La Statistique : Une Science Omniprésente

Dhafer Malouche

Ecole Supérieure de la Statistique et de l'Analyse de l'Information
Université de Carthage

اَلمجمع التونسي للّعلوم و الأَدب
«بيت الحكمـــة»
تونس، 7 ديسمبر 2013

La Statistique :

La Statistique :

- Objectif : Mener, grâce à l'observation d'un phénomène **aléatoire**, une inférence sur la **distribution** probabiliste à l'origine de ce phénomène.

La Statistique :

- Objectif : Mener, grâce à l'observation d'un phénomène **aléatoire**, une inférence sur la **distribution** probabiliste à l'origine de ce phénomène.

- $\Rightarrow$ Fournir
  - une analyse ou une description d'un phénomène passé

La Statistique :

- Objectif : Mener, grâce à l'observation d'un phénomène **aléatoire**, une inférence sur la **distribution** probabiliste à l'origine de ce phénomène.

- $\Rightarrow$ Fournir
  - une analyse ou une description d'un phénomène passé
  - et/ou une prédiction d'un phénomène à venir de nature similaire.

La Statistique :

- Objectif : Mener, grâce à l'observation d'un phénomène **aléatoire**, une inférence sur la **distribution** probabiliste à l'origine de ce phénomène.

- $\Rightarrow$ Fournir
  - une analyse ou une description d'un phénomène passé
  - et/ou une prédiction d'un phénomène à venir de nature similaire.
- Aspects supplémentaires : Sondages, Plan d'expérience, Collectes de données..

Ces analyses doivent d'abord être motivées par un objectif

Ces analyses doivent d'abord être motivées par un objectif

- Comment peut-on prédire les retards des vols pour une compagnie aérienne ?

Ces analyses doivent d'abord être motivées par un objectif

- Comment peut-on prédire les retards des vols pour une compagnie aérienne ?
- Peut-on expliquer la rentabilité des clients d'une banque à partir de leurs caractéristiques socio-démographiques ?

Ces analyses doivent d'abord être motivées par un objectif

- Comment peut-on prédire les retards des vols pour une compagnie aérienne ?
- Peut-on expliquer la rentabilité des clients d'une banque à partir de leurs caractéristiques socio-démographiques ?
- Quelle est la variété de huile d'olive la plus préférée parmi un ensemble de variétés de huiles d'olive ? Ceci peut-il être expliqué par un ensemble de paramètres physicochimiques ?

Ces analyses doivent d'abord être motivées par un objectif

- Comment peut-on prédire les retards des vols pour une compagnie aérienne ?
- Peut-on expliquer la rentabilité des clients d'une banque à partir de leurs caractéristiques socio-démographiques ?
- Quelle est la variété de huile d'olive la plus préférée parmi un ensemble de variétés de huiles d'olive ? Ceci peut-il être expliqué par un ensemble de paramètres physicochimiques ?
- Peut-on donner une chronologie des œuvres de Platon ?

...etc..

Les feux de forêts

Les feux de forêts

- C'est un phénomène aléatoire

Les feux de forêts

- C'est un phénomène aléatoire
- Favorisé par des facteurs écologiques et/ou atmosphériques.

Les feux de forêts

- C'est un phénomène aléatoire
- Favorisé par des facteurs écologiques et/ou atmosphériques.
- Détermination de p (la probabilité d'apparition d'un feu) en fonction des divers facteurs.

Les feux de forêts

- C'est un phénomène aléatoire
- Favorisé par des facteurs écologiques et/ou atmosphériques.
- Détermination de p (la probabilité d'apparition d'un feu) en fonction des divers facteurs.
- Une approche : imposer une forme paramétrique à p

$$p = \exp(\alpha_1 h + \alpha_2 t + \alpha_3 x) / (1 + \exp(\alpha_1 h + \alpha_2 t + \alpha_3 x))$$

ou

$$\text{logit}(p) = \alpha_1 h + \alpha_2 t + \alpha_3 x$$

où h est l'humidité, t est la température et x est degré de gestion de la forêt.

Les feux de forêts

- C'est un phénomène aléatoire
- Favorisé par des facteurs écologiques et/ou atmosphériques.
- Détermination de p (la probabilité d'apparition d'un feu) en fonction des divers facteurs.
- Une approche : imposer une forme paramétrique à p

$$p = \exp(\alpha_1 h + \alpha_2 t + \alpha_3 x) / (1 + \exp(\alpha_1 h + \alpha_2 t + \alpha_3 x))$$

ou

$$\text{logit}(p) = \alpha_1 h + \alpha_2 t + \alpha_3 x$$

où h est l'humidité, t est la température et x est degré de gestion de la forêt.

- Phase statistique : évaluer α_1, α_2, α_3.

Détermination du nombre de bus dans une ville

Détermination du nombre de bus dans une ville

- Stratégie :
  1. suivre pendant toute une journée les bus et noter leurs numéros.
  2. répéter la même chose le lendemain en relevant les numéros déjà observés la veille : n.

Détermination du nombre de bus dans une ville

- Stratégie :
  1. suivre pendant toute une journée les bus et noter leurs numéros.
  2. répéter la même chose le lendemain en relevant les numéros déjà observés la veille : n.
- Supposons qu'on a observé 20 le 1er jour et 30 le 2nd jour dont n sont marquées.

Détermination du nombre de bus dans une ville

- Stratégie :
  1. suivre pendant toute une journée les bus et noter leurs numéros.
  2. répéter la même chose le lendemain en relevant les numéros déjà observés la veille : n.
- Supposons qu'on a observé 20 le 1er jour et 30 le 2nd jour dont n sont marquées.
- n est un nombre aléatoire de loi *Hypergéométrique*.

$$p(n) = \frac{\binom{20}{n}\binom{N-20}{30-n}}{\binom{N}{30}}$$

Détermination du nombre de bus dans une ville

- Stratégie :
  1. suivre pendant toute une journée les bus et noter leurs numéros.
  2. répéter la même chose le lendemain en relevant les numéros déjà observés la veille : n.
- Supposons qu'on a observé 20 le 1er jour et 30 le 2nd jour dont n sont marquées.
- n est un nombre aléatoire de loi *Hypergéométrique*.

$$p(n) = \frac{\binom{20}{n}\binom{N-20}{30-n}}{\binom{N}{30}}$$

- On montre

$$N \approx \frac{20 \times 30}{n}$$

Détermination du nombre de bus dans une ville

- Stratégie :
  1. suivre pendant toute une journée les bus et noter leurs numéros.
  2. répéter la même chose le lendemain en relevant les numéros déjà observés la veille : n.
- Supposons qu'on a observé 20 le 1er jour et 30 le 2nd jour dont n sont marquées.
- n est un nombre aléatoire de loi *Hypergéométrique*.

$$p(n) = \frac{\binom{20}{n}\binom{N-20}{30-n}}{\binom{N}{30}}$$

- On montre

$$N \approx \frac{20 \times 30}{n}$$

- Méthode *capture-recapture* couramment utilisée en écologie pour estimer la taille d'une population animale.

