Cours 9 – Tests statistiques: généralités.

Eya ZOUGAR *

Institut National des Sciences appliqués-INSA

Génie mathématiques GM3
Thursday 30th March, 2023

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN NORMANDIE

*Basé sur le cours de Bruno PORTIER

1. Introduction.
1.1. C'est quoi un test statistique ?

Un test statistique est une procédure mathématique qui permet, gràce à un modèle probabiliste, de conclure, avec un risque connu qu'on a fixé à l'avance, quant à l'acceptation ou au rejet d'une hypothèse posée sur le modèle au départ.

Que fait un test ?

En pratique, on cherche à expliquer les différences observées entre une hypothèse posée à priori sur le modèle et le résultat correspondant obtenu à partir des données.

Le test statistique permet de décider si les différences observées sont dues au seul fait du hasard (fluctuations d'échantillonnage) ou bien si elles sont significatives du fait que l'hypothèse posée à priori n'est pas fondée.

1.2. Principe de base d'un test statistique.

D'après l'ouvrage 'Statistiques pour statophobes', un test consiste à:

1. Poser une hypothèse, nommée H_0 ou hypothèse nulle;
2. Calculer, dans la gamme des résultats expérimentaux possibles, ceux qui sont tellement éloignés du résultat moyen attendu selon H_0, que ces résultats n'ont presque aucune chance de se produire si H_0 est vraie;
3. Comparer ces résultats avec celui qui a été réellement obtenu;
4. Conclure que H_0 est peu crédible et donc la rejeter si le résultat obtenu appartient aux résultats qui n'avaient presque aucune chance de se produire si H_0 était vraie;
5. Ou conclure que H_0 reste crédible et donc ne pas la rejeter si le résultat obtenu appartient aux résultats qui avaient une chance de se produire si H_0 était vraie.

En statistique, le "presque aucune chance" se traduit par "dans moins de 5% des cas où H_0 est vraie".

1.3. Exemple: Jet d'une pièce.

On considère une pièce bien équilibrée (non truquée).

En la jetant sur une table, on a donc une chance sur deux d'obtenir un pile et une chance sur deux d'obtenir un face.

On jette 10 fois de suite cette pièce et on comptabilise le nombre de "Face" obtenu. On observe 9 fois le coté Face, soit une proportion de 9/10.

La pièce est-elle vraiment bien équilibrée (non truquée)?

En effet, on se serait attendu à trouver une proportion de faces de 1/2, voire 2/5 ou 3/5, mais pas 9/10.

Ce résultat surprenant est-il dû au seul fait du hasard ou bien est-il dû au fait qu'en réalité la pièce utilisée n'est pas bien équilibrée ?

2. Détails des étapes d'un test statistique.
2.1. Les données et le modèle probabiliste.

Le point de départ, c'est toujours les données du problème.

Pour simplifier, on se place dans le cadre univarié.

On dispose de n données unidimensionnelles $x_1, x_2, \ldots, x_n$ qui sont des mesures ou des observations d'une variable qualitative ou quantitative.

Introduire un modèle probabiliste, c'est faire l'hypothèse fondamentale que:

- Ces données sont en fait n réalisations indépendantes d'une variable aléatoire X de loi continue ou discrète,
- Ou bien encore que ces données sont les réalisations de n variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et de même loi.

Exemple:

Dans le cas du lancer de pièces, on peut associer aux résultats obtenus la suite $x_1, x_2, \ldots, x_{10}$ avec pour tout entier $j = 1, \ldots, 10$, $x_j = 1$ si le résultat obtenu au $j$ème lancer est Face, et $x_j = 0$ si le resultat est Pile.

On peut alors considérer que les données (x_j) sont les réalisations de variables aléatoires (X_j) indépendantes et de même loi de Bernouilli de paramètre p, c'est à dire qu'on a

$$\mathbb{P}[X_j = 1] = p \quad et \quad \mathbb{P}[X_j = 0] = 1 - p.$$

Nous disposons donc d'un modèle probabiliste.

Si la pièce n'est pas truquée, les variables aléatoires $X_1, X_2, \ldots, X_{10}$ sont de loi de Bernouilli de paramètre $p = 1/2$.

2.2. Les hypothèses nulle et alternative.
2.2.1. Définitions.

Disposant d'un modèle probabiliste, on souhaite vérifier une hypothèse posée à priori sur ce modèle:

- On appelle cette hypothèse, l'hypothèse nulle, et on la notera H_0.
- On notera H_1 l'hypothèse alternative que l'on souhaite confronter à cette hypothèse nulle.
  En général, H_1 est la négation de H_0.

