreur de 2nde espèce
Probabilité $\alpha$ d'arriver	Probabilité $\beta$ d'arriver
Risque du test	Puissance du test
$\alpha$ contrôlé (donné par l'énoncé)	$\beta$ non calculable de manière générique

# Questions sur les tests statistiques ?

# Test de la moyenne

# Un exemple

Une usine produit des tablettes de chocolats de 100g en moyenne. Les 100 dernières tablettes produites pèsent :

105	102	96	103	101
⋮	⋮	⋮	⋮	⋮
101	98	100	98	104

- Question : la machine est-elle dérégulée ? On prend  $\alpha = 5\%$ .



# Formalisation des hypothèses

On a une **valeur de référence**  $m_0 = 100\text{g}$ .

On produit des plaquettes de chocolats de loi inconnue.

On va tester l'hypothèse nulle

- $H_0 : m = m_0$ , c'est-à-dire que la moyenne des plaques de chocolats est de 100g (on continue la production)

**versus**

- $H_1 : m \neq m_0$ , c'est-à-dire que la machine est défectueuse (on arrête la production, on recalibre la machine et on fait repartir la production)

Calcul de  $\hat{m}$  et  $\hat{\sigma}$ 

- Moyenne estimée :

$$\hat{m} = \frac{X_1 + \cdots + X_n}{n} = 100.8.$$

- Écart-type estimé :

$$\hat{\sigma} = \sqrt{\frac{(X_1 - \hat{m})^2 + \cdots + (X_n - \hat{m})^2}{n - 1}} = 2.3.$$

## Flashback : TCL

## Théorème

Soit  $X_1, \dots, X_n$   $n$  variables aléatoires indépendantes de même loi **inconnue** de moyenne  $m_0$  et d'écart-type  $\sigma$ .

Alors, quand  $n$  est grand (ici  $n \geq 30$ ), pour tous nombres réels  $a, b$ ,

$$P\left(a \leq \frac{X_1 + \dots + X_n - m_0}{\frac{\sigma}{\sqrt{n}}} \leq b\right) \simeq P(a \leq \mathcal{N}(0, 1) \leq b).$$

Chez nous, cela va donner :

$$P\left(a \leq \frac{\hat{m} - m_0}{\frac{\hat{\sigma}}{\sqrt{n}}} \leq b\right) \simeq P(a \leq \mathcal{N}(0, 1) \leq b).$$

Ainsi, sous l'hypothèse  $H_0$ , par le TCL,

$$\hat{z} = \frac{\hat{m} - m_0}{\frac{\hat{\sigma}}{\sqrt{n}}} = \frac{100.8 - 100}{2.3/\sqrt{100}} = \frac{0.8}{0.23} \simeq 3.48$$

est censé suivre une loi gaussienne centrée réduite. En particulier, si  $H_0$  est vrai, 95% du temps,  $\hat{z}$  est compris entre  $-1.96$  et  $1.96$  (voir table).

Comme ce n'est pas le cas ici,  $H_1$  est accepté. La machine est dérégulée.

# Flashback : lecture table gaussienne

AFFICHAGE DE LA TABLE

# Les variantes

- Différents types d'hypothèses alternatives :
  - $H_1 : m \neq m_0$  ou
  - $H_1 : m < m_0$  ou
  - $H_1 : m > m_0$ .
- Différents types de données possibles :
  - Loi inconnue et  $n \geq 30$ .
  - Loi de Bernoulli (pourcentage) et  $n \geq 30$ .
  - Loi normale et  $n \leq 30$ .

9 types d'exercices possibles !

## Exemple : pourcentage avec $H_1 : p > p_0$

Un fournisseur vous fournit des boulons dont la probabilité d'être défectueux est de 5%.

Vous avez reçu un lot de 10000 boulons et vous souhaitez savoir si vous l'acceptez ou le refusez.

Pour cela, vous prenez 200 boulons et constatez que 11 d'entre eux sont défectueux.

