Chapitre I : Introduction aux tests statistiques

Cours de tests d'hypothèses pour l'analyse bivariée - Olivier Bouaziz

2022-2023

Programme tests

- Chapitre I : Introduction sur les tests
- Chapitre II : Tests paramétriques de comparaison de deux échantillons
- Chapitre III : Tests du Chi-deux
- Chapitre IV : Tests de corrélation

Planning tests

Au total, 10 séances de 3h découpés en cours/TD/TP :

- Interro : 6ème semaine, le mardi 13 décembre 2022.
- 1 compte rendu de TP.

Exemple sur la salmonelle

Plusieurs personnes ont soufferts d'intoxication alimentaire de salmonelle. Après enquête, on suspecte une marque de glaces d'être responsable de l'intoxication. Pour vérifier cela, des scientifiques ont mesuré le niveau de salmonelle dans 9 pots de glace, tirés de façon aléatoire chez le fabricant de glace. On obtient les mesures suivantes, exprimées en NPP (Nombre le Plus Probable) par gramme :

```
x<-c(0.175,0.205,0.76,0.719,0.199,0.529,0.306,0.52,0.01)
```

- On sait que le niveau de référence que doit respecter un fabricant de glaces est de 0.3 NPP/g.
- La question que l'on se pose donc, est : le niveau moyen de salmonelle dans les pots de glace est-il supérieur à 0.3 ?

```
mean(x)
```

```
## [1] 0.3803333
```

Exemple sur la salmonelle

- Même si sur cet échantillon de 9 pots de glace, la moyenne est supérieur à 0.3, est-ce que cela suffit pour demander au fabricant d'arrêter la production de glace ?
- Imaginons que l'on recommence l'expérience sur 9 autres pots de glace et que l'on trouve cette fois une moyenne de salmonelle égale à 0.22 NPP/g. Que doit-on alors conclure ?
- Que se passe-t-il si on recommence l'expérience cette fois-ci sur 100 pots de glace (au lieu de seulement 9) ? Sera-t-il plus facile de conclure ?

Exemple sur l'effet d'insecticides

Pour comparer l'effet entre deux insecticides, on applique l'insecticide A et B sur 36 groupes de 30 insectes et après avoir laissé agir pendant une heure, on compte le nombre d'insectes morts.

On obtient pour l'insecticide A, les nombres d'insectes tués suivants :

```
##  [1] 14  2 21  5 20  2 17  3 11  7 11  3  5 23  1  4 16 10  3  6  4
## [26] 13  4  3 24 14  9  5 21 17 15  6
```

dont la moyenne est :

```
## [1] 9.44
```

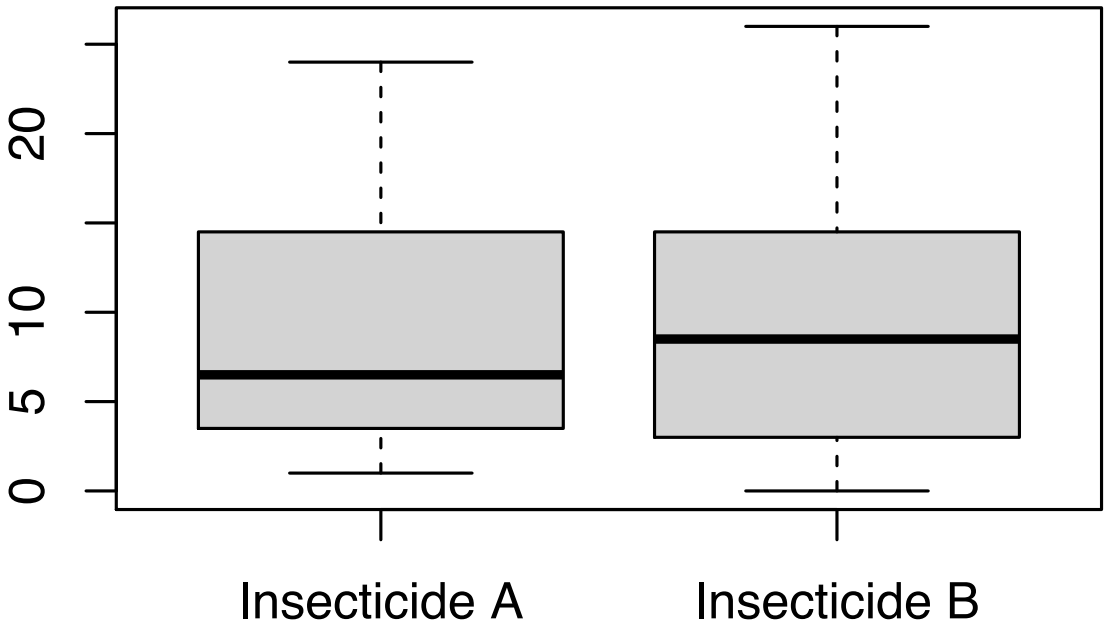
Pour l'insecticide B, les nombres d'insectes tués sont :

```
##  [1]  1 26  2  5 17 13 15  3 26 14  1 19  5  1 11  2  4 20  3 10  3
## [26] 12 13  5 13  6  0 10 16  7 12 17
```