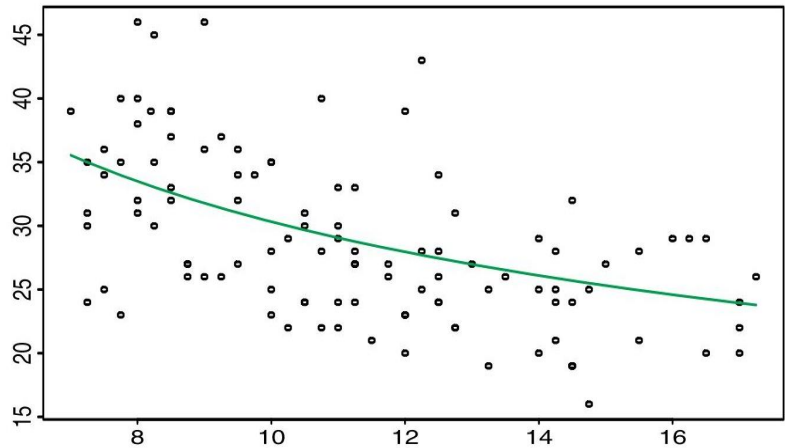
Taux de chômage, Nombre d'accidents

Taux de chômage, Nombre d'accidents

Lenk 1999, données recueillit dans le Michigan entre 1978 et 1987

Taux de chômage, Nombre d'accidents

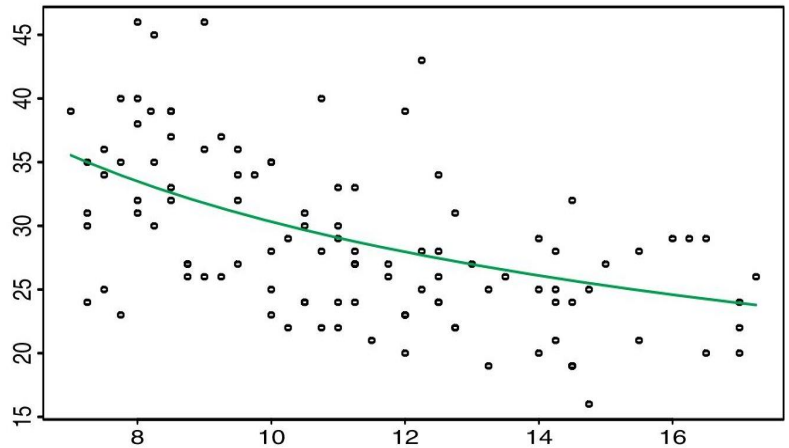
Lenk 1999, données recueillit dans le Michigan entre 1978 et 1987



Taux de chômage, Nombre d'accidents

Lenk 1999, données recueillit dans le Michigan entre 1978 et 1987

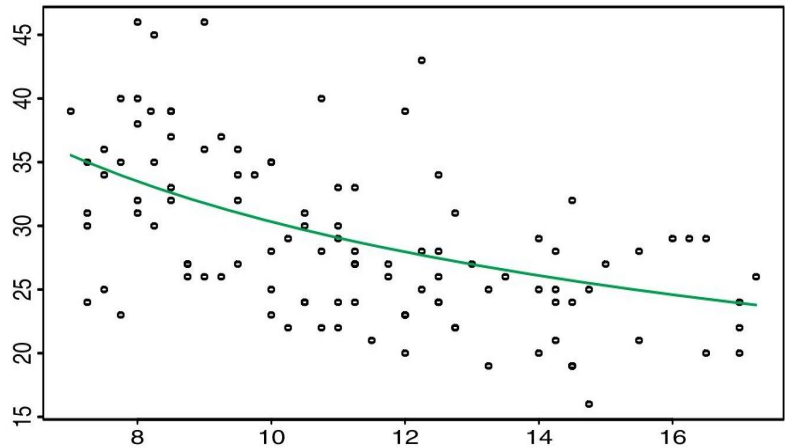
Une hausse du chômage entraîne-t-elle une baisse des accidents de la route ?



Taux de chômage, Nombre d'accidents

Lenk 1999, données recueillit dans le Michigan entre 1978 et 1987

Une hausse du chômage entraîne-t-elle une baisse des accidents de la route ?



Un structure paramétrique de la dépendance entre les deux variables

$$N \mid \tau \sim \text{Poisson}(\lambda)$$

où

$$\log \lambda = \beta_0 + \beta_1 \log(\tau)$$

C'est le modèle de *régression de Poisson.*

Algorithme d'amélioration des radiographies

Radiographie des poumons

Algorithme d'amélioration des radiographies

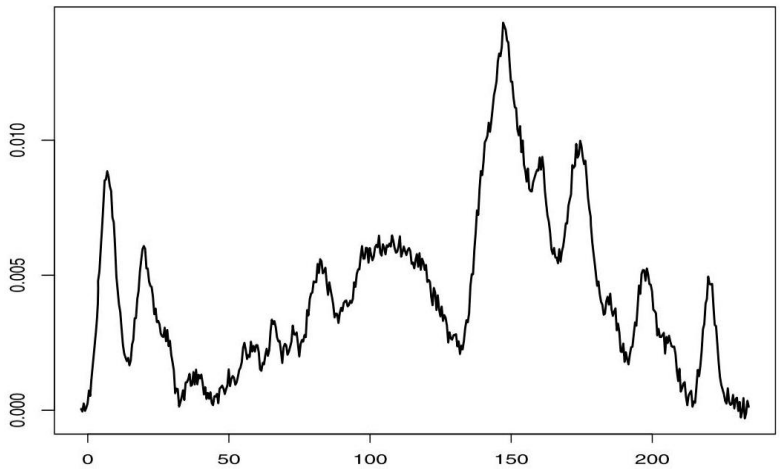
C'est une grille de 1000 $\times$ 1200 points (*pixels*) avec un niveau de gris associé à un nombre compris entre 0 et 256

Radiographie des poumons

Algorithme d'amélioration des radiographies

C'est une grille de 1000 $\times$ 1200 points (*pixels*) avec un niveau de gris associé à un nombre compris entre 0 et 256

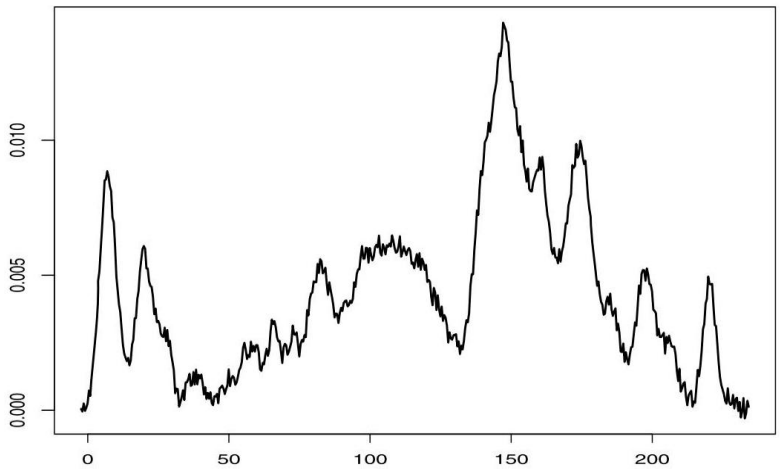
L'histogramme donne une approximation de la distribution



