Science des données I : module 12

Critique statistique

Philippe Grosjean & Guyliann Engels

Université de Mons, Belgique
Laboratoire d'Écologie numérique

https://wp.sciviews.org
sdd@sciviews.org

UMONS

Que faire face à un nouveau jeu de données ?

UMONS

Que faire face à une nouveau jeu de données ?

Face à un nouveau jeu de données, il faut **se poser d'abord les questions suivantes** :

- A **quelle(s) question(s)** le biologiste cherche-t-il à répondre en ayant réalisé ces mesures ?
- Quelle est la **variable réponse (= dépendante)** et quelles sont les **variables explicatives (= indépendantes)** dans ce jeu de données ?
- Le jeu de données est-il **représentatif** de la population (échantillonnage aléatoire et observations indépendantes) ?
- Quelle **information** ce jeu de données contient-t-il ? (exploration, statistiques descriptives, représentations graphiques)
- Comment la question biologique se traduit-elle en **formulation statistique**, et par conséquent, **quelle méthode doit être appliquée** pour répondre à la question ?

UMONS

Que faire face à une nouveau jeu de données ?

Face à un nouveau jeu de données, il faut **se poser d'abord les questions suivantes** :

- A **quelle(s) question(s)** le biologiste cherche-t-il à répondre en ayant réalisé ces mesures ?
- Quelle est la **variable réponse (= dépendante)** et quelles sont les **variables explicatives (= indépendantes)** dans ce jeu de données ?
- Le jeu de données est-il **représentatif** de la population (échantillonnage aléatoire et observations indépendantes) ?
- Quelle **information** ce jeu de données contient-t-il ? (exploration, statistiques descriptives, représentations graphiques)
- Comment la question biologique se traduit-elle en **formulation statistique**, et par conséquent, **quelle méthode doit être appliquée** pour répondre à la question ?

UMONS

Que faire face à une nouveau jeu de données ?

Face à un nouveau jeu de données, il faut **se poser d'abord les questions suivantes** :

- A **quelle(s) question(s)** le biologiste cherche-t-il à répondre en ayant réalisé ces mesures ?
- Quelle est la **variable réponse (= dépendante)** et quelles sont les **variables explicatives (= indépendantes)** dans ce jeu de données ?
- Le jeu de données est-il **représentatif** de la population (échantillonnage aléatoire et observations indépendantes) ?
- Quelle **information** ce jeu de données contient-t-il ? (exploration, statistiques descriptives, représentations graphiques)
- Comment la question biologique se traduit-elle en **formulation statistique**, et par conséquent, **quelle méthode doit être appliquée** pour répondre à la question ?

UMONS

Que faire face à une nouveau jeu de données ?

Face à un nouveau jeu de données, il faut **se poser d'abord les questions suivantes** :

- A **quelle(s) question(s)** le biologiste cherche-t-il à répondre en ayant réalisé ces mesures ?
- Quelle est la **variable réponse (= dépendante)** et quelles sont les **variables explicatives (= indépendantes)** dans ce jeu de données ?
- Le jeu de données est-il **représentatif** de la population (échantillonnage aléatoire et observations indépendantes) ?
- Quelle **information** ce jeu de données contient-t-il ? (exploration, statistiques descriptives, représentations graphiques)
- Comment la question biologique se traduit-elle en **formulation statistique**, et par conséquent, **quelle méthode doit être appliquée** pour répondre à la question ?

UMONS

Que faire face à une nouveau jeu de données ?

Face à un nouveau jeu de données, il faut **se poser d'abord les questions suivantes** :

- A **quelle(s) question(s)** le biologiste cherche-t-il à répondre en ayant réalisé ces mesures ?
- Quelle est la **variable réponse (= dépendante)** et quelles sont les **variables explicatives (= indépendantes)** dans ce jeu de données ?
- Le jeu de données est-il **représentatif** de la population (échantillonnage aléatoire et observations indépendantes) ?
- Quelle **information** ce jeu de données contient-t-il ? (exploration, statistiques descriptives, représentations graphiques)
- Comment la question biologique se traduit-elle en **formulation statistique**, et par conséquent, **quelle méthode doit être appliquée** pour répondre à la question ?