- Au risque 5%, que faire ?
- Au risque 5%, à partir de combien de boulons défectueux, refusez-vous le lot ?

WOOCLAP

# Formalisation des hypothèses

On a la valeur de référence  $p_0 = 0.05$ .

On a un pourcentage/loi de Bernoulli.

On va tester l'hypothèse

- $H_0 : p = p_0$ , on accepte le lot  
**versus**
- $H_1 : p > p_0$ , on refuse le lot (en effet, si  $H_1$  est vrai alors on a plus de boulons défectueux qu'attendu). En revanche, si  $p < p_0$ , ce n'est pas grave, on a moins de boulons défectueux qu'attendu.



## Flashback du S3 : loi de Bernoulli

- $\Omega(X) = \{0; 1\}$ , deux alternatives seulement.
- $P(X = 1) = p$  et  $P(X = 0) = 1 - p$ .
- $E[X] = 0 \times (1 - p) + 1 \times p = p$ .
- $\text{Var}(X) = E[X^2] - E[X]^2 = (0^2 \times (1 - p) + 1^2 \times p) - p^2 = p - p^2 = p(1 - p)$ .
- $\sigma = \sqrt{\text{Var}(X)} = \sqrt{p(1 - p)}$ .

Calcul de  $\hat{p}$  et de  $\sigma_0$ .

- Moyenne estimée :  $\hat{p} = 11/200 = 0.055$ .
- Écart-type sous  $H_0$ , on va prendre

$$\sigma_0 = \sqrt{p_0(1 - p_0)} = \sqrt{0.05 \times 0.95} = 0.22.$$

- Par le TCL sous  $H_0$ ,

$$\hat{z} = \frac{0.055 - 0.05}{0.22/\sqrt{200}} = \frac{0.005}{0.0155} \simeq 0.322.$$

suit une loi normale centrée réduite.

# Comparaison gaussienne : la table

- $H_0 : p = p_0 = 0.05$  versus  $H_1 : p > p_0 = 0.05$  au risque 5%.
- On détermine la zone de rejet.

PRENEZ DES NOTES

- Comme  $\hat{z} < 1.645$ , on accepte le lot.
- On refuse le lot si  $\hat{z} = \frac{\hat{p} - 0.05}{0.0155} > 1.645$ .  
Donc en résolvant l'inéquation, on trouve  
 $\hat{p} > 0.05 + 1.645 \times 0.0155 = 0.0755$ .  
Ainsi, à partir de  $200 \times 0.0755 = 15.1$ , soit 16 boulons défectueux, on refuse le lot au risque 5%.

# PAUSE

# Test du $\chi^2$

# Prononciation de $\chi$

- $\chi$  est une lettre grecque.
- Cela se prononce Khi.
- On la trouve dans le mot “chiromancie” en français et aussi dans “chirurgie”.

# Utilité d'un test du $\chi^2$

- Adéquation en loi.

Le marché automobile de 2000 ressemblait à

Essence	Diesel	Électrique/Hybride
60%	30%	10%

Est-ce toujours le cas en 2020 ?

On se donne un risque de 10%, et on a les données suivantes, sur 150 véhicules vendus, on a observé

Essence	Diesel	Électrique/Hybride
72	37	41

- Test d'indépendance.

On fait une campagne de publicité, a-t-elle fonctionné ?



# Test d'adéquation en loi

# Test d'adéquation en loi

- Une loi de probabilité qui est donnée (vente de véhicules en 2000) :

Essence	Diesel	Électrique/Hybride
60%	30%	10%

- Données : vente de 150 véhicules en 2020, on a

Essence	Diesel	Électrique/Hybride
72	37	41

- Question : est-ce que la répartition de ventes des véhicules entre 2000 et 2020 a significativement changé ?

