Rappels III : Test de comparaison de proportion à une valeur de référence

Master 2 Recherche SES-IES Analyse de données

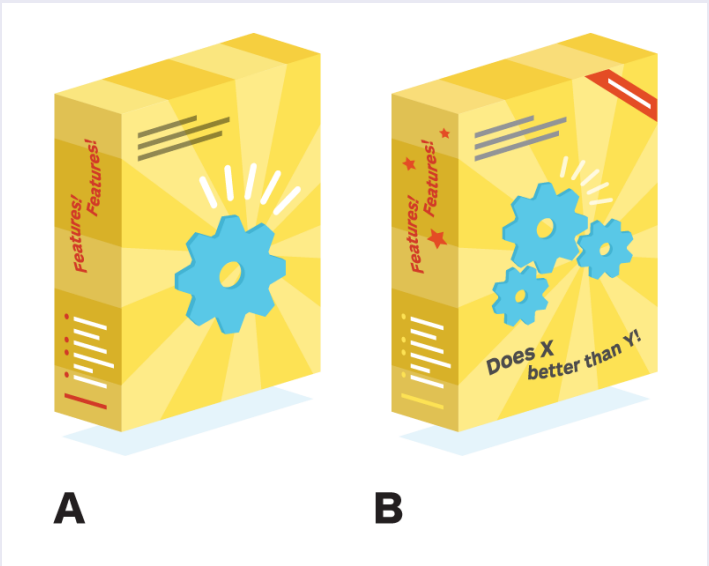
Ana Karina Fermin

Université Paris Nanterre

http://fermin.perso.math.cnrs.fr/

Packaging A ou packaging B

On demande à des consommateurs s'ils préfèrent, pour un produit de grande consommation qu'on veut relooker, le packaging A ou le packaging B.



Des données à une modélisation

Exemple : Packaging A ou packaging B

- On demande à des consommateurs s'ils préfèrent, pour un produit de grande consommation qu'on veut relooker, le packaging A ou le packaging B.
- On interroge n personnes dans un panel de consommateurs et on inscrit les résultats dans un tableau.

Consommateur n°	1	2	3	4	5	6	. . .
Résultat	A	A	B	A	B	B	. . .

- Problème métier : choisir entre deux packaging. Choisir le packaging qui se vend le mieux !
- Idée : se baser sur des données pour prendre la décision

Quel est le travail du statisticien ?

- Quel est le problème métier ?
- Donner un modèle
- Proposer une méthodologie
- Répondre au problème métier à partir des résultats obtenus en prenant en compte l'incertitude (on ne pas sure à 100 %, il faut prendre des risques et accepter de ne pas avoir toujours raison)

Exercice : supposons qu'en 100 consommateurs sondés, on ait eu 58 votes en faveur de A et 42 en faveur de B. À partir de cet échantillon,

1. que peut-on déduire sur la valeur de p ?
2. peut on conclure "statistiquement" que le packaging A est préféré au packaging B ?

Codage : on code la préférence pour A par 0 et celle pour B par 1.

Consommateur numéro	1	2	3	4	5	6	. . .
Résultat	0	0	1	0	1	1	. . .

Contexte :

- Population : Consommateurs
- Variable : réponse du consommateur, codée par une variable X qui prends de valeurs 0 ou 1.
- On note x_1, x_2, x_3, ..., les résultats successifs.
  Par exemple, le troisième consommateur interrogé préfère le packaging B ($x_3 = 1$).

Le mathématicien se place aux instants avant les interrogations d'un consommateur donné et considère celles-ci comme des expériences aléatoires : il note $X_1, X_2, X_3, \ldots$ les variables aléatoires correspondantes.

Imaginons qu'on tire un petit nombre de consommateurs dans un grand panel.

Exemple : Packaging A ou B

Cons. n°	Résultat
1	0
2	0
3	1
⋮	⋮

La mesure n°1 (x_1) est la réalisation d'une variable aléatoire X_1, la mesure n°2 (x_2) est la réalisation d'une variable aléatoire X_2, la mesure n°3 (x_3) est la réalisation d'une variable aléatoire X_3 etc.

Pour cet exemple, quel est la loi de chaque variable ?

Modèle statistique paramétrique

La réponse d'un consommateur choisi au hasard est vue comme la valeur d'une variable aléatoire de loi Bernoulli.