Algorithme d'amélioration des radiographies

C'est une grille de 1000 $\times$ 1200 points (*pixels*) avec un niveau de gris associé à un nombre compris entre 0 et 256

L'histogramme donne une approximation de la distribution

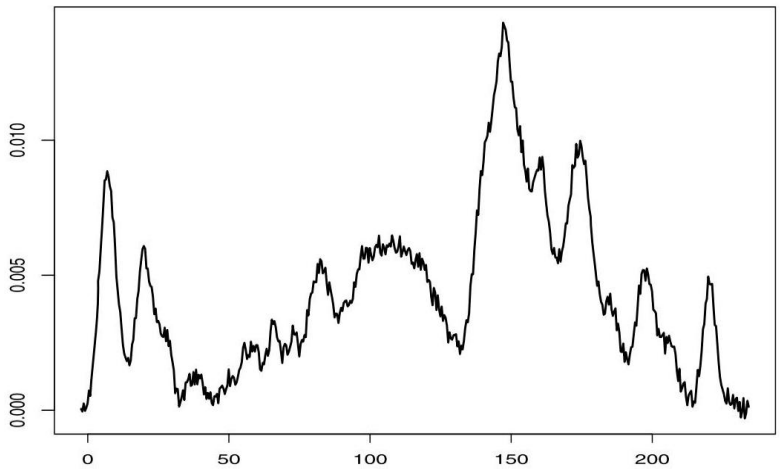


Une distribution approximativement **bimodale**,

Algorithme d'amélioration des radiographies

C'est une grille de 1000 $\times$ 1200 points (*pixels*) avec un niveau de gris associé à un nombre compris entre 0 et 256

L'histogramme donne une approximation de la distribution



Une distribution approximativement **bimodale**,
Modélisation par un mélange de lois Normales

$$f(x) = \frac{p}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] + \frac{1-p}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right]$$

Algorithme d'amélioration des radiographies

- Les deux modes correspondent à deux régions de la poitrine : les poumons et le médiastin (la région de la cage thoracique située entre les deux poumons contenant le cœur, l'œsophage, la trachée et les deux bronches souches).

Algorithme d'amélioration des radiographies

- Les deux modes correspondent à deux régions de la poitrine : les poumons et le médiastin (la région de la cage thoracique située entre les deux poumons contenant le cœur, l'œsophage, la trachée et les deux bronches souches).

- C'est une technique de lissage utilisée dans l'algorithme d'amélioration des radiographies : *Parametric Histogram Specification*, Plessis 1989.

Deux approches...

Deux approches...

- **Statistique non paramétrique** :

Deux approches...

- **Statistique non paramétrique** :
  - Prendre en compte de la complexité de l'expérience aléatoire tout en cherchant à estimer la distribution sous jacente sous des hypothèses minimales : Estimation fonctionnelle.

Deux approches...

- **Statistique non paramétrique** :
  - Prendre en compte de la complexité de l'expérience aléatoire tout en cherchant à estimer la distribution sous jacente sous des hypothèses minimales : Estimation fonctionnelle.
  - Par exemple, l'estimation de la distribution :

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

où $h > 0$ est appelée la *fenêtre* (*bandwidth*) et K est la fonction *noyau*

Deux approches...

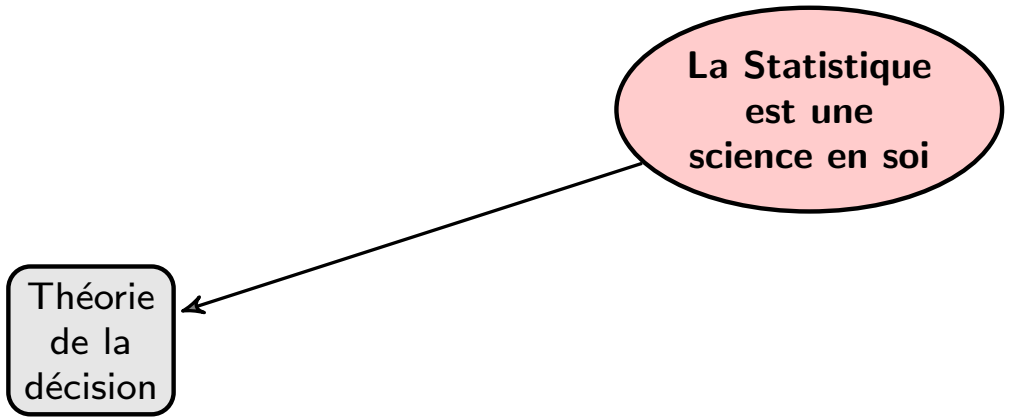
- **Statistique non paramétrique** :
  - Prendre en compte de la complexité de l'expérience aléatoire tout en cherchant à estimer la distribution sous jacente sous des hypothèses minimales : Estimation fonctionnelle.
  - Par exemple, l'estimation de la distribution :

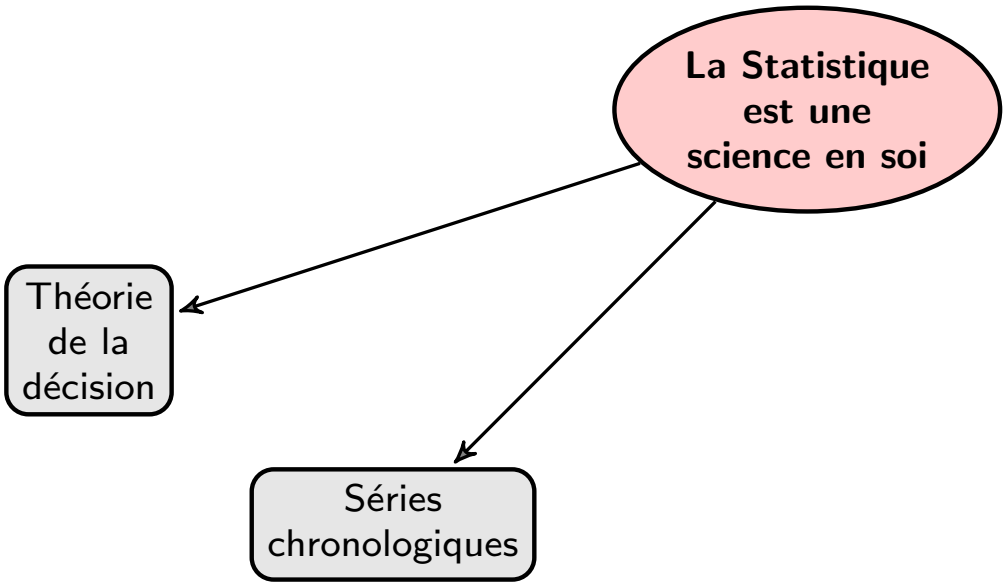
$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