UMONS

Que faire face à une méthode statistique inconnue ?

UMONS

Astuces face à une méthode statisitque inconnue

1. Rechercher un **“tutorial”** avec exemple **réalisé dans le logiciel** (aide en ligne, manuel de R, etc…).
2. L’appliquer aussi à des **jeux de données qu’on connaît bien**, pour voir comment elle se comporte.
3. Travailler avec des **jeux de données artificiels** – que se passe-t-il si… ?
4. Rechercher dans la **bibliographie** des cas similaires où la méthode a déjà été appliquée (**Attention ! Dans les revues scientifiques en biologie, il y a parfois des erreurs graves…** rester critiques) !
5. Si la nouvelle méthode s’apparente à une ou plusieurs autres connues, **comparer**.
6. Si possible : **décortiquer le calcul** pour comprendre comment elle fonctionne.

UMONS

Astuces face à une méthode statisitque inconnue

1. Rechercher un **“tutorial”** avec exemple **réalisé dans le logiciel** (aide en ligne, manuel de R, etc…).
2. L’appliquer aussi à des **jeux de données qu’on connaît bien**, pour voir comment elle se comporte.
3. Travailler avec des **jeux de données artificiels** – que se passe-t-il si… ?
4. Rechercher dans la **bibliographie** des cas similaires où la méthode a déjà été appliquée (**Attention ! Dans les revues scientifiques en biologie, il y a parfois des erreurs graves…** rester critiques) !
5. Si la nouvelle méthode s’apparente à une ou plusieurs autres connues, **comparer**.
6. Si possible : **décortiquer le calcul** pour comprendre comment elle fonctionne.

UMONS

Astuces face à une méthode statisitque inconnue

1. Rechercher un **“tutorial”** avec exemple **réalisé dans le logiciel** (aide en ligne, manuel de R, etc…).
2. L’appliquer aussi à des **jeux de données qu’on connaît bien**, pour voir comment elle se comporte.
3. Travailler avec des **jeux de données artificiels** – que se passe-t-il si… ?
4. Rechercher dans la **bibliographie** des cas similaires où la méthode a déjà été appliquée (**Attention ! Dans les revues scientifiques en biologie, il y a parfois des erreurs graves…** rester critiques) !
5. Si la nouvelle méthode s’apparente à une ou plusieurs autres connues, **comparer**.
6. Si possible : **décortiquer le calcul** pour comprendre comment elle fonctionne.

UMONS

Astuces face à une méthode statisitque inconnue

1. Rechercher un **“tutorial”** avec exemple **réalisé dans le logiciel** (aide en ligne, manuel de R, etc…).
2. L’appliquer aussi à des **jeux de données qu’on connaît bien**, pour voir comment elle se comporte.
3. Travailler avec des **jeux de données artificiels** – que se passe-t-il si… ?
4. Rechercher dans la **bibliographie** des cas similaires où la méthode a déjà été appliquée (**Attention ! Dans les revues scientifiques en biologie, il y a parfois des erreurs graves…** rester critiques) !
5. Si la nouvelle méthode s’apparente à une ou plusieurs autres connues, **comparer**.
6. Si possible : **décortiquer le calcul** pour comprendre comment elle fonctionne.

UMONS

Astuces face à une méthode statisitque inconnue

1. Rechercher un **“tutorial”** avec exemple **réalisé dans le logiciel** (aide en ligne, manuel de R, etc…).
2. L’appliquer aussi à des **jeux de données qu’on connaît bien**, pour voir comment elle se comporte.
3. Travailler avec des **jeux de données artificiels** – que se passe-t-il si… ?
4. Rechercher dans la **bibliographie** des cas similaires où la méthode a déjà été appliquée (**Attention ! Dans les revues scientifiques en biologie, il y a parfois des erreurs graves…** rester critiques) !
5. Si la nouvelle méthode s’apparente à une ou plusieurs autres connues, **comparer**.
6. Si possible : **décortiquer le calcul** pour comprendre comment elle fonctionne.