WOOCLAP

# Formalisation des hypothèses

On va tester l'hypothèse

- $H_0$  : la vente des véhicules en 2020 ressemble à celle des années 2000 versus
- $H_1$  : ce n'est pas le cas.

# Test d'adéquation en loi

- Une loi de probabilité qui est donnée (vente de véhicules en 2000) :

Essence	Diesel	Électrique/Hybride
60%	30%	10%

- 1ère étape : calcul des valeurs théoriques pour 150 véhicules :

Essence	Diesel	Électrique/Hybride
$150 \times 0.6 = 90$	$150 \times 0.3 = 45$	$150 \times 0.1 = 15$

- Vérification que tous les chiffres dans ce tableau sont  $\geq 5$  et que leur somme est  $\geq 30$ .

# Test d'adéquation en loi

- Valeurs théoriques et valeurs réelles :

	Essence	Diesel	Électrique/Hybride
Théorique	90	45	15
Observé	72	37	41

- Statistique que l'on regarde est

$$\hat{z} = \frac{(90 - 72)^2}{90} + \frac{(45 - 37)^2}{45} + \frac{(15 - 41)^2}{15} = 3.6 + 1.4 + 45.1 = 50.1.$$

De manière générique :

$$\hat{z} = \sum_{i=1}^k \frac{(n_{\text{théorique},i} - n_{\text{observé},i})^2}{n_{\text{théorique},i}}.$$

- Pourquoi cette statistique ?

# Le théorème de probabilité

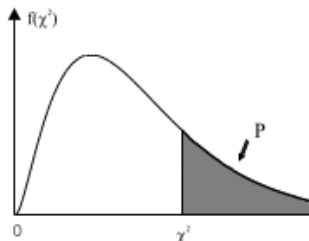
- Remarque : plus  $\hat{z}$  est proche de 0, mieux cela doit être.

Un résultat en probabilité dit que,

- sous  $H_0$ , la statistique  $\hat{z}$  suit une loi du  $\chi^2$  à  $k - 1$  degré de liberté.

## Table de la loi du Khi-deux

Valeurs de  $\chi^2$  ayant la probabilité  $P$  d'être dépassées



Lecture dans la table de la  $\chi^2$ 

AFFICHAGE DE LA TABLE

# Test d'adéquation en loi : conclusion

- Par la table des quantiles de la loi du  $\chi^2$ , un risque de 5% à 2 degrés de liberté correspond à 5.991.
- Comme  $50.1 > 5.991$ , on accepte, au risque 5%, l'hypothèse  $H_1$  qui est que le marché automobile a changé entre 2000 et 2020.



# Test d'indépendance

## Exo 6 du TD2 du S3

Le département GEA2 2A accueille 86 étudiants répartis comme dans le tableau ci-dessous :

filière \ langue	allemand	espagnol
CG2P	3	27
GEMA	6	50

- Le choix entre la filière et la langue est-il indépendant au risque 10% ?

# Flashback : l'indépendance

- Deux événements  $A$  et  $B$  sont indépendants si

$$P(A \text{ et } B) = P(A) \times P(B).$$

- Filière et langue indépendantes, cela veut dire que les 4 choses suivantes sont vraies :
  - $P(\text{CG2P et allemand}) = P(\text{CG2P}) \times P(\text{allemand})$ ,
  - $P(\text{CG2P et espagnol}) = P(\text{CG2P}) \times P(\text{espagnol})$ ,
  - $P(\text{GEMA et allemand}) = P(\text{GEMA}) \times P(\text{allemand})$ ,
  - $P(\text{GEMA et espagnol}) = P(\text{GEMA}) \times P(\text{espagnol})$ .

# Formalisation du test

On va tester l'hypothèse

- $H_0$  : filière et langue sont indépendantes versus
- $H_1$  : ce n'est pas le cas.