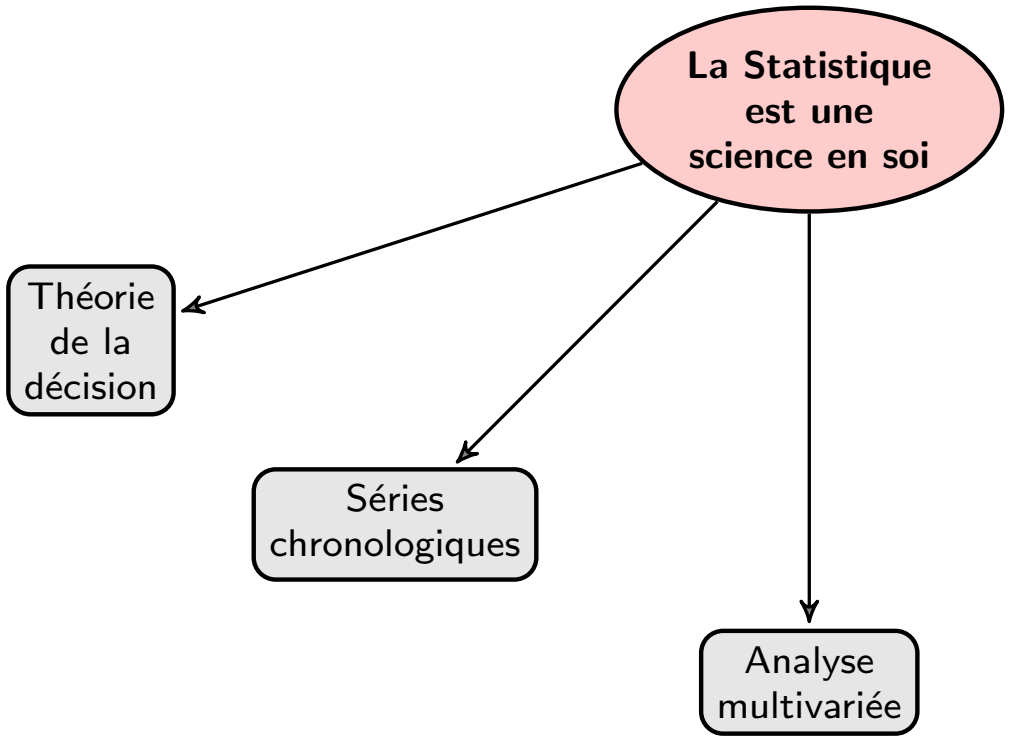
où $h > 0$ est appelée la *fenêtre* (*bandwidth*) et K est la fonction *noyau*

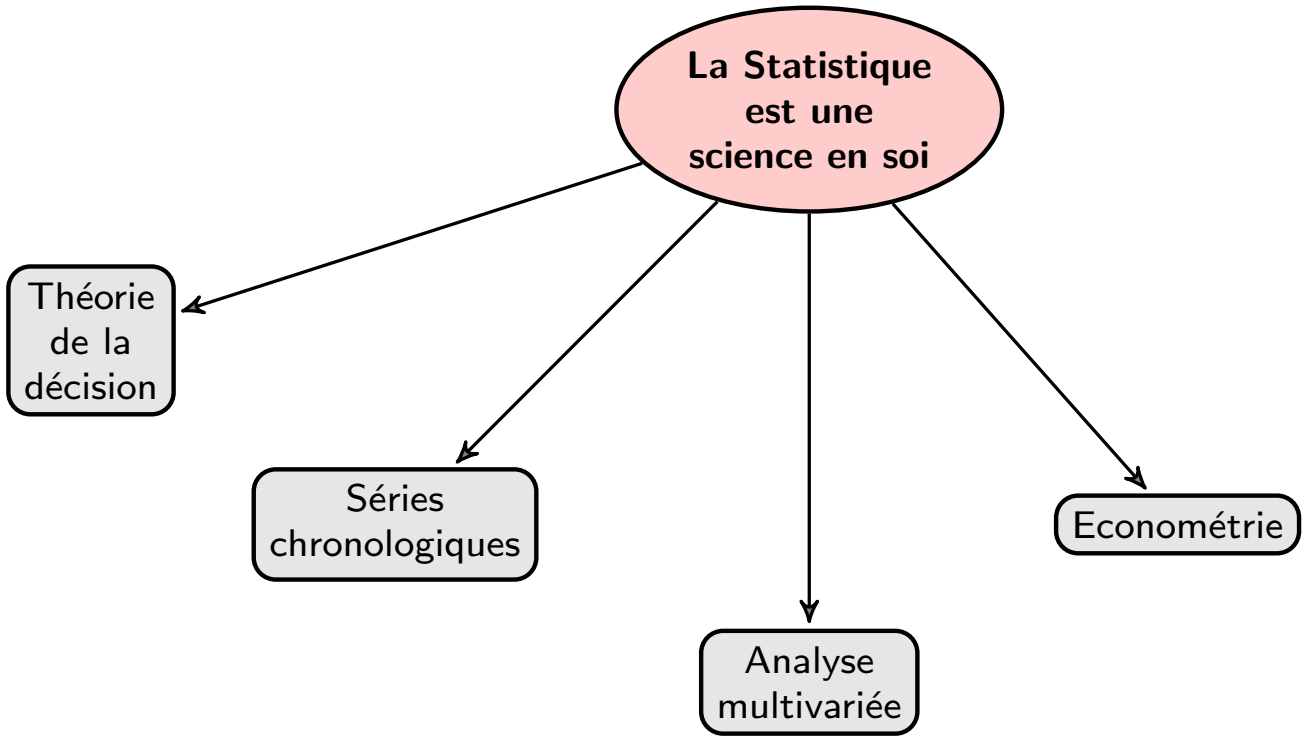
- **Statistique paramétrique** : suppose que les observations $x_1, \ldots, x_n$ proviennent de lois de probabilités $f_i\left(x_i \mid \theta_i, x_1, \ldots, x_{i-1}\right)$ où
  - θ_i est inconnue
  - f_i est connue.