UMONS

Astuces face à une méthode statisitque inconnue

1. Rechercher un **“tutorial”** avec exemple **réalisé dans le logiciel** (aide en ligne, manuel de R, etc…).
2. L’appliquer aussi à des **jeux de données qu’on connaît bien**, pour voir comment elle se comporte.
3. Travailler avec des **jeux de données artificiels** – que se passe-t-il si… ?
4. Rechercher dans la **bibliographie** des cas similaires où la méthode a déjà été appliquée (**Attention ! Dans les revues scientifiques en biologie, il y a parfois des erreurs graves…** rester critiques) !
5. Si la nouvelle méthode s’apparente à une ou plusieurs autres connues, **comparer**.
6. Si possible : **décortiquer le calcul** pour comprendre comment elle fonctionne.

UMONS

Découvrir une méthode statistique

1. A quel type de **question** répond-t-elle ?
2. **Quel(s) type(s) de variable(s)** faut-il ?
3. Quelles sont les **contraintes** qu'il faut rencontrer (et donc vérifier !) pour pouvoir appliquer cette méthode (**normalité, homoscédasticité, etc...**) ?
4. Existe t'il des **variantes (bi- / unilatéral, apparié / non apparié)** ?
5. Comment **interpréter** le résultat ?

UMONS

Découvrir une méthode statistique

1. A quel type de **question** répond-t-elle ?
2. **Quel(s) type(s) de variable(s)** faut-il ?
3. Quelles sont les **contraintes** qu'il faut rencontrer (et donc vérifier !) pour pouvoir appliquer cette méthode (**normalité, homoscédasticité, etc...**) ?
4. Existe t'il des **variantes (bi- / unilatéral, apparié / non apparié)** ?
5. Comment **interpréter** le résultat ?

UMONS

Découvrir une méthode statistique

1. A quel type de **question** répond-t-elle ?
2. **Quel(s) type(s) de variable(s)** faut-il ?
3. Quelles sont les **contraintes** qu'il faut rencontrer (et donc vérifier !) pour pouvoir appliquer cette méthode (**normalité, homoscédasticité, etc...**) ?
4. Existe t'il des **variantes** (bi- / unilatéral, apparié / non apparié) ?
5. Comment **interpréter** le résultat ?

UMONS

Découvrir une méthode statistique

1. A quel type de **question** répond-t-elle ?
2. **Quel(s) type(s) de variable(s)** faut-il ?
3. Quelles sont les **contraintes** qu'il faut rencontrer (et donc vérifier !) pour pouvoir appliquer cette méthode (**normalité, homoscédasticité, etc...**) ?
4. Existe t'il des **variantes (bi- / unilatéral, apparié / non apparié)** ?
5. Comment **interpréter** le résultat ?

UMONS

Découvrir une méthode statistique

1. A quel type de **question** répond-t-elle ?
2. **Quel(s) type(s) de variable(s)** faut-il ?
3. Quelles sont les **contraintes** qu'il faut rencontrer (et donc vérifier !) pour pouvoir appliquer cette méthode (**normalité, homoscédasticité, etc...**) ?
4. Existe t'il des **variantes (bi- / unilatéral, apparié / non apparié)** ?
5. Comment **interpréter** le résultat ?

UMONS

Critique statistique - mise en situation

UMONS

Doit-on croire les études statistiques ?

Oui, mais toujours avec un regard critique *(Bennett, Briggs & Triola, Statistical reasoning for everyday life, 2nd ed.)* :

1. Bien **identifier la question posée** et la population étudiée
2. **Vérifier les sources** (est-il possible qu'il y ait des biais ?)
3. Analyser la **méthode d'échantillonnage** utilisée
4. Rechercher des problèmes éventuels dans la **définition ou la mesure des variables** (vérifier les ordres de grandeurs)
5. Être attentif aux **variables confondantes** qui peuvent invalider les conclusions (imaginer des scénarios alternatifs)
6. Comment l'étude a-t-elle été **mise en place et réalisée** ?
7. Vérifier que les **graphiques** proposés représentent les données de manière objective
8. Vérifier que les **conclusions finales** répondent bien à la question posée

UMONS

Doit-on croire les études statistiques ?