dont la moyenne est :

```
## [1] 9.56
```

Exemple sur l'effet d'insecticides



- A partir de ces résultats, peut-on dire que l'insecticide B marche mieux que le A ?
- Si on pouvait recommencer l'expérience, pourrait-on parfois avoir la moyenne du nombre d'insectes tués par l'insecticide A qui soit supérieure au nombre d'insectes tués par l'insecticide B ?
- Le problème est qu'en pratique, on ne dispose que d'un échantillon.

Idée des tests

A partir de ces exemples, on voit bien que les résultats observés dépendent :

- des échantillons choisis. En pratique, on ne dispose que d'un échantillon ! Est-ce que l'échantillon est bien représentatif des données ? Est-on en présence d'un échantillon atypique ?
- de la taille des échantillons. On a l'intuition que plus l'échantillon est grand plus il sera facile de discriminer entre deux hypothèses.
- de la variance des observations. On a l'intuition que plus la variance est faible plus il sera facile de discriminer entre deux hypothèses.

Idée des tests : risques de première et deuxième espèces

En pratique, on dispose de deux hypothèses, l'hypothèse **nulle** (H_0) et l'hypothèse **alternative** (H_1) et le but est d'essayer de discriminer entre ces deux hypothèses.

Dans l'exemple sur la salmonelle, on représente par X la variable aléatoire (v.a) "quantité de salmonelle dans un pot de glace (en NPP/g)' '. On note $\mu = \mathbb{E}[X]$ et $\sigma^2 = \mathbb{V}[X]$. On souhaite tester :

$$(H_0) \ : \ \mu \leq 0.3 \text{ contre } (H_1) \ : \ \mu > 0.3$$

Il existe deux *mondes* :

- dans le *monde* de (H_0), la quantité de salmonelle des pots de glace ne dépasse pas le niveau réglementaire de 0.3 NPP/g.
- dans le *monde* de (H_1), la quantité de salmonelle des pots de glace dépasse le niveau réglementaire de 0.3 NPP/g et une personne peut tomber **gravement** malade si elle mange de la glace !

Idée des tests : risques de première et deuxième espèces

Deux mondes d'hypothèses = deux risques possibles :

1. dans le monde de (H_0) (les pots de glace ne présentent pas de risque d'intoxication), on peut se **tromper** en choisissant (H_1) et en disant que la glace peut être dangereuse pour la santé (alors que ce n'est pas le cas) !

Ce risque s'appelle le risque de **1ère espèce**.

2. dans le monde de (H_1) (les pots de glace sont dangereux pour la santé), on peut se **tromper** en choisissant (H_0) à la place et en disant que la glace n'est pas dangereuse pour la santé (alors qu'elle l'est) !

Ce risque s'appelle le risque de **2ème espèce**.

Idée des tests : risques de première et deuxième espèces

On voit que les deux risques ne sont pas symmétriques ! Selon le consommateur ou le fabricant il y a un risque plus **grave** que l'autre.

- Pour le fabricant, le plus **grave** est de dire que sa glace peut être dangereuse pour la santé alors qu'elle ne l'est pas !
- Pour le consommateur, le plus **grave** est de dire que la glace n'est pas dangereuse pour la santé alors qu'elle l'est !
- En pratique on n'arrive généralement pas à contrôler les deux risques. Il faut choisir ce que l'on veut montrer et quel risque on veut contrôler.