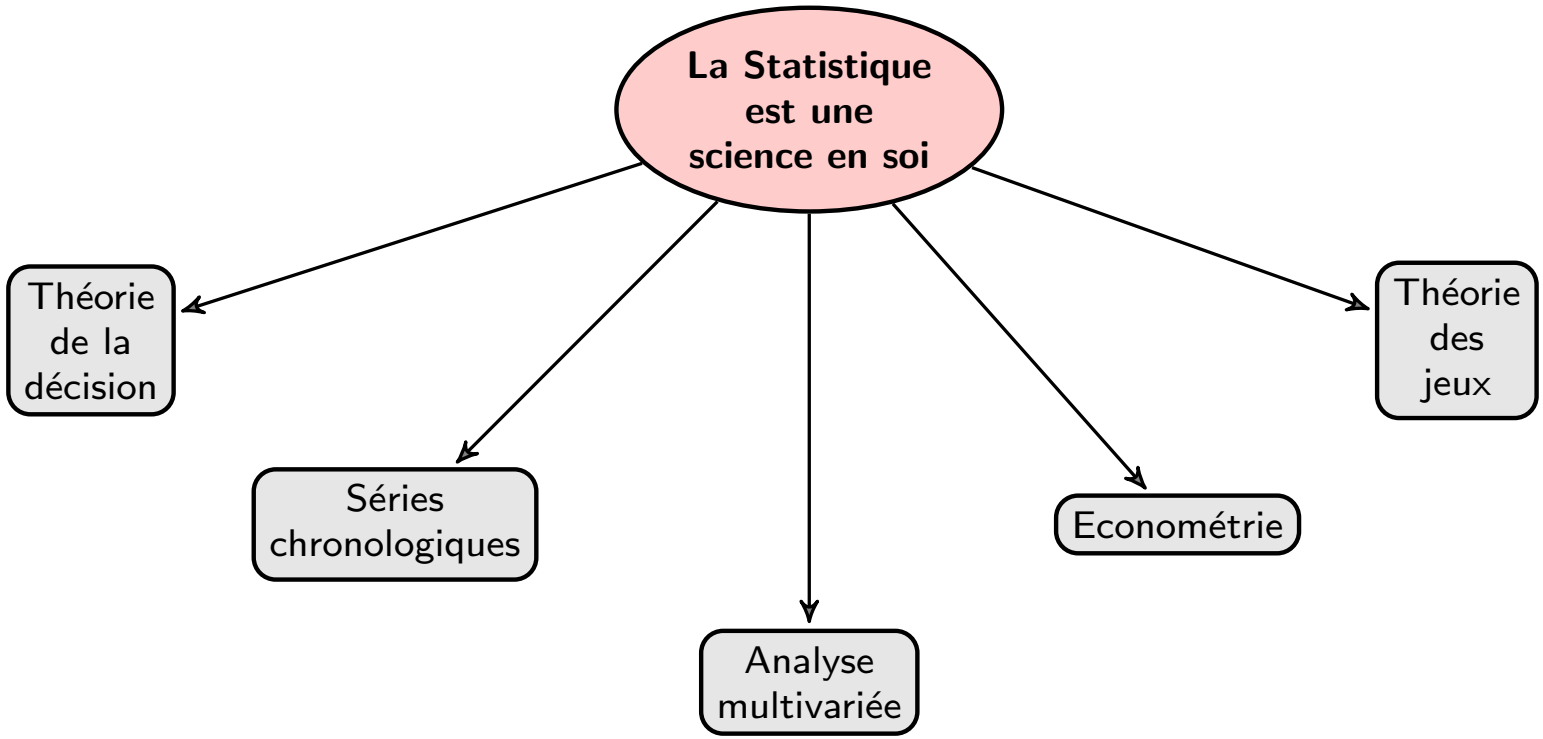

La Statistique
est une
science en soi

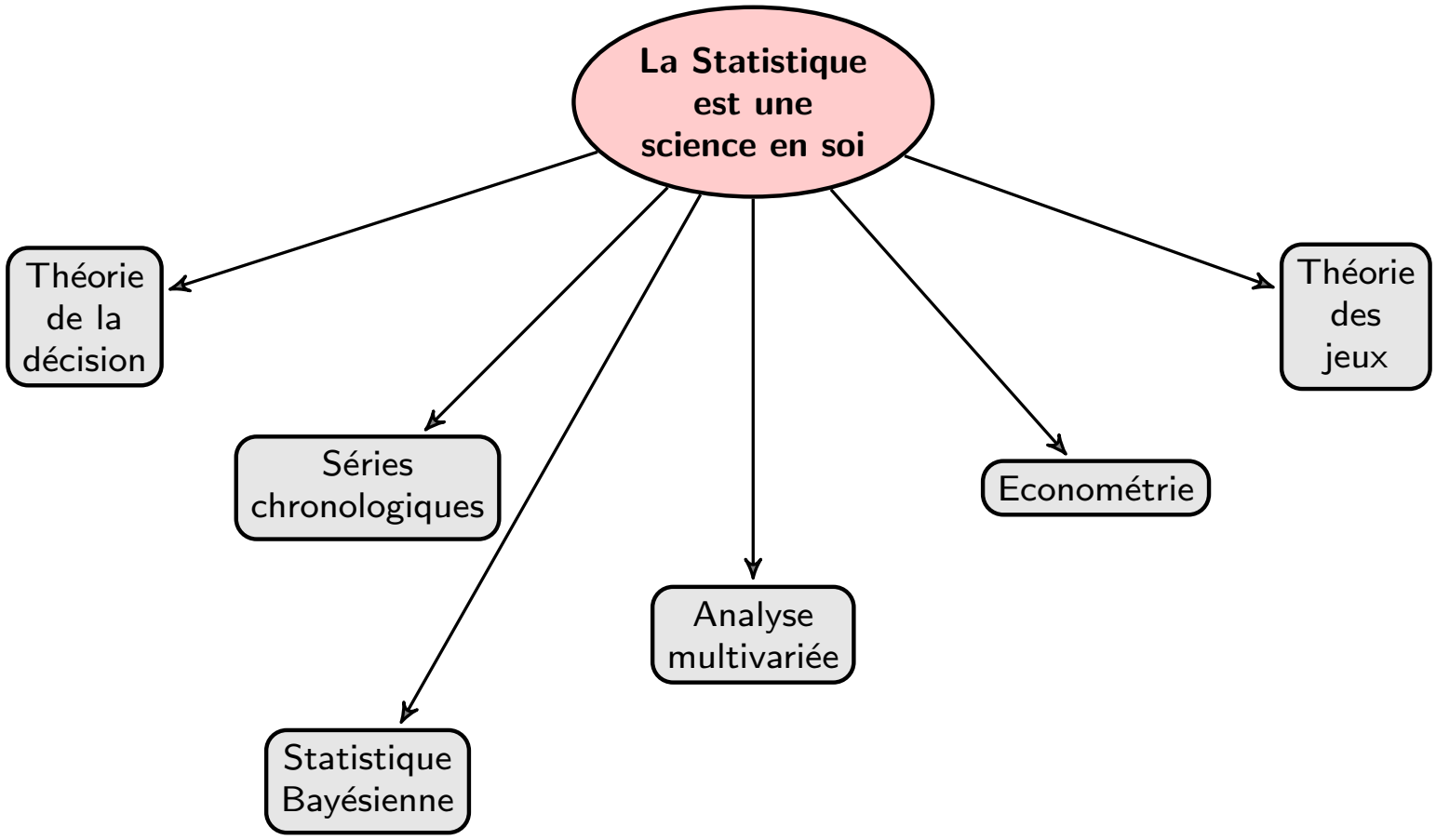

La Statistique est une science en soi
Théorie de la décision

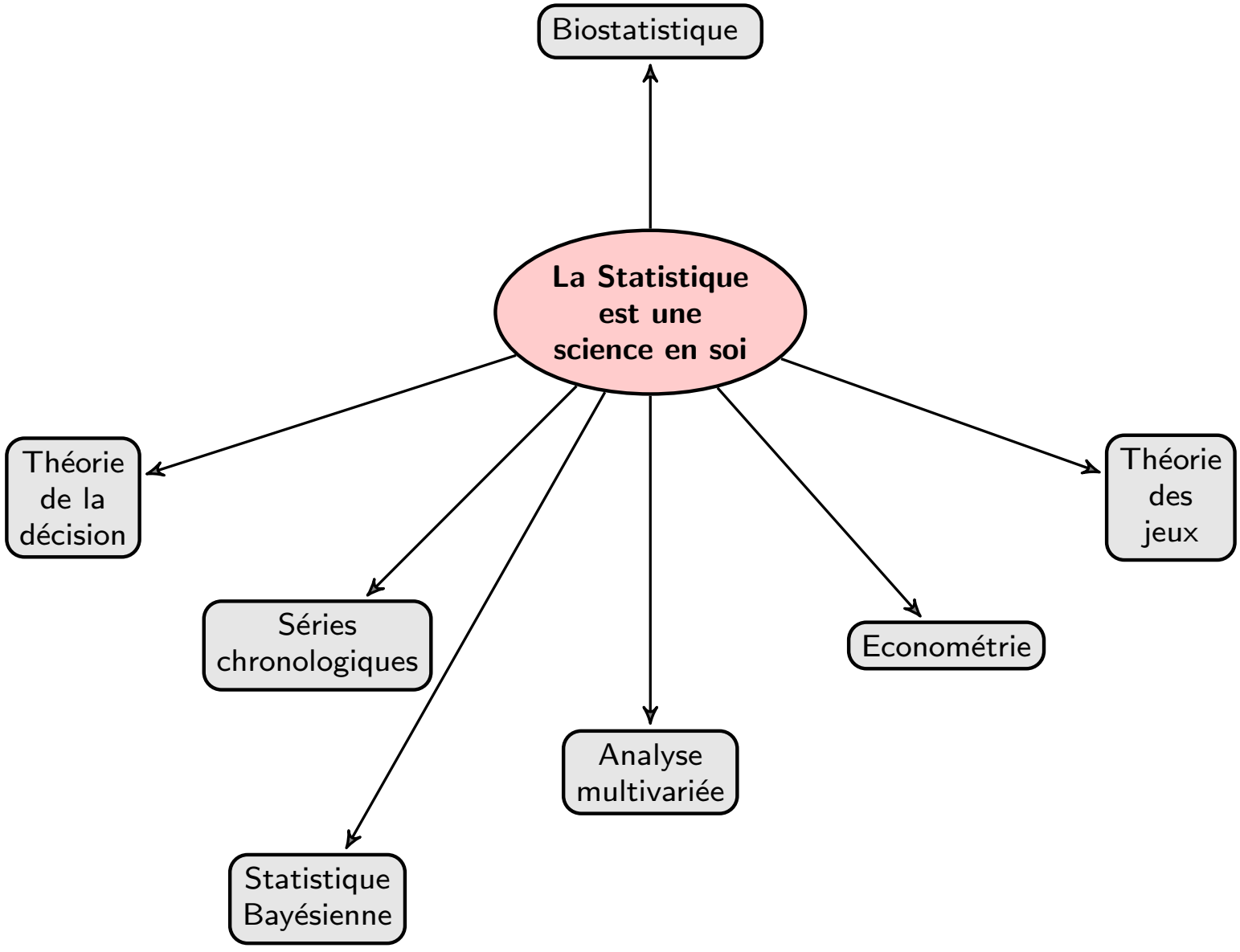

