

Démarche statistique

Premiers pas avec

David Causeur

Agrocampus Ouest

IRMAR CNRS UMR 6625

<http://math.agrocampus-ouest.fr/infoglueDeliverLive/membres/david.causeur>

Objectifs d'apprentissage

A la fin de ce module, vous serez capables :

- d'aborder les problèmes les plus courants d'analyse de données ;
- d'argumenter le choix de procédures d'analyse ;
- d'évaluer les performances d'une règle de décision statistique ;
- de mettre en œuvre une démarche d'analyse de données avec R.



Evaluation (voir aussi document distribué)

- Deux courts contrôles continus des connaissances - 50%
- Etude de cas - 50%

Pourquoi ?

- **R** couvre un large panel de domaines d'application ;
- **R** est libre ;
- Connaître **R** est très souvent exigé dans les offres d'emploi ;

Ressources

- Télécharger 
- Télécharger  Studio

Plan

1 Effet à l'échelle d'une population

2 Décider à partir de données

3 Effet 'groupe'

- Comparaison de groupes

- Analyse de variance à un facteur

- Estimation des paramètres d'effet

- Test de Fisher

- Le cas particulier de la comparaison de 2 groupes

- Décrire un effet groupe

- Test avec des données appariées

4 Effet linéaire

Effet à l'échelle d'une population

france
inter

Info

Culture

Humour

Musique

Plus ▾

Programmes

Replay

Le direct
and bien vous fasse

Publicité

Accueil > Sciences > Groupe sanguin et coronavirus, un hasard génétique

Groupe sanguin et coronavirus, un hasard génétique

par **Sophie Bécherel** publié le 21 mars 2020 à 8h06

D'après une étude chinoise, les personnes de groupe sanguin O sont mieux immunisées contre le coronavirus que les autres groupes. Elles ont un risque d'infection 33% moindre. A contrario, les personnes de groupe A ont 20% de risque supplémentaire d'être infectées. Cette inégalité s'explique par l'action des anticorps.

Publicité

La newsletter d'Inter

Recevez du lundi au vendredi à 12h
une sélection toute fraîche à lire ou à
écouter.

Effet à l'échelle d'une population

La principale problématique en analyse de données :

'y-a-t'il un effet de ceci sur cela ?'

- le groupe sanguin influence-t'il la susceptibilité à la COVID 19 ?
- L'augmentation de la dose d'un médicament modifie-t'elle la tension artérielle ?
- La concentration en azote dans le sol a-t'elle un impact sur le rendement des cultures ?
- Le genre du consommateur est-il un déterminant de sa propension à acheter un produit ?

Effet à l'échelle d'une population

La principale problématique en analyse de données :

'y-a-t'il un effet de ceci sur cela ?'

Exemple illustratif : Le pourcentage de viande maigre (**LMP** pour Lean Meat Percentage) d'un porc détermine sa valeur commerciale. Il est mesuré *indirectement* par des épaisseurs de tissus gras et maigres.

- Dans quelle mesure les épaisseurs de tissus prédisent-elles le LMP ?
- Les épaisseurs de tissus dépendent-elles du type génétique ?

Effet à l'échelle d'une population

On identifie **deux types de variables** :

- la **variable réponse** Y ,
- la **variable explicative** X .

... les variations de Y pouvant dépendre des variations de X

' X a un effet sur Y '

peut être formulé mathématiquement par

'la distribution de Y parmi les individus ayant la même valeur x de X dépend de la valeur x '.

Effet à l'échelle d'une population

On identifie **deux types de variables** :

- la **variable réponse** Y ,
- la **variable explicative** X .

... les variations de Y pouvant dépendre des variations de X

'Le type génétique a un effet sur l'épaisseur de gras'

peut être formulé mathématiquement par

'les distributions de l'épaisseur de gras diffèrent selon le type génétique'.