Loi Bernoulli $\mathcal{B}(p)$, avec $p \in [0,1]$

On dit que la variable X suit une loi Bernoulli si

$$\mathbb{P}(X=1) = p \quad \text{et} \quad \mathbb{P}(X=0) = 1-p,$$

où p est un nombre réel compris entre 0 et 1 appelé paramètre de la loi. On note cette loi $\mathcal{B}(p)$.

- Dans notre exemple, le paramètre p correspond à la proportion de consommateurs dans le grand panel en faveur de B.
- Si le panel de consommateurs est vraiment très grand, les enquêteurs n'ont pas le temps d'interroger tout le monde et ne peuvent accéder à la vraie valeur de p.
- Il est important de remarquer que ici p est inconnu.
- On a besoin de la théorie pour malgré tout, pouvoir dire des choses avec un degré de confiance raisonnable sur p.

Estimation d'une proportion

- Exemple : $n = 100$ consommateurs sondés, 58 votes en faveur de A et 42 en faveur de B.
  - A partir de notre échantillon on peut dire que p est estimé à

$$42/100 = 0.42$$

  - Peut-on conclure au risque $\alpha = 5\%$ que le packaging A est préférable au packaging B ? On a besoin d'appliquer un certain test d'hypothèses.
  - Quelle confiance accorder à cette estimation ? On fait donc appel à la théorie d'estimation d'une proportion par intervalle de confiance (on verra plus tard).

Test d'hypothèses

Se poser la question si le packaging A est préférable au packaging B revient à se demander si $p = 1/2$ ou $p < 1/2$.

Hypothèses : H_0 et H_1

- Un test oppose deux hypothèses : l'hypothèse nulle H_0 et l'hypothèse alternative H_1.
- A l'issue du test, on va décider de rejeter ou pas H_0. Quelle que soit la décision on peut se tromper.

Valeur critique ou p-valeur

- L'usage ancien des tables statistiques donnant les quantiles des différentes lois usuelles n'a plus lieu d'être avec la pratique d'un logiciel statistique. En effet, ceux-ci fournissent directement la probabilité critique ou p-valeur (en anglais p-value) associée à un test donné.
- Il suffit de comparer la p-valeur fournit avec le seuil ou niveau de test α fixé.

Moyenne empirique

Definition

Soit $X_1, \ldots, X_n$ un échantillon d'une loi X.

On appelle moyenne empirique la variable aléatoire $\bar{X}_n$ définie par :

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

- Estimateur ponctuelle de p : $\bar{X}_n$.

$$\widehat{p} = \bar{x}_{100} = 42/100 = 0.42$$

- Notez bien la différence entre $\bar{X}_{100}$ et $\bar{x}_{100}$!

Remarque (une convention utile pour la suite) : on mettra un petit chapeau à toutes les quantités qui ne dépendent que des observations.

- Avec les valeurs observées $x_1, \ldots, x_{100}$ de 100 consommateurs on a teste $\mathrm{H}_0 : p = 0.5$ contre $\mathrm{H}_1 : p < 0.5$ au seuil $\alpha = 5\%$.

Critère de décision basée sur la p-valeur:

On rejette l'hypothèse nulle H_0 si p-valeur $\leqslant \alpha$.

- Résultat du test (avec la loi exacte de $\mathcal{B}(100, 0.42)$) obtenu avec la fonction **binom.test** de la libraire(binom) du logiciel R