La Statistique est une science en soi
Théorie de la décision
Séries chronologiques

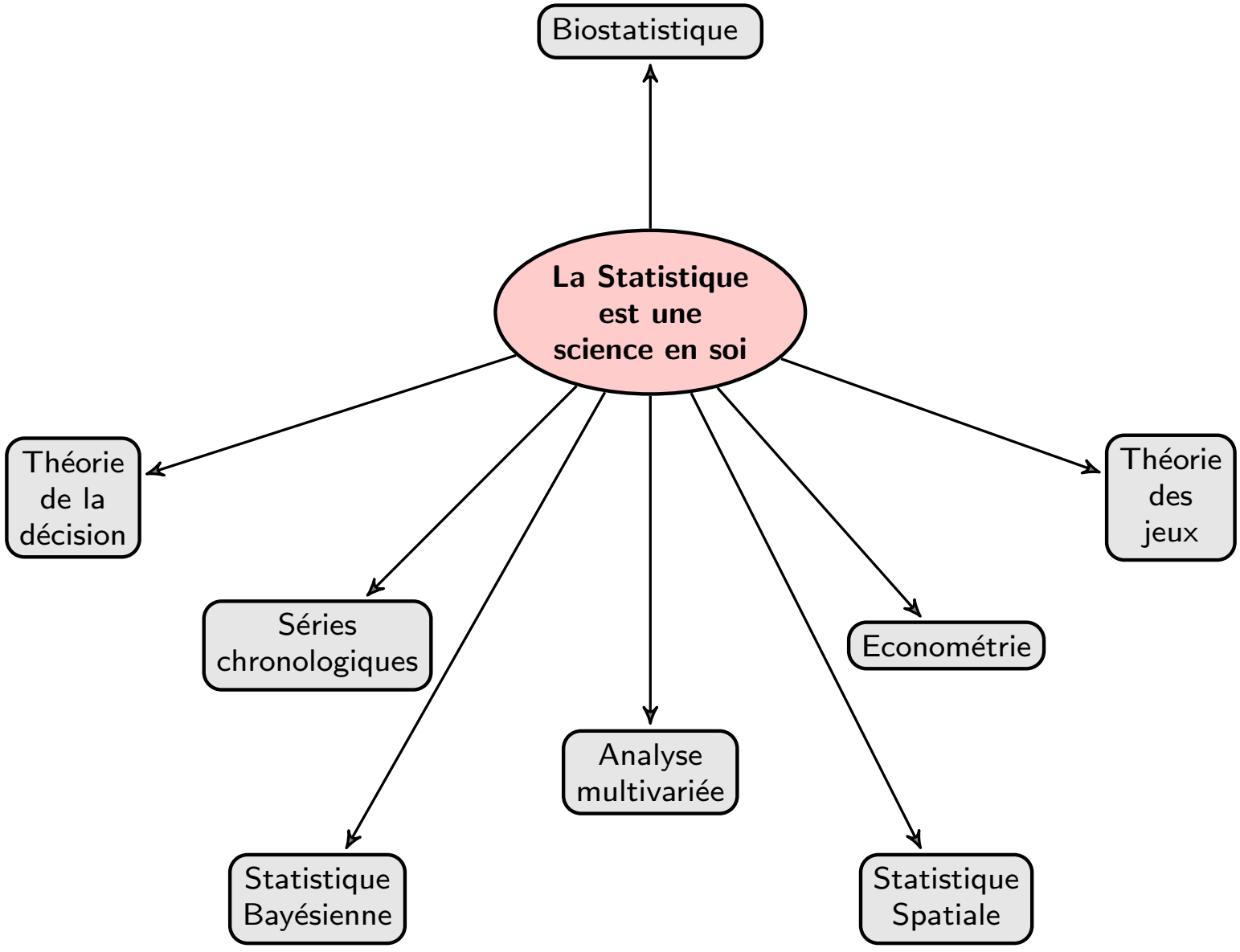

La Statistique est une science en soi
Théorie de la décision
Séries chronologiques
Analyse multivariée

