# Conditions d'application des méthodes statistiques paramétriques

Dr. Ir. Epiphane SODJINOU

Agroéconomiste, Biostatisticien

## Plan

1. Introduction: principales méthodes statistiques avec les critères de choix
2. Principales méthodes statistiques paramétriques avec leurs conditions d'application
3. Importance pratique du respect des conditions d'application
4. Conséquences pratiques du non-respect des conditions d'application
5. Tests d'hypothèses pour la vérification des conditions d'application

# 1. Introduction

- Qu'est-ce que la statistique?
  - La statistique est à la fois une science formelle, une méthode et une technique.
  - Ne doit pas être confondue avec une statistique qui est un nombre calculer à partir d'observation
  - Science et techniques d'interprétation mathématique de données complexes et nombreuses, permettant de faire des prévisions
- Il y a, en général, deux approches en statistiques, entre lesquelles on jongle constamment :
  - les statistiques *exploratoires* : on explore d'abord les données pour avoir une idée qualitative de leurs propriétés
  - les statistiques *confirmatoires* (*exploratory and confirmatory statistics*) : ensuite on fait des hypothèses de comportement que l'on confirme ou infirme en recourant à d'autres techniques statistiques

# 1. Introduction

- Le choix d'une méthode d'analyse statistique dépend de vos objectifs et hypothèses qui peuvent être regroupés en :
  - Décrire
  - Modéliser
  - Analyser
  - Tester

# 1. Introduction

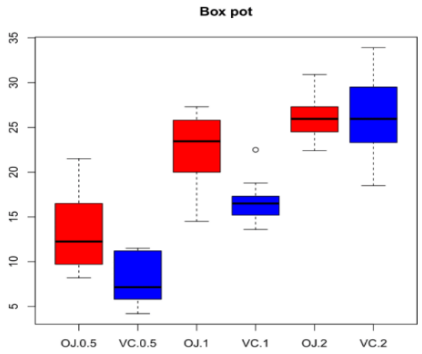
**Décrire**

- Statistiques descriptives
  - Ensemble d'indicateurs calculés sur un échantillon permettant de définir de manière univariée une variable
  - Comprennent :
    - Pour une variable quantitative : moyenne, médiane, écart-type, variance, quantiles, aplatissement et l'asymétrie
    - Pour une variable qualitative : mode, proportion par modalité
- Quantiles
  - Ex : On appelle centiles ou percentiles les quantiles avec la valeur du paramètre ramenée en pourcentages

# 1. Introduction

**Décrire**

- Box-plot




- Histogramme
- Corrélation

# 1. Introduction

**Modéliser**

- La plupart des méthodes économétriques se retrouvent dans ce lot
- Incluent globalement:
  - Le modèle linéaire général
    - modèle le plus utilisé en statistique
    - Comprend comme cas particulier : l'analyse de la variance (ANOVA), la régression linéaire, l'analyse de la covariance (ANCOVA).
    - La seule différence entre ces trois approches réside dans la nature des variables explicatives (qualitatives pour l'ANOVA, quantitatives pour la régression linéaire et des deux types pour l'ANCOVA).
    - But: Identifier une relation linéaire entre des variables explicatives indépendantes (X) et une variable à expliquer dépendante (Y).
    - Faire recours souvent à la méthode des moindres carrés

# 1. Introduction

**Modéliser**

- Incluent globalement:
  - Le modèle linéaire mixte
    - Il s'agit d'une variante du modèle linéaire général dans lequel on introduit les notions de facteurs répétés et de facteurs aléatoires.
    - Un **facteur répété** = facteur qui identifie des répétitions d'observations sur les mêmes individus
    - Un **facteur aléatoire** = facteur pour lequel on considère que les modalités présentes dans l'échantillon représentent un échantillon issu d'une population beaucoup plus large

# 1. Introduction

**Modéliser**

- Incluent globalement:
  - Les modèles logistiques
    - modèles de régression pour lesquels la variable dépendante à expliquer est une variable qualitative
    - **Régression logistique** : suppose une variable dépendante binaire et va permettre de classer des individus dans l'une des deux classes de la variable binaire
    - **Modèle logit multinomial** : suppose une variable dépendante à expliquer qualitative avec plus de deux modalités.
    - **Modèle logit ordinal** : suppose une variable dépendante à expliquer ordinale
    - **Modèle logit conditionnel** : est moins connu que les précédents et qui suppose une variable dépendante qualitative non ordonnée.
      Ce modèle de régression va permettre d'expliquer le choix d'une modalité en utilisant les informations liées à toutes les modalités de la variable dépendante

# 1. Introduction

**Modéliser**

- Incluent globalement:
  - La régression sur les composantes principales
    - permet de combler deux faiblesses de la régression linéaire : la multicolinéarité entre les variables indépendantes et le cas d'un tableau avec plus de colonnes que de lignes.
    - Principe : faire une ACP sur les variables explicatives, puis appliquer une régression linéaire entre les composantes et la variable dépendante à expliquer.