Idée des tests : risques de première et deuxième espèces

(H_0) : $\mu \leq 0.3$ contre (H_1) : $\mu > 0.3$

	Décision : $\mu \leq 0.3$	Décision : $\mu > 0.3$
Réalité : $\mu \leq 0.3$	$1-\alpha$	α
Réalité : $\mu > 0.3$	β	$1-\beta$

- α s'appelle le risque de **1ère espèce**. $\alpha = \mathbb{P}_{H_0}[\text{rejeter } H_0]$.
- β s'appelle le risque de **2ème espèce**. $\beta = \mathbb{P}_{H_1}[\text{ne pas rejeter } H_0]$.

Idée des tests : risques de première et deuxième espèces

En pratique, le risque de **1ère espèce** α est fixé à l'avance $(5\%, 10\%)$, c'est un risque que l'on contrôle.

- Si on décide de rejeter (H_0) on connait le pourcentage d'erreur de se tromper.
- Par contre, si on décide de ne pas rejetter (H_0), on ne connait généralement pas la marge d'erreur que l'on commettrait en choisissant (H_0).
- C'est pourquoi, on décide toujours de mettre dans (H_1) l'hypothèse que l'on veut montrer.

Vers une règle de décision

- Pour illustrer la méthode sur les données de salmonelle, on suppose que la variance de la quantité de salmonelle dans un pot de glace est connue et est égale à 0.08. On suppose également que la loi de la quantité de salmonelle dans un pot de glace est une loi gaussienne.
- Soient $X_1, \ldots, X_n$, n variables aléatoires (v.a) de loi $\mathcal{N}(\mu, 0.08)$, μ inconnue. Ici $n = 9$.
- $x_1, \ldots, x_n$ sont les réalisations de $X_1, \ldots, X_n$ (ce sont les valeurs observées sur l'échantillon : $x_1 = 0.175, x_2 = 0.205 \ldots$)
- La règle de décision est basée sur μ qui est inconnue. On estime donc μ par :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Rappeler pourquoi $\bar{X}$ est un bon estimateur de μ (voir cours d'ISI).

Vers une règle de décision

Pour fixer les idées on suppose que l'on teste :

$$(H_0) \ : \ \mu = 0.3 \text{ contre } (H_1) \ : \ \mu = 0.4$$

- Sous (H_0), $\bar{X} \sim$?
- Sous (H_1), $\bar{X} \sim$?
- Que peut-on dire, sous (H_0), de la loi de

$$\sqrt{9}(\bar{X} - 0.3)/\sqrt{0.08} = 10.61 \times (\bar{X} - 0.3) \ ?$$

- Que peut-on dire, sous (H_1), de la loi de

$$\sqrt{9}(\bar{X} - 0.3)/\sqrt{0.08} = 10.61 \times (\bar{X} - 0.3) \ ?$$

- $T_n = \sqrt{9}(\bar{X} - 0.3)/\sqrt{0.08}$ s'appelle la statistique de test, sa loi sous (H_0) doit être connue.

Vers une règle de décision

- On rejettera donc (H_0) dès que T_n est *grand*.
- La question est de savoir à partir de quel **seuil** on choisit (H_1) plutôt que (H_0).
- On cherche $c_{0.05}$ tel que si $T_n > c_{0.05}$, on rejette (H_0) et si $T_n \leq c_{0.05}$, on ne rejette pas (H_0).
- Plus précisément : si T_n est supérieur à un seuil $c_{0.05}$, ce qui n'a que 5% de chances d'arriver sous (H_0), alors on choisira (H_1). Si $T_n \leq c_{0.05}$, on ne rejettera pas (H_0) faute de preuves suffisantes.
- On trouve $c_{0.05}$ à l'aide de la table de la loi normale ! On conclut : $R_{0.05} = \{T_n > 1.645\}$.
- Sur notre échantillon, $\bar{x} = 0.38$ et donc $t_n = \sqrt{9}(0.38 - 0.3)/\sqrt{0.08} \approx 0.852$. On ne rejette donc pas (H_0) !
- Au vue des données, nous ne disposons pas de suffisamment d'informations pour pouvoir rejetter (H_0). On dit que le test n'est pas significatif au niveau 5%.

Degré de signification ou p-valeur du test

- Même si on ne rejette pas (H_0), on peut se demander si au risque 10%, 15% on rejetterait (H_0).
- D'une manière générale, en augmentant le risque de se tromper, on a plus de chances de pouvoir rejetter (H_0).
- Un indicateur intéressant est le degré de signification ou p-valeur du test qui représente le plus petit seuil pour lequel on rejette (H_0).
- De manière équivalente, celà revient à se demander dans quelle mesure la moyenne $\bar{x} = 0.38$ trouvée sur l'échantillon est elle compatible avec l'hypothèse (H_0) : $\mu = 0.3$.