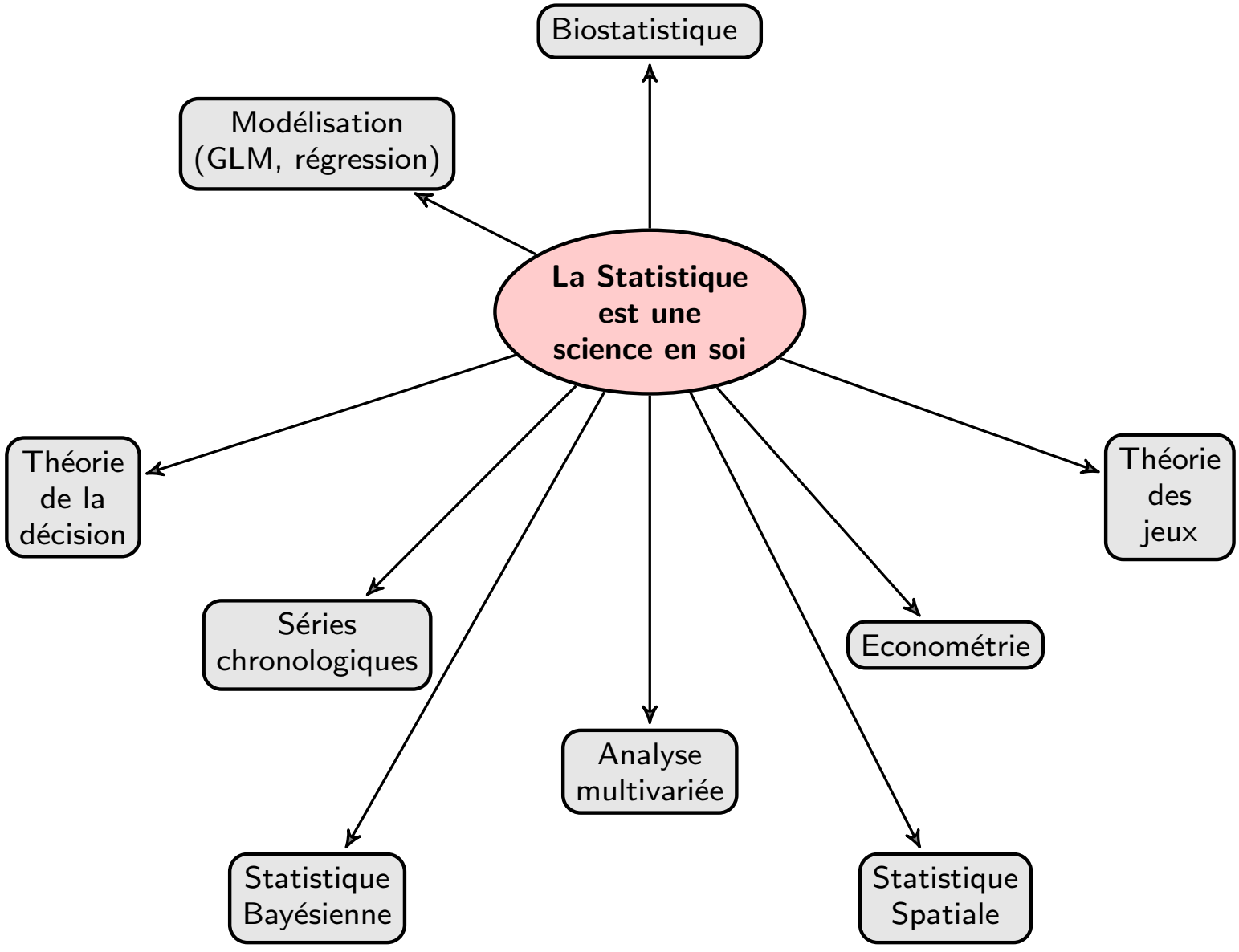

La Statistique est une science en soi
Théorie de la décision
Séries chronologiques
Analyse multivariée
Econométrie