Oui, mais toujours avec un regard critique *(Bennett, Briggs & Triola, Statistical reasoning for everyday life, 2nd ed.)* :

1. Bien **identifier la question posée** et la population étudiée
2. **Vérifier les sources** (est-il possible qu'il y ait des biais ?)
3. Analyser la **méthode d'échantillonnage** utilisée
4. Rechercher des problèmes éventuels dans la **définition ou la mesure des variables** (vérifier les ordres de grandeurs)
5. Être attentif aux **variables confondantes** qui peuvent invalider les conclusions (imaginer des scénarios alternatifs)
6. Comment l'étude a-t-elle été **mise en place et réalisée** ?
7. Vérifier que les **graphiques** proposés représentent les données de manière objective
8. Vérifier que les **conclusions finales** répondent bien à la question posée

UMONS

Doit-on croire les études statistiques ?

Oui, mais toujours avec un regard critique *(Bennett, Briggs & Triola, Statistical reasoning for everyday life, 2nd ed.)* :

1. Bien **identifier la question posée** et la population étudiée
2. **Vérifier les sources** (est-il possible qu'il y ait des biais ?)
3. Analyser la **méthode d'échantillonnage** utilisée
4. Rechercher des problèmes éventuels dans la **définition ou la mesure des variables** (vérifier les ordres de grandeurs)
5. Être attentif aux **variables confondantes** qui peuvent invalider les conclusions (imaginer des scénarios alternatifs)
6. Comment l'étude a-t-elle été **mise en place et réalisée** ?
7. Vérifier que les **graphiques** proposés représentent les données de manière objective
8. Vérifier que les **conclusions finales** répondent bien à la question posée

UMONS

Doit-on croire les études statistiques ?

Oui, mais toujours avec un regard critique *(Bennett, Briggs & Triola, Statistical reasoning for everyday life, 2nd ed.)* :

1. Bien **identifier la question posée** et la population étudiée
2. **Vérifier les sources** (est-il possible qu'il y ait des biais ?)
3. Analyser la **méthode d'échantillonnage** utilisée
4. Rechercher des problèmes éventuels dans la **définition ou la mesure des variables** (vérifier les ordres de grandeurs)
5. Être attentif aux **variables confondantes** qui peuvent invalider les conclusions (imaginer des scénarios alternatifs)
6. Comment l'étude a-t-elle été **mise en place et réalisée** ?
7. Vérifier que les **graphiques** proposés représentent les données de manière objective
8. Vérifier que les **conclusions finales** répondent bien à la question posée

UMONS

Doit-on croire les études statistiques ?

Oui, mais toujours avec un regard critique *(Bennett, Briggs & Triola, Statistical reasoning for everyday life, 2nd ed.)* :

1. Bien **identifier la question posée** et la population étudiée
2. **Vérifier les sources** (est-il possible qu'il y ait des biais ?)
3. Analyser la **méthode d'échantillonnage** utilisée
4. Rechercher des problèmes éventuels dans la **définition ou la mesure des variables** (vérifier les ordres de grandeurs)
5. Être attentif aux **variables confondantes** qui peuvent invalider les conclusions (imaginer des scénarios alternatifs)
6. Comment l'étude a-t-elle été **mise en place et réalisée** ?
7. Vérifier que les **graphiques** proposés représentent les données de manière objective
8. Vérifier que les **conclusions finales** répondent bien à la question posée

UMONS

Doit-on croire les études statistiques ?