# 1. Introduction

**Modéliser**

- Incluent globalement:
  - La régression PLS
    - développée par Svante Wold et permet de remédier à de nombreuses faiblesses des méthodes de régression plus classiques :
      - Le cas de multicolinéarité entre les variables indépendantes.
      - Le cas d'un tableau avec plus de colonnes que de lignes.
      - La présence de données manquantes.
      - Le traitement d'un bloc de variables dépendantes (Y est alors une matrice)

# 1. Introduction

**Modéliser**

- Incluent globalement:
  - La régression non linéaire
  - La régression LASSO
  - Le modèle elasticnet
  - La régression Ridge
  - La régression log-linéaire
  - La régression quantile
  - La régression non paramétrique
  - Les modèles de survie
  - Le modèle de Cox
  - La régression de Weibull

# 1. Introduction

**Analyser**

- Analyse en composantes principales
- Analyse discriminante
- Classification k-means
- Classification hierarchique
- Cartes de Kohonen
- Analyse factorielle multiple (AFM)
- Analyses de tableaux multiples
- ACP mixte
- Méthode Statis
- Analyse factorielle des correspondances

# 1. Introduction

**Tester**

- Comparaison de moyennes
- Comparaison de variances
- Comparaison de distribution
- Test du chi² sur un tableau de contingence
- Tests non paramétriques de comparaison de 2 échantillons (Mann-Whitney, Wilcoxon)
- Tests non paramétriques de comparaisons de k échantillons (Kruskall-Wallis, Friedman)
- Tests non paramétriques de comparaisons de 2 distributions (Kolmogorov-Smirnov)
- Test de Mantel
- Tests de corrélations
- Tests de normalité (Shapiro-Wilk, Jacque-Bera, etc.)

# 2. **Principales méthodes avec leurs conditions d'application**

## *2.0. Introduction*

- Méthodes statistiques paramétriques nécessitent le respect des hypothèses de base faites lors de leur conception
- La violation des conditions d'application de ces méthodes donne souvent lieu à de fausses interprétations des résultats obtenus puisque rien ne garantit la précision des méthodes en dehors de leurs hypothèses d'utilisation
- La méconnaissance par l'utilisateur des hypothèses l'amène souvent à ignorer cette étape importante du traitement des données de recherche

# 2. **Principales méthodes avec leurs conditions d'application**

## *2.1. Méthodes statistiques paramétriques pour une variable et conditions d'utilisation*

| Méthodes statistiques paramétriques | Conditions d'application |
|---|---|
| Test de conformité d'une proportion<br>Test d'égalité de 2 ou plusieurs proportions | - Echantillons aléatoires simples et indépendants. |
| Test de conformité d'une moyenne | - Echantillon aléatoire simple.<br>- Echantillon tiré de population normale. |
| Test d'égalité de deux moyennes | - Echantillons aléatoires simples et indépendants.<br>- Echantillons tirés de populations normales. |
| Test t pour données appariées | - Echantillons aléatoires simples et dépendants.<br>- Echantillons tirés de populations normales. |
| Test de conformité d'un ou de deux écarts-types (ou variances) | - Echantillons aléatoires, simples et indépendants (ou non)[1].<br>- Echantillons tirés de populations normales. |

# 2. Principales méthodes avec leurs conditions d'application

## *2.1. Méthodes statistiques paramétriques pour une variable et conditions d'utilisation*

| | |
|---|---|
| Test de conformité du rapport de deux écarts-types ou de deux variances à une valeur théorique. | - Echantillons aléatoires et simples.<br>- Echantillons tirés de populations normales. |
| Test d'égalité de plusieurs écarts-types ou de plusieurs variances (test de Bartlett, test de Levene, etc.). | - Echantillons aléatoires, simples et indépendants.<br>- Echantillons tirés de populations normales ou non[2]. |
| Analyse de la variance à $p$ critères de classification. | - Echantillons aléatoires et indépendants.<br>- Echantillons tirés de populations normales.<br>- Egalité des variances des populations. |