Exemple : Dans le cas du lancer de pièce, notre hypothèse de base est que la pièce n'est pas truquée et donc que $p = 0,5$.

Pour vérifier cette hypothèse, nous allons tester l'hypothèse nulle $H_0 : "p = 0,5"$ contre une hypothèse alternative qui pourra être $H_1 : "p \neq 0,5"$ ou bien $H_1 : "p > 0,5"$ en fonction de l'objectif recherché.

2.3. Statistique de test et sa loi sous H_0.

Une fois :

- le modèle probabiliste introduit,
- l'hypothèse nulle H_0 définie,

Il faut pouvoir disposer d'une variable aléatoire encore appelée statistique de test, que l'on nommera Z

- dont on connait la loi sous H_0,
- et dont la loi diffère sous H_1, pour pouvoir séparer H_0 de H_1.

Bien évidemment,

- Z est une fonction mesurable des variables aléatoires $X_1, X_2, \ldots, X_n$, c'est à dire que $Z = T(X_1, X_2, \ldots, X_n)$.
- Z doit pouvoir être calculée sur les données observées. Notons z_{obs} la valeur de Z sur les données : $z_{obs} = T(x_1, x_2, \ldots, x_n)$.

2.4. Zone de rejet de H_0.

- Connaissant la loi de Z, on peut déterminer ses valeurs les plus extrêmes , les plus éloignées de la valeur moyenne attendue, autrement dit les valeurs qui n'ont presque aucune chance d'être observées si H_0 est vraie.
- Notre connaissance de la distribution de Z sous H_0 nous permet en particulier de calculer avec précision les gammes de valeurs extrêmes qui seront observées avec une probabilité α que nous pouvons choisir librement.

L'idée est bien entendu de choisir α petit.
On prend généralement $\alpha = 0,05 = 5\%$ et on définit 2 zones de rejet de H_0, contenant chacune $\alpha/2 = 2,5\%$ de la distribution de Z sous H_0:

- L'une de ces zones de rejet concerne les 2,5% des valeurs extrêmes de Z "incroyablement élevées" par rapport à la valeur moyenne attendue sous H_0 ;
- L'autre zone de rejet concerne les 2,5% des valeurs extrêmes de Z "incroyablement basses" par rapport à la valeur moyenne attendue sous H_0.

2.5. Conclusion du test et Risque de première espèce.

On conclut le test, à partir de la valeur observé z_{obs} de Z, de la manière suivante :

- Si z_{obs} appartient à une des deux zones de rejet de H_0, on rejette H_0 : selon l'argument que si H_0 était vraie, on aurait (presque) jamais pu observer une telle valeur de Z. Il est donc plus raisonnable d'accepter l'hypothèse selon laquelle H_0 est probablement fausse (ce qui expliquerait très facilement le résultat obtenu) ;
- Si z_{obs} n'appartient à aucune des deux zones de rejet de H_0, on ne rejette pas H_0, ce qui signifie qu'on considère ne pas avoir d'éléments suffisamment solides pour la déclarer peu crédible. Mais attention, cela ne signifie pas qu'on a montré que H_0 est vraie.

Bien évidemment cette décision ne va pas sans risque :
En effet, on peut décider de rejeter H_0 même si elle est vraie. Mais ce risque est maîtrisé puisque c'est nous qui l'avons choisi.
Ce risque est appelé **Risque de première espèce**.

2.6. Résumé de la construction d'un test.

En résumé, la mise en oeuvre d'un test nécessite :

1. Associer aux données de l'étude des variables aléatoires (introduction du modèle) ;
2. On définit les hypothèses H_0 et H_1 ;
3. On choisit le risque α ;
4. On introduit la statistique de test Z et on précise sa loi sous H_0 ;
5. On calcule la zone de rejet $ZR_{H_0,\alpha}$ de H_0 au risque α ;
6. On calcule la réalisation z_{obs} de Z sur les données ;
7. On conclut :
   - si $z_{obs} \notin ZR_{H_0,\alpha}$, alors on ne peut pas rejeter H_0 ;
   - si $z_{obs} \in ZR_{H_0,\alpha}$, alors on rejette H_0 au risque α.