# Estimation de la loi CG2P/GEMA et de la loi espagnol/allemand

filière \ langue	allemand	espagnol	Total	Proportion
CG2P	3	27	30	$30/86 \simeq 0.349$
GEMA	6	50	56	$56/86 \simeq 0.651$
Total	9	77	86	100
Proportion	0.105	0.895	100	

Ainsi, les probabilités estimées de chacun est

- $P(\text{CG2P}) \simeq 0.349$  et  $P(\text{GEMA}) \simeq 0.651$ ,
- $P(\text{allemand}) \simeq 0.105$  et  $P(\text{espagnol}) \simeq 0.895$ .

## Effectif théorique si indépendant

Si c'était indépendant, on aurait

- $P(\text{CG2P et allemand}) = 0.349 \times 0.105 \simeq 0.0366$ , et donc  $86 \times 0.0366 = 3.2$  étudiants CG2P et allemand.
- ...

Ainsi le calcul des effectifs théoriques donne :

F \ L	allemand	espagnol
CG2P	$0.349 \times 0.105 \times 86 \simeq 3.2$	$0.349 \times 0.895 \times 86 \simeq 26.9$
GEMA	$0.651 \times 0.105 \times 86 \simeq 5.9$	$0.651 \times 0.895 \times 86 \simeq 50.1$

Il faut vérifier que tous les effectifs théoriques soient  $\geq 5$ .

Ici : on fera comme si,

mais, en vrai, un effectif est  $< 5$ , donc on ne peut pas faire.

## Comparaisons entre les données et le théorique

F \ L	allemand	espagnol
CG2P	obs : 3	27
	théo : 3.2	26.9
	$\hat{z}_{c,a} = \frac{(3-3.2)^2}{3.2} = 0.0125$	0.0004
GEMA	6	50
	5.9	50.1
	$\frac{(6-5.9)^2}{5.9} = 0.0017$	0.0002

- Calcul des  $\hat{z}_{i,j} : \frac{(n_{\text{observé}} - n_{\text{théorique}})^2}{n_{\text{théorique}}}$
- Puis  $\hat{z}$  est la somme des  $\hat{z}_{i,j}$  :

$$\hat{z} = 0.0125 + 0.0004 + 0.0017 + 0.0002 = 0.0148.$$

# Probabilité

- $\hat{z}$  suit une loi du  $\chi^2$  à  $(2 - 1) \times (2 - 1) = 1$  degré de liberté.
- De manière générique : à  $(k_1 - 1)(k_2 - 1)$  degrés de liberté où  $k_1$  c'est le nombre de lignes du tableau et  $k_2$  le nombre de colonnes.
- Dans la table de la  $\chi^2$ , on lit 2.706 pour 1 d.l. et  $\alpha = 0.1$ .
- Comme  $\hat{z} = 0.0148 \leq 2.706$ , on ne rejette pas  $H_0$ .
- Conclusion : le choix de la langue et de la filière semble indépendant.



# Résumé des méthodes vues

# Intervalle de confiance

- 1 Identifier la taille  $n$  de l'échantillon ( $n \leq 30$  ou  $n \geq 30$ ).
- 2 Identifier la loi des données (gaussienne, pourcentage, inconnue).
- 3 Calculer la moyenne empirique  $\hat{m}$  ou  $\hat{p}$ .
- 4 Calculer l'écart-type estimé  $\hat{\sigma}$ .
- 5 La loi de  $\frac{m - \hat{m}}{\hat{\sigma}/\sqrt{n}}$  ?
  - $n \geq 30$  : loi normale.
  - $n \leq 30$  et loi des données gaussienne : loi de Student.
- 6 Trouver  $z$  via la bonne table tel que

$$P\left(-z \leq \frac{m - \hat{m}}{\hat{\sigma}/\sqrt{n}} \leq z\right) = \alpha.$$

- 7 En déduire que l'IC au niveau de confiance  $\alpha$  est

$$\left[ \hat{m} - z \frac{\hat{\sigma}}{\sqrt{n}}; \hat{m} + z \frac{\hat{\sigma}}{\sqrt{n}} \right].$$