Données

Données : observations $(x_i, y_i)_{i=1, \dots, n}$ de X et Y , avec $n \geq 2$

Échantillon : ensemble des n individus pour lesquels on dispose des observations $(x_i, y_i)_{i=1, \dots, n}$, où $n \geq 2$ est la **taille d'échantillon**.

► Importation de données dans R

Inférence statistique

Statistique inférentielle : méthodologie statistique visant à décider pour une population à partir d'un échantillon.

... suppose que l'échantillon est représentatif d'une population plus large.

Description statistique

L'**analyse exploratoire des données** vise à décrire un effet par la synthèse orientée de données :

- par des représentations graphiques
- par des **résumés statistiques**

... permet de construire des hypothèses de travail sur la nature d'un effet.

Nature des variables

La distribution d'une variable aléatoire dépend de sa nature,

- soit **quantitative/numérique** : mesurée sur une échelle continue ou discrète (LMP, épaisseurs tissulaires, ...)
- soit **qualitative/catégorielle** : qui définit des sous-groupes de la population (sexe, type génétique, ...)

... parfois ambigu ... par exemple pour le nombre d'enfants.

Plan

1 Effet à l'échelle d'une population

2 Décider à partir de données

3 Effet 'groupe'

- Comparaison de groupes

- Analyse de variance à un facteur

- Estimation des paramètres d'effet

- Test de Fisher

- Le cas particulier de la comparaison de 2 groupes

- Décrire un effet groupe

- Test avec des données appariées

4 Effet linéaire

Effet significatif

L'effet de X sur Y est **significatif** si l'analyse des données conclut à l'existence de cet effet à l'échelle de la population.

Erreurs de décision

Deux types d'erreurs inhérentes à la prise de décision :

- **Type I** : déclarer un effet significatif alors qu'il n'existe pas à l'échelle de la population.
- **Type II** : déclarer un effet non-significatif alors qu'il existe à l'échelle de la population.

On dit aussi :

- **Type I** : faux positif.
- **Type II** : faux négatif.

Erreurs de décision

Les deux types d'erreur sont antagonistes :

- **Règle de décision libérale** : l'effet est déclaré significatif même s'il est peu évident
... le risque de l'erreur de type I est alors grand et le risque de l'erreur de type II est faible.
- **Règle de décision conservative** : l'effet n'est déclaré significatif que s'il est très évident
... le risque de l'erreur de type I est alors faible et le risque de l'erreur de type II est grand.

Règle de décision conservative

Objectif principal : faible risque de l'erreur de type I

Deux hypothèses asymétriques :

$$\begin{cases} H_0 : \text{l'effet n'existe pas à l'échelle de la population} \\ H_1 : \text{l'effet existe à l'échelle de la population} \end{cases}$$

H_0 , **hypothèse nulle**, n'est rejetée que s'il est évident que les observations ne sont pas en cohérence avec elle.

The null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation (R.A. Fisher, 'The Design of Experiments', 1935).

Règle de décision conservative

Test de H_0 : règle de rejet ou non de H_0 à partir des données.

Trois clés pour un test statistique

- **Statistique de test** T . Elle mesure l'effet : plus l'effet est évident, plus la valeur de T est grande.
- **Distribution** de T sous H_0 .
Supposez que $\mathbb{P}_{H_0}(T \geq 2) \leq 0.05$, alors observer que $T = 3$ doit conduire à rejeter H_0 .
- **p-value** du test : la probabilité, calculée sous l'hypothèse nulle, que la statistique de test soit plus grande que la valeur observée de T .

Si la **p-value** est plus petite qu'un seuil α (le plus souvent $\alpha = 0.05$), alors l'effet est significatif **au seuil** α .

α : seuil ou niveau du test.

Trois clés pour un test statistique

Rendez-vous en terre inconnue : vous prenez un vol pour une destination inconnue, les yeux bandés.