2.7. Exemple sur le lancer d'une pièce.

Pour vérifier que la pièce n'est pas truquée, on test au risque 5%,

$$H_0 : "p = 0,5" \quad \text{contre} \quad H_1 : "p \neq 0,5"$$

1. Puisque les var. a. $X_1, X_2, \ldots, X_{10}$ sont iid de même loi de Bernouilli de paramètre p, alors $Z = \sum_{j=1}^{10} X_j$ suit une loi binomiale $B(10, p)$.
2. Sous H_0, le paramètre p est égal à $1/2$.
   - Donc sous H_0, Z suit une loi $B(10, 1/2)$.
   - Bien évidemment, sous H_1, Z ne suit plus une loi $B(10, 1/2)$.
3. Les quantiles à 0,025 et 0,975 d'une $B(10, 1/2)$ sont respectivement 2 et 8. La zone de rejet de H_0 au risque 5% est donc de la forme $ZR_{H_0} = \{0, 1\} \bigcup \{9, 10\}$.
4. Nous avons observé 9 faces, donc la valeur de la statistique de test Z sur les donnée est égale à $z_{obs} = 9$
5. Puisque $z_{obs} \in ZR_{H_0}$, on rejette H_0 au risque de 5%. On considère donc que la pièce n'est pas équilibrée ou truquée.

3. Test de conformité d'une moyenne (Test de Student).

3.1. Le problème.

On considère n données réelles $x_1, x_2, \ldots, x_n$ qui sont les mesures d'une variable quantitative.

On se place dans le cadre de l'échantillon gaussien, c'est à dire que l'on suppose que les données $x_1, \ldots, x_n$ sont les réalisations de n variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$, les paramètres m et σ^2 étant supposés inconnus.

On veut savoir au risque α si la moyenne théorique m des données est différente ou non d'une valeur m_0 donnée a priori (valeur de référence ou supposée : $m_0 = 0$ par exemple).

Pour cela, on teste, au risque α, l'hypothèse nulle

$$H_0 : " m = m_0 "$$

contre l'hypothèse alternative

$$H_1 : " m \neq m_0 "$$

3.2. Un exemple d'application.

- A la suite d'un traitement (régime alimentaire) sur une variété de porcs, on prélève un échantillon de 5 porcs et on les pèse.
- On obtient les poids suivants (en kg) : 83 ; 81 ; 84 ; 80 ; 85.
- On sait que le poids moyen pour cette variété de porcs est de 87,6 kg.
- On suppose que le poids de cette variété de porcs est normalement distribué.
- Le poids moyen des porcs traités diffère-t-il significativement de cette norme au seuil de 5 % ?

3.3. Choix de la statistique de test.
3.3.1. Le résultat clé.

Il nous faut donc maintenant choisir une statistique de test.
On sait que pour tout $m \in \mathbb{R}$,

$$\frac{\sqrt{n}(\overline{X}_n - m)}{S} \sim T_{n-1}$$

$$\text{avec} \quad \overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \quad \text{et} \quad S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

3.3.2. Choix de la statistique de test. Loi sous H_0.

Sous H_0, le paramètre m est égal à m_0.

Donc, sous H_0, la variable Z définie par

$$Z = \frac{\sqrt{n}(\overline{X}_n - m_0)}{S}$$

suit une loi de Student à $n-1$ ddl, et on note $Z \underset{H_0}{\sim} T_{n-1}$.

3.3.3. Loi de la statistique de test sous H_1.

En revanche, sous H_1, Z ne suit plus une T_{n-1} puisque on a la décomposition suivante :

$$Z = \underbrace{\frac{\sqrt{n}(\overline{X}_n - m)}{S}}_{\sim T_{n-1}} + \underbrace{\frac{\sqrt{n}(m - m_0)}{S}}_{\neq 0 \text{ sous } H_1}$$

ainsi Z satisfait les conditions pour être une statistique de test.

3.4. Construction du test de conformité.

• Donc, pour tester au risque α l'hypothèse nulle H_0 : "$m = m_0$" contre l'hypothèse alternative H_1 : "$m \neq m_0$",

on utilise la statistique de test $Z = \dfrac{\sqrt{n}(\overline{X}_n - m_0)}{S} \underset{H_0}{\sim} T_{n-1}$.

• La zone de rejet est de la forme $\{|Z| > t_{n-1}\}$ où t_{n-1} est le quantile d'ordre $(1 - \alpha/2)$ de la loi de Student à $n - 1$ ddl, c'est à dire t_{n-1} est tel que

$$\mathbb{P}_{H_0}\left[|Z| \leq t_{n-1}\right] = \mathbb{P}\left[|T_{n-1}| \leq t_{n-1}\right] = 1 - \alpha$$

• On calcule alors la valeur z_{obs} de Z sur les données :

$$z_{obs} = \frac{\sqrt{n}(\overline{x}_n - m_0)}{s}.$$

• Si $|z_{obs}| \leq t_{n-1}$, alors on ne peut pas rejeter l'hypothèse nulle H_0 et on considère que la moyenne observée n'est pas significativement différente de m_0.