# 2. Principales méthodes avec leurs conditions d'application

## *2.2. Méthodes statistiques pour deux variables observées simultanément et conditions d'utilisation*

| Méthodes statistiques paramétriques | Conditions d'application |
|---|---|
| Test d'indépendance ou test du Chi2. | - Echantillons aléatoires et simples. |
| Test de signification ou de conformité d'un coefficient de corrélation.<br><br>Test d'égalité de deux coefficients de corrélation. | - Echantillons aléatoires et simples.<br>- Echantillons tirés de populations normales bivariées.<br>- Valeurs de variables connues sans erreurs de mesure. |
| Régression linéaire simple.<br><br>Test d'égalité de deux coefficients de régression.<br><br>Test de conformité d'un coefficient de régression. | - Normalité des résidus de régression.<br>- Nullité de la moyenne des résidus.<br>- Homogénéité des résidus de régression.<br>- Indépendance des résidus de régression. |

# 2. Principales méthodes avec leurs conditions d'application

## *2.3. Méthodes statistiques multivariées et conditions d'utilisation*

| Méthodes statistiques paramétriques | Conditions d'application |
|---|---|
| Analyse en composantes principales (ACP) | - Aucune condition |
| Analyse factorielle des correspondances (AFC) | - Tableau de contingence. |
| La classification numérique | - Aucune condition |
| Analyse discriminante linéaire et quadratique | - Echantillons aléatoires et simples.<br>- Echantillons tirés de populations multinormales.<br>- Egalité ou non[1] des matrices de variances-covariances. |
| Analyse de la variance multivariée et analyse canonique discriminante | - Echantillons aléatoires et simples.<br>- Echantillons tirés de populations multinormales.<br>- Egalité des matrices de variances-covariances. |
| L'analyse de la corrélation canonique | - Echantillons aléatoires et simples.<br>- Echantillons tirés de populations multinormales. |

# 2. Principales méthodes avec leurs conditions d'application

## *2.4. Synthèse des conditions d'utilisation des méthodes statistiques*

- Certaines méthodes statistiques ne nécessitent que la condition d'échantillons aléatoires et simples
  - Ce sont les tests relatifs à une proportion, le test d'indépendance et les méthodes multivariées descriptives comme l'ACP, l'AFC et la classification numérique.
- Les conditions d'utilisation des méthodes statistiques paramétriques peuvent être résumées de façon générale en 3 groupes :
  - le caractère aléatoire et simple ainsi que l'indépendance des échantillons soumis aux méthodes paramétriques ou encore l'indépendance des résidus de régression
  - la normalité à une ou plusieurs dimensions ou encore la normalité des résidus de régression
  - caractère homoscédastique des échantillons :
    - égalité des variances ou encore égalité des matrices de variances-covariances ; de même, l'homogénéité des variances résiduelles.

# 3. Importance pratique du respect des conditions d'application

***3.1. Importance de la normalité en inférence statistique***

- Normalité de la population dont est issu l'échantillon
  - une des conditions les plus importantes dans l'utilisation des méthodes paramétriques.
- En inférence statistique
  - le calcul de la probabilité associée à un test, de même que l'estimation et la détermination des limites de confiance d'une moyenne ou d'un écart-type se basent sur l'hypothèse de normalité des observations.
- Lorsqu'une telle hypothèse (la normalité) n'est pas satisfaite, l'utilisation de la méthode statistique peut conduire à des résultats biaisés.

# 3. Importance pratique du respect des conditions d'application

***3.1. Importance de la normalité en inférence statistique***

- La propriété de normalité asymptotique de la distribution d'échantillonnage de la moyenne rend moins importante la condition de normalité pour de grands échantillons dans le cas des tests d'égalité de moyennes ou de vecteurs de moyennes
- Il n'en est pas de même lorsqu'on s'intéresse à la structuration de moyennes après une analyse de la variance (tests de Newman et Keuls, de Dunnett, de Bonferroni, de Tukey, etc.)
- Les tests d'égalité des variances, écarts-types ou matrices de variances-covariances sont nettement plus sensibles à la non-normalité des populations-parents.

# 3. Importance pratique du respect des conditions d'application

***3.2. Importance de la condition d'homoscédasticité en inférence statistique***

- Condition d'homoscédasticité en inférence statistique
  - concerne l'hypothèse d'égalité des variances ou écarts-types des échantillons dans le cas des tests univariés, et l'égalité des matrices de variances-covariances dans le cas des tests multivariés
- En inférence statistique à deux ou plusieurs dimensions,
  - la condition d'homoscédasticité est nécessaire surtout en cas de structuration des vecteurs de moyennes avec l'analyse canonique discriminante puisqu'elle utilise l'estimation commune des matrices de variances-covariances des populations.

# 3. Importance pratique du respect des conditions d'application

***3.2. Importance de la condition d'homoscédasticité en inférence statistique***

- En inférence statistique à une dimension (tests de comparaison de moyennes, analyse de la variance etc.),
  - l'hypothèse d'égalité des variances des échantillons prend toute son importance dans l'estimation et la détermination des limites de confiance ainsi que la détermination