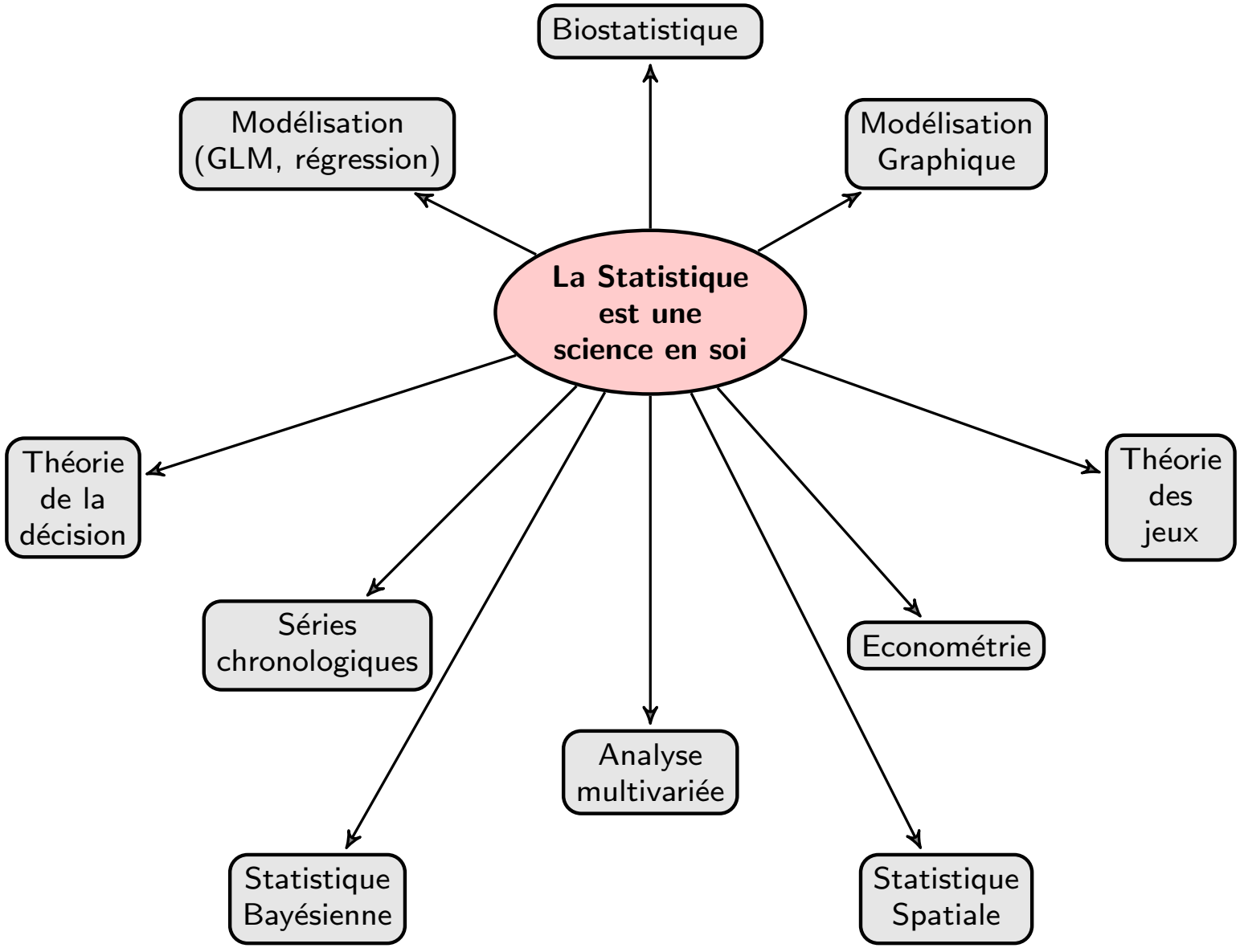

La Statistique est une science en soi
Théorie de la décision
Séries chronologiques
Analyse multivariée
Econométrie
Théorie des jeux

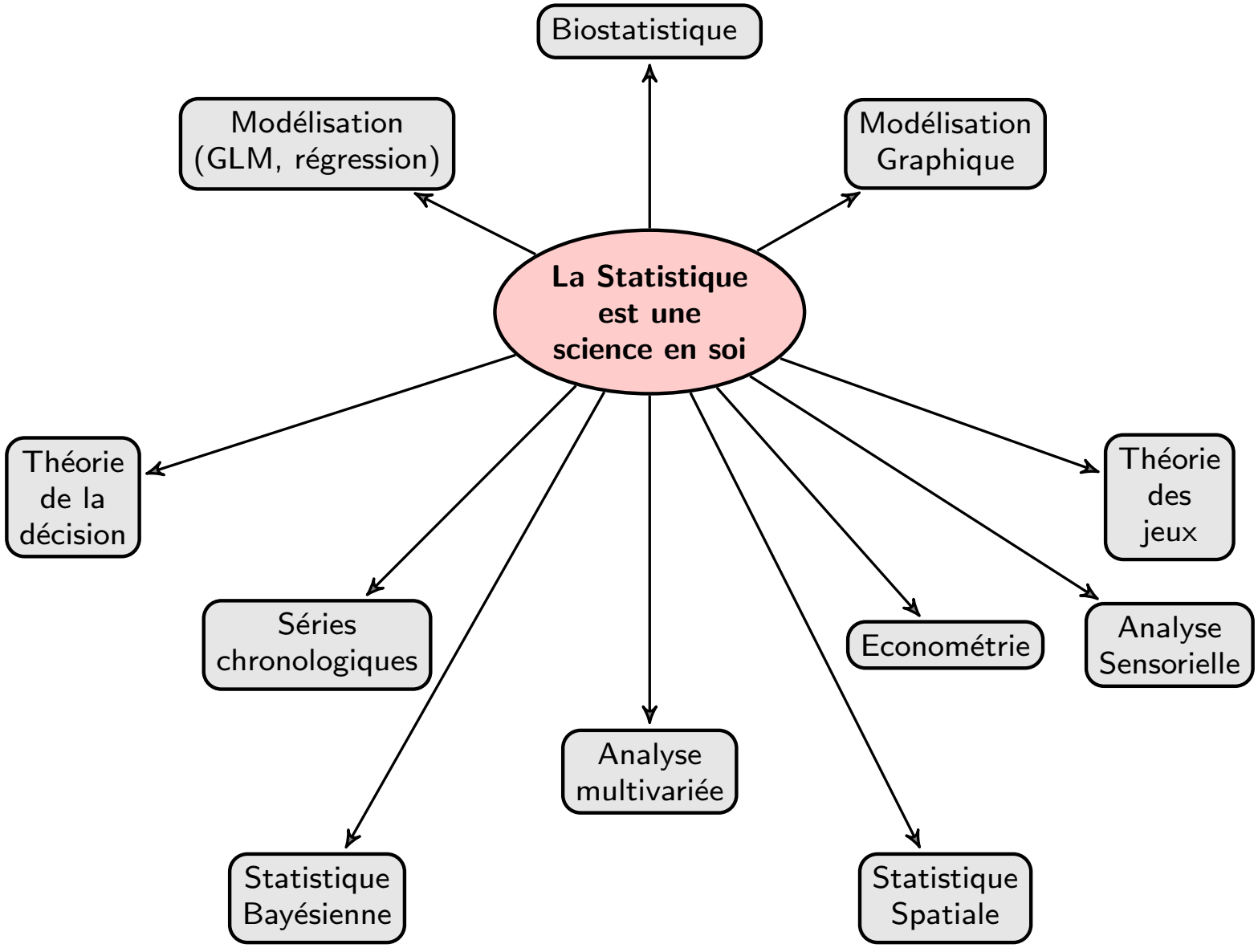

La Statistique est une science en soi
Théorie de la décision
Séries chronologiques
Statistique Bayésienne
Analyse multivariée
Econométrie
Théorie des jeux


Biostatistique
La Statistique est une science en soi
Théorie de la décision
Théorie des jeux
Séries chronologiques
Econométrie
Analyse multivariée
Statistique Bayésienne


Biostatistique
La Statistique est une science en soi
Théorie de la décision
Théorie des jeux
Séries chronologiques
Econométrie
Analyse multivariée
Statistique Bayésienne
Statistique Spatiale


Biostatistique
Modélisation (GLM, régression)
La Statistique est une science en soi
Théorie de la décision
Théorie des jeux
Séries chronologiques
Econométrie
Analyse multivariée
Statistique Bayésienne
Statistique Spatiale


Biostatistique
Modélisation (GLM, régression)
Modélisation Graphique
La Statistique est une science en soi
Théorie de la décision
Théorie des jeux
Séries chronologiques
Econométrie
Analyse multivariée
Statistique Bayésienne
Statistique Spatiale


Biostatistique
Modélisation (GLM, régression)
Modélisation Graphique
La Statistique est une science en soi
Théorie de la décision
Théorie des jeux
Séries chronologiques
Econométrie
Analyse Sensorielle
Analyse multivariée
Statistique Bayésienne
Statistique Spatiale