Oui, mais toujours avec un regard critique *(Bennett, Briggs & Triola, Statistical reasoning for everyday life, 2nd ed.)* :

1. Bien **identifier la question posée** et la population étudiée
2. **Vérifier les sources** (est-il possible qu'il y ait des biais ?)
3. Analyser la **méthode d'échantillonnage** utilisée
4. Rechercher des problèmes éventuels dans la **définition ou la mesure des variables** (vérifier les ordres de grandeurs)
5. Être attentif aux **variables confondantes** qui peuvent invalider les conclusions (imaginer des scénarios alternatifs)
6. Comment l'étude a-t-elle été **mise en place et réalisée** ?
7. Vérifier que les **graphiques** proposés représentent les données de manière objective
8. Vérifier que les **conclusions finales** répondent bien à la question posée

UMONS

Doit-on croire les études statistiques ?

Oui, mais toujours avec un regard critique *(Bennett, Briggs & Triola, Statistical reasoning for everyday life, 2nd ed.)* :

1. Bien **identifier la question posée** et la population étudiée
2. **Vérifier les sources** (est-il possible qu'il y ait des biais ?)
3. Analyser la **méthode d'échantillonnage** utilisée
4. Rechercher des problèmes éventuels dans la **définition ou la mesure des variables** (vérifier les ordres de grandeurs)
5. Être attentif aux **variables confondantes** qui peuvent invalider les conclusions (imaginer des scénarios alternatifs)
6. Comment l'étude a-t-elle été **mise en place et réalisée** ?
7. Vérifier que les **graphiques** proposés représentent les données de manière objective
8. Vérifier que les **conclusions finales** répondent bien à la question posée

UMONS

Doit-on croire les études statistiques ?

Oui, mais toujours avec un regard critique *(Bennett, Briggs & Triola, Statistical reasoning for everyday life, 2nd ed.)* :

1. Bien **identifier la question posée** et la population étudiée
2. **Vérifier les sources** (est-il possible qu'il y ait des biais ?)
3. Analyser la **méthode d'échantillonnage** utilisée
4. Rechercher des problèmes éventuels dans la **définition ou la mesure des variables** (vérifier les ordres de grandeurs)
5. Être attentif aux **variables confondantes** qui peuvent invalider les conclusions (imaginer des scénarios alternatifs)
6. Comment l'étude a-t-elle été **mise en place et réalisée** ?
7. Vérifier que les **graphiques** proposés représentent les données de manière objective
8. Vérifier que les **conclusions finales** répondent bien à la question posée

UMONS

Critique statistique (1)

Un chercheur compile les statistiques de longévité de diverses professions. Pour ce faire, il encode les données des certificats de décès (nom, âge au moment du décès et profession). Il calcule ensuite l'âge moyen de décès par profession. Il constate que la valeur minimale est observée **chez les étudiants**, avec une valeur moyenne de seulement 20,7 ans (Wainer, Palmer & Bradlow, A selection of selection anomalies, *Chance*, vol. 11, n°2).

La « profession » d'étudiant est-elle réellement plus dangereuse que celle de policier, chauffeur de taxi, ou cascadeur ? Expliquez...

UMONS

Critique statistique (2)

Un biologiste étudie une chauve-souris insectivore naine. Il trouve dans la littérature que la biomasse totale de cette chauve-souris varie de 0,23 à 1,95 kg/ha dans les forêts recensées. Afin de calculer l'abondance de ces populations de chauve-souris, il détermine le poids moyen d'un individu comme étant (moyenne $\pm$ écart type) 55 ± 13 mg ($n = 45$). Il utilise ces données pour comparer les populations de chauve-souris aux autre animaux présents dans cette forêt. Il en conclu, enfin, que la population de chauve-souris dans ces forêts est très nettement supérieure à celle des oiseaux et équivalente à celle des insectes dans la région étudiée. Ce résultat est inattendu et permet de considérer cette chauve-souris comme espèce clé dans la chaîne trophique, alors que son effet a toujours été négligé auparavant, tant elle est discrète et passe inaperçu la plupart du temps.

Vous travaillez aussi sur les chaînes trophiques de ces mêmes forêts. Comment réagissez-vous à la lecture de ce rapport ? Que faites-vous ensuite ?