Sinon, on rejette l'hypothèse nulle H_0 au risque α, et on considère que la moyenne des données est significativement différente de m_0.

3.5. Illustration sur l'exemple des porcs.
3.5.1. Les hypothèses et la statistique de test.

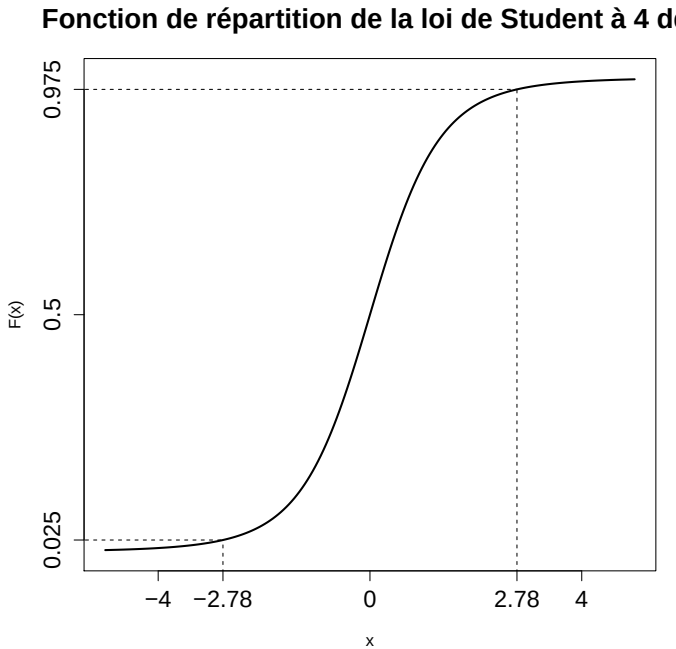
• On note x_1, x_2, x_3, x_4, x_5 les 5 poids mesurés. Puisque l'on peut considérer que le poids est distribué selon une loi normale, on peut considérer que les données $x_1, \ldots, x_5$ sont en fait les réalisations de 5 variables aléatoires $X_1, \ldots, X_5$ indépendantes et identiquement distribuées de loi $\mathcal{N}(m, \sigma^2)$.

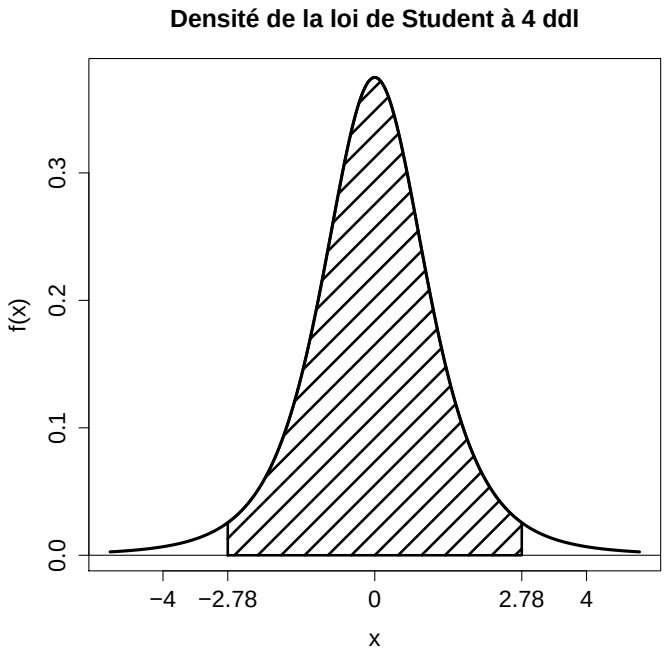
• Pour savoir si le poids moyen observé diffère de la norme ou non, nous allons tester au risque 5% l'hypothèse nulle H_0 : "$m = 87,5$" contre H_1 : "$m \neq 87,5$".

• Pour ce faire, on utilise la statistique de test $Z = \dfrac{\sqrt{5}\,(\overline{X}_5 - 87,5)}{S}$ qui suit sous H_0 une loi de Student à 4 ddl.

3.5.2. Zone de rejet et conclusion.

La zone de rejet de H_0 à 5% est de la forme $\{|Z| > t_4 = 2,776\}$.





On calcule maintenant la valeur z de la statistique de test Z sur les données : on a $z = \dfrac{\sqrt{5}\,(\bar{x}_5 - 87,5)}{s} = \dfrac{\sqrt{5}\,(82,6 - 87,5)}{2,074} = -5,2838.$

Puisque $|z| = 5,2838 > 2,776$, on rejette l'hypothèse nulle au risque 5% et on considère que le poids moyen de la variété de porcs diffère de la norme.