# Test de la moyenne ou pourcentage

- 1 Identifier la taille  $n$  de l'échantillon ( $n \leq 30$  ou  $n \geq 30$ ).
- 2 Identifier la loi des données (gaussienne, pourcentage, inconnue).
- 3 Identifier la valeur de référence  $m_0$  ou  $p_0$  et définir l'hypothèse nulle  $H_0 : m = m_0$ .
- 4 **Réfléchir** pour trouver quel  $H_1$  est le mieux entre  $\neq$ ,  $<$  et  $>$ .
- 5 Calculer la moyenne empirique  $\hat{m}$  ou  $\hat{p}$ .
- 6 Calculer l'écart-type estimé  $\hat{\sigma}$  ou  $\sigma_0 = \sqrt{p_0(1 - p_0)}$ .
- 7 La loi de  $\hat{z} = \frac{m_0 - \hat{m}}{\hat{\sigma}/\sqrt{n}}$  ?
  - $n \geq 30$  : loi normale.
  - $n \leq 30$  et loi des données gaussienne : loi de Student.
- 8 Zone de rejet  $Z_r$  à partir de  $\alpha$ ,  $H_1$  et la loi de  $\hat{z}$ .
- 9 Conclure si  $\hat{z} \in Z_r$  ou  $\hat{z} \notin Z_r$  ?

# Test du $\chi^2$ : adéquation en loi

- ① Identifier la taille  $n$  de l'échantillon et vérifier  $n \geq 30$ .
- ② Identifier la loi de référence  $\mathcal{L}_0$ .
- ③ Déterminer  $k$  le nombre de classes de  $\mathcal{L}_0$ .
- ④ Poser  $H_0 : \mathcal{L} = \mathcal{L}_0$  et  $H_1 : \mathcal{L} \neq \mathcal{L}_0$ .
- ⑤ Calculer les effectifs théoriques de chaque classe sous  $\mathcal{L}_0$ .
- ⑥ Vérifier qu'ils sont tous  $\geq 5$ .
- ⑦ Pour chaque classe, calculer  $\frac{(n_{\text{théo}} - n_{\text{obs}})^2}{n_{\text{théo}}}$ .
- ⑧ Les sommer pour trouver  $\hat{z}$ .
- ⑨ Lecture table pour la zone de rejet  $Z_r$  :
  - colonne :  $\alpha$ ,
  - ligne :  $k - 1$ , le degré de liberté de la  $\chi^2$ .
- ⑩ Conclure si  $\hat{z} \in Z_r$  ou  $\hat{z} \notin Z_r$  ?

# Test du $\chi^2$ : test d'indépendance

- ① Identifier la taille  $n$  de l'échantillon et vérifier  $n \geq 30$ .
- ② Déterminer  $k_1$  et  $k_2$  les nombres de classes (lignes et colonnes du tableau).
- ③ Poser  $H_0$  : indépendance et  $H_1$  : non indépendance.
- ④ Estimer les probabilités de chaque classe :
  - probabilité d'une ligne et
  - probabilité d'une colonne.
- ⑤ Calculer les effectifs théoriques de chaque case en supposant l'indépendance :
 

$n \times \text{proba d'être dans la ligne} \times \text{proba d'être dans la colonne.}$
- ⑥ Pour chaque case, calculer  $\frac{(n_{\text{théo}} - n_{\text{obs}})^2}{n_{\text{théo}}}$ .
- ⑦ Les sommer pour trouver  $\hat{z}$ .
- ⑧ Lecture table pour la zone de rejet  $Z_r$  :
  - colonne :  $\alpha$ ,
  - ligne :  $(k_1 - 1)(k_2 - 1)$ , le degré de liberté de la  $\chi^2$ .
- ⑨ Conclure si  $\hat{z} \in Z_r$  ou  $\hat{z} \notin Z_r$  ?

Bonne fin de semaine !