UMONS

Critique statistique (3)

Le magazine “Men’s Health” a publié des statistiques qui décrivent l’“homme moyen”. Celui-ci a 34,4 ans, pèse 79,4kg, mesure 177,8cm, dors 6,9 heures chaque nuit, bois 3,3 tasses de café par jour et consomme 1,2 boisson alcoolique quotidiennement.

Sachant que toutes les distributions sont unimodales, donc que les valeurs moyennes correspondent toutes à des observations effectivement mesurées en grand nombres (identiques ou très proches) sur des hommes réels, ce portrait robot de l’“homme moyen” décrit-il effectivement un grand nombre d’individus réellement existants ? Justifiez. Qu’en serait-il de l’“homme médian” ?

UMONS

Critique statistiques (4)

L'espérance de vie est une donnée statistique qui permet de connaître la durée de vie moyenne qu'on peut espérer atteindre à un moment donné pour une nation donnée. Cette statistique est calculée et publiée par de nombreux organismes, incluant l'OMS. Les statistiques indiquent que l'espérance de vie des hommes dans nos pays est de 75,5 ans, et des femmes de 83,5 ans.

Calculez le temps que vous pouvez espérer encore vivre en fonction de votre âge. Que pensez-vous de ce calcul ?

UMONS

Critique statistique (5)

Un scientifique mesure la stabilité de la membrane lysosomale (indice de stress des cellules utilisé en écotoxicologie : on sait que les polluants étudiés tendent à déstabiliser la membrane des lysosomes) chez la moule *Mytilus edulis* en Mer du Nord. Deux régions sont comparées : la pleine mer (A), et l'embouchure de l'Escault dans sa partie considérée comme la plus polluée (B). Cinq moules sont prélevées aléatoirement sur les deux sites, et dix mesures sont réalisées sur chaque individu. Le scientifique conclu à une stabilité lysosomale significative plus faible au seuil alpha de 5% dans le site B (test de Student non apparié et unilatéral à gauche, t = -6,5, ddl = 49, $P < 0.001$).

Que pensez-vous de cette étude ?

UMONS

Critique statistiques (6)

Trois clients dans un restaurant payent leur repas : 30€ (10€ par personne). Le serveur se rend compte qu'en fait leur repas n'a coûté que 25€ en tout. Comme il ne pourra diviser les 5€ à rendre en trois facilement, il décide de garder 2€ dans sa poche et rend 1€ à chaque client. Donc, chaque client a payé $10 - 1 = 9$€, soit un total de 27€. Avec les 2€ que le serveur a gardé dans sa poche, cela fait 29€. Alors, où est passé l'euro manquant par rapport aux 30€ payés initialement ?

Réfléchissez et dénoncer l'erreur de raisonnement dans le récit précédent.

UMONS

Critique statistique (7)

Un chercheur dans une industrie chimique s'intéresse à l'effet d'un nouvel insecticide à effet progressif. Il teste son produit sur des drosophiles et observe une mortalité de 10% par jour, et ce, quel que soit le moment où il effectue les mesures après avoir mis les mouches en contact avec l'insecticide. Il en conclu qu'il faut 10 jours pour tuer toutes les mouches. Ce résultat est meilleur que le produit du concurrent, car ce dernier tue 80% des mouches sur la même durée de 10 jours.

Que pensez-vous de la façon dont cette expérience a été menée et de ses conclusion ?

UMONS

Critique statistique (8)

Par le plus grand des hasards, le numéro 8 est sorti 6 fois en 7 tirages successifs du lotto. Sachant qu'une vérification de ce que ce numéro n'a pas plus de chances que les autres d'être tiré au sort, vous ne manquerez pas de constater en bon statisticien(ne) que le numéro 8 est très nettement sur-représenté dans les tirages.

La prochaine fois que vous remplirez votre grille de lotto, jouerez-vous le numéro 8 ? Pourquoi ?

Vous est-il arriver de jouer la suite 1, 2, 3, 4, 5, 6, 7, 8 au lotto (ou rempliriez-vous une grille avec ces nombres si vous deviez y jouer) ? Pourquoi ?

UMONS