3.5.3. Illustration avec le logiciel R.

Avec le logiciel R, on résoud le problème avec les instructions suivantes :

```
X = c(83 , 81 , 84 , 80 , 85)
t.test(X, mu= 87.5 , conf.level=0.95)
```

On obtient le résultat suivant :

```
> t.test(X, mu= 87.5 , conf.level=0.95)

^^IOne Sample t-test

data:  X
t = -5.2838, df = 4, p-value = 0.006154
alternative hypothesis: true mean is not equal to 87.5
95 percent confidence interval:
 80.02523 85.17477
sample estimates:
mean of x
     82.6
```

3.6. Illustration par simulations avec le logiciel R.
3.6.1. Pourquoi faire des simulations.

Les simulations permettent, dans le cadre des tests statistiques, d'étudier

- le niveau empirique du test
- ainsi que la puissance empirique du test.

Pour étudier le niveau empirique du test, on se place sous H_0, et on compare le pourcentage de rejet à tort, au niveau théorique que l'on a fixé.

Pour étudier la puissance empirique du test, on se place sous H_1, et on compare le pourcentage de rejet (bonnes décisions) à 100%.

3.6.2. Non rejet de H_0 à raison.

On simule 20 réalisations d'une loi normale centrée:

-0.56	-0.23	1.56	0.07	0.13	1.72	0.46	-1.27	-0.69	-0.45
1.22	0.36	0.40	0.11	-0.56	1.79	0.50	-1.97	0.70	-0.47

On veut savoir au risque 5% si la moyenne théorique de la loi de cet échantillon est nulle ou non.

On met en oeuvre le test précédent :
On teste donc H_0 : "$m = 0$" contre H_1 : "$m \neq 0$".
On prend la statistique de test

3.6.2. Non rejet de H_0 à raison.

On simule 20 réalisations d'une loi normale centrée:

-0.56	-0.23	1.56	0.07	0.13	1.72	0.46	-1.27	-0.69	-0.45
1.22	0.36	0.40	0.11	-0.56	1.79	0.50	-1.97	0.70	-0.47

On veut savoir au risque 5% si la moyenne théorique de la loi de cet échantillon est nulle ou non.

On met en oeuvre le test précédent :
On teste donc H_0 : "$m = 0$" contre H_1 : "$m \neq 0$".
On prend la statistique de test

$$Z = \frac{\sqrt{20}\,\overline{X}_{20}}{S} \underset{H_0}{\sim} T_{19}$$

3.6.2. Non rejet de H_0 à raison.

On simule 20 réalisations d'une loi normale centrée:

-0.56	-0.23	1.56	0.07	0.13	1.72	0.46	-1.27	-0.69	-0.45
1.22	0.36	0.40	0.11	-0.56	1.79	0.50	-1.97	0.70	-0.47

On veut savoir au risque 5% si la moyenne théorique de la loi de cet échantillon est nulle ou non.

On met en oeuvre le test précédent :
On teste donc H_0 : "$m = 0$" contre H_1 : "$m \neq 0$".
On prend la statistique de test

$$Z = \frac{\sqrt{20}\,\overline{X}_{20}}{S} \underset{H_0}{\sim} T_{19}$$

La zone de rejet est $\{|Z| > 2,09\}$.
Pour cet échantillon, on trouve $z = 0,651$.

Puisque $|z| \leq 2,09$ on ne rejette pas (à raison) l'hypothèse nulle H_0, et on considère que la moyenne est non significativement différente de 0.

3.6.3. Rejet de H_0 à tort, puis à raison.

On simule un nouvel échantillon :

1.48	0.87	0.54	0.04	0.63	1.15	-0.71	0.11	0.57	0.44
1.36	-0.21	0.31	0.33	0.96	0.73	1.92	0.00	0.29	1.54

Pour cet échantillon, nous sommes amenés à rejeter à tort H_0. En effet, on a $\overline{x}_{20} = 0,62$ et une valeur de $z = 4,247 > 2,09$ (p-value $= 4.10^{-4}$).

Pour finir, on simule un échantillon de 20 valeur.

-0.07	0.78	-0.03	0.27	0.37	-0.69	1.84	1.15	-0.14	2.25
1.43	0.70	1.90	1.88	1.82	1.69	1.55	0.94	0.69	0.62

Vérifions que le test rejette l'hypothèse nulle H_0.
Pour cet échantillon, on trouve $\overline{x}_{20} = 0,95$ et une valeur de $z = 5,1123 > 2,09$