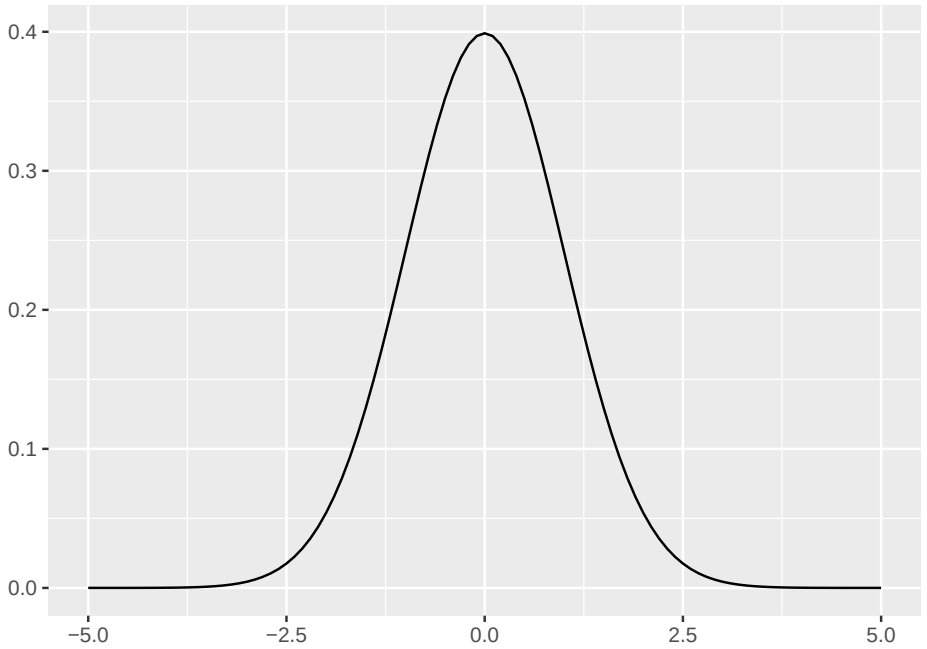
```
>binom.test(42,100,p=0.5,alternative='less',conf.level=0.95)

number of successes = 42, number of trials = 100,
p-value = 0.06661
alternative hypothesis: true probability of success is
less than 0.5
95 percent confidence interval:
0.0000000 0.5071585
sample estimates:
probability of success
0.42
```

- Dans cette expérience, statistiquement on ne rejette pas H_0.

Le théorème centrale limite (TCL)

TCL

Si $X_1, \ldots, X_n$ est une suite de variables aléatoires réelles indépendantes et identiquement distribuées, alors la loi de probabilité de la quantité $\frac{(\bar{X}_n - \mathbb{E}(\bar{X}_n))}{\sqrt{Var(\bar{X}_n)}}$ se rapproche de la loi normale $\mathcal{N}(0,1)$ lorsque n es suffisamment grand.



Application du TLC :

Si l'on prend $X_i \sim \mathcal{B}(p)$, $i = 1, \ldots, n$ on retrouve qu'une Binomiale $\mathcal{B}(n, p)$ approche à une normale (lorsque n est grand).

$\frac{(\bar{X}_n - p)}{\sqrt{\frac{p(1-p)}{n}}}$ se rapproche de la loi normale $\mathcal{N}(0, 1)$

Estimation d'une proportion par intervalle de confiance

Intervalle de confiance asymptotique de niveau $(1-\alpha)\%$

Pour une valeur α fixée,

$$\left[\bar{X}_n - q_Z(1-\frac{\alpha}{2})\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + q_Z(1-\frac{\alpha}{2})\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right]$$

où Z est une v.a. suivant une loi normale centrée réduite.

Estimation d'une proportion par intervalle de confiance

- Dans notre exemple, pour $\alpha = 0.05$, on obtient
  - quantile d'ordre 0.975 de la loi normale centrée réduite vaut 1.96
  - $\sqrt{\frac{0.42*(1-0.42)}{100}} = 0.049$
  - Réalisation de l'intervalle

$$[0.42 - 1.96 \times 0.049, 0.42 + 1.96 \times 0.049] \approx [0.42 - 0.10, 0.42 + 0.10]$$
$$= [0.32, 0.52]$$

- Avec les valeurs observées $x_1, \ldots, x_{100}$ de 100 consommateurs on a construit un intervalle de confiance de p. On peut (seulement) dire avec une grande confiance que le vrai paramètre p , représentant le taux de préférence de B sur le panel, est estimé à $42\% \pm 10\%$.

- Avec les valeurs observées $x_1, \ldots, x_{100}$ de 100 consommateurs on a testé $\mathrm{H}_0 : p = 0.5$ contre $\mathrm{H}_1 : p < 0.5$ au seuil $\alpha = 5\%$. Dans cette expérience, le statisticien dit, on ne rejette pas l'hypothèse $\mathrm{H}_0 : p = 0.5$.
- Voir avec le directeur du marketing ! Tout va dépendre si le directeur de marketing est prudent et ne vaut pas se risquer au changement (à cause du coût au changement du nouveau packaging) ou si c'est le directeur de marketing qui propose ce changement en essayant de faire passer le nouveau packaging.

Bibliographie

- Notes de cours L3 de Gilles Stolz (Éléments de statistique pour citoyens d'aujourd'hui et managers de demain)