- Votre hypothèse (**H_0**) : la destination finale est la Bretagne ;
- A votre arrivée, vous évaluez la température extérieure (**statistique de test**) à 40° ;
- Cette observation, $T = 40^\circ$, est-elle en cohérence avec votre hypothèse (**test de l'hypothèse nulle**) ?
- Vous savez aussi que la probabilité que la température excède 40° en Bretagne est très faible (**distribution de T sous l'hypothèse nulle**).
- Votre conclusion : **rejet** de l'hypothèse nulle.

Plan

1 Effet à l'échelle d'une population

2 Décider à partir de données

3 Effet 'groupe'

- Comparaison de groupes

- Analyse de variance à un facteur

- Estimation des paramètres d'effet

- Test de Fisher

- Le cas particulier de la comparaison de 2 groupes

- Décrire un effet groupe

- Test avec des données appariées

4 Effet linéaire

Description de différences entre groupes

'**X a un effet sur Y**'

peut être reformulé par

'la **distribution de Y parmi les individus d'un même groupe** diffère selon le groupe'.

► Description statistique d'un effet groupe dans \mathbb{R}

Description de différences entre groupes

Moyenne et **médiane** sont des indicateurs de la position d'une série x_1, \dots, x_n sur l'axe réel.

La **moyenne** de (x_1, \dots, x_n) est définie par :

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}.$$

La **médiane** est définie de manière moins explicite :

médiane(x) $\in [x_{(n/2)}; x_{(n/2+1)}[$ si n est pair,

médiane(x) = $x_{((n+1)/2)}$ si n est impair,

où $(x_{(1)}, \dots, x_{(n)})$ est la série triée par ordre croissant.

► Moyenne et médiane dans \mathbb{R}

Description de différences entre groupes

Le **quantile** d'ordre α de (x_1, \dots, x_n) est défini par :

$$q_\alpha(x) \in [x_{(i)}; x_{(i+1)}[, \text{ si } \alpha < 0.5$$

$$q_\alpha(x) \in]x_{(i-1)}; x_{(i)}], \text{ si } \alpha > 0.5$$

où i est le plus petit entier entre 1 et n tel que plus de $100\alpha\%$ d'éléments de la série sont inférieurs à $x_{(i)}$

► Quantiles dans \mathbb{R}

Description de différences entre groupes

L'écart-type est défini par :

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}.$$

Plus s_x est grand, plus la série est dispersée autour de \bar{x} .

Variance s_x^2 : *presque* la moyenne des carrés des écarts $x_i - \bar{x}$.

Pourquoi divise-t'on par $n - 1$ et pas par n ?

... $x_i - \bar{x}$ contient $(n - 1)$ écarts linéairement indépendants.

On dit que $(x_1 - \bar{x}, \dots, x_n - \bar{x})$ a $n - 1$ degrés de liberté.

Description de différences entre groupes

► Boîte de dispersion dans R

Une **boîte de dispersion** résume la répartition de valeurs numériques par les éléments graphiques suivants :

- **la boîte**, dont la limite inférieure est $q_{0.25}$ et la limite supérieure est $q_{0.75}$. Un segment tracé dans la boîte localise la médiane ;
- **la moustache inférieure**, qui s'étend jusqu'à la plus petite valeur moins éloignée que $1.5 \times \text{IQR}$ de la médiane ;
- **la moustache supérieure**, qui s'étend jusqu'à la plus grande valeur moins éloignée que $1.5 \times \text{IQR}$ de la médiane ;
- **des points isolés** pour chaque valeur en dehors des limites des moustaches.

Description de différences entre groupes

On en déduit que les types génétiques se différencient essentiellement par leur épaisseur moyenne de gras dorsal.

... peut-être $P_0 > P_{25} \approx P_{50}$?

Plan

1 Effet à l'échelle d'une population

2 Décider à partir de données

3 Effet 'groupe'

- Comparaison de groupes

- Analyse de variance à un facteur

- Estimation des paramètres d'effet

- Test de Fisher

- Le cas particulier de la comparaison de 2 groupes

- Décrire un effet groupe

- Test avec des données appariées

4 Effet linéaire