On calcule :

$$\text{p-val} = \mathbb{P}_{H_0}[T_n > 0.852] = \cdots = 0.197$$

Celà signifie que si (H_0) était vraie (i.e la glace ne contient pas trop de salmonelle), alors les seules fluctuations d'échantillonnage auraient environ 19.7 chances sur 100 de conduire à une valeur moyenne de salmonelle dans les pots de glace aussi élevée.

Risque de deuxième espèce

On peut maintenant essayer de calculer le risque de se tromper que l'on prend en choisissant (H_0) à la place de (H_1).

On écrit d'abord : $R_{0.05} = \{T_n > 1.645\} = \{\bar{X} > 0.455\}$.

$$\begin{aligned}\beta &= \mathbb{P}_{H_1}[\text{ne pas rejetter } (H_0)] \\ &= \mathbb{P}_{H_1}[\bar{X} \leq 0.455] \\ &= \cdots \\ &= 0.721\end{aligned}$$

On a environ 72.1% de chances de se tromper en choisissant (H_0) alors que c'est (H_1) qui est vraie !

Puissance de test

La puissance d'un test représente la probabilité de prendre la bonne décision en rejettant l'hypothèse nulle : c'est la probabilité de rejetter (H_0) à raison :

$$\begin{aligned}\pi &= \mathbb{P}_{H_1}[\text{rejetter } (H_0)] \\ &= 1 - \beta \\ &= 1 - 0.721 = 0.279\end{aligned}$$

On a environ 0% de chances de détecter à raison le rejet de (H_0).

Remarques :

- La puissance d'un test est d'autant plus grande que l'hypothèse alternative est éloignée de l'hypothèse nulle !
- La forme de la zone de rejet est indiquée par (H_1) mais le seuil c_α ne dépend que de (H_0) et de α (le risque de 1ère espèce).
- Les deux hypothèses (H_0) et (H_1) ne jouent pas le même rôle. On contrôle avant tout le risque de rejetter à tort (H_0) ! En pratique, on n'est généralement pas capable de contrôler le risque de deuxième espèce.

Influence du niveau α du test

Le risque de 1ère espèce α s'appelle également le **niveau** du test.

Que se passe-t-il quand le niveau du test change ?

- Refaire le test au niveau 10%. Au niveau 20%.
- Si α grandit, on tolère une plus grande probabilité d'erreur en rejettant (H_0) et donc β diminue et la puissance augmente.
- Si α diminue, la règle de décision est plus stricte : on n'abandonne moins souvent (H_0) au profit de (H_1) et donc β augmente et la puissance diminue.
- L'idéal serait d'avoir un niveau de test *petit* et une puissance *grande*. Un moyen d'y parvenir est d'augmenter la taille d'échantillon.

Influence de la taille d'échantillon sur le test

Plus la taille d'échantillon est grande et plus les estimateurs sont précis et donc plus il est facile de discriminer entre deux hypothèses.

- Pour le niveau $\alpha = 5\%$, quel est le nombre d'observations nécessaires pour avoir une puissance de 80% ?
- Pour le niveau $\alpha = 5\%$, quel est le nombre d'observations nécessaires pour avoir une puissance de 90% ?

Influence de la formulation des hypothèses

Que se passe-t-il si l'on échange les hypothèses du test. Supposons que l'on teste :

$$(H_0) : \mu > 0.3 \text{ contre } (H_1) : \mu \leq 0.3$$

ou pour faciliter les calculs,

$$(H_0) : \mu = 0.4 \text{ contre } (H_1) : \mu = 0.3$$

- La conclusion du test est-elle la même que précédemment ?
- Calculer la p-valeur du test.
- Calculer la puissance du test.

Application du test à un échantillon sous R

Généralement, en pratique, la variance des observations n'est pas connue. On verra comment faire dans ce cas là au prochain chapitre. Sous R, la procédure est facile à implémenter.

```
x<-c(0.175,0.205,0.76,0.719,0.199,0.529,0.306,0.52,0.01)
t.test(x=x,alternative="greater",mu=0.3)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 0.92005, df = 8, p-value = 0.1922
## alternative hypothesis: true mean is greater than 0.3
## 95 percent confidence interval:
##  0.2179689       Inf
## sample estimates:
## mean of x
## 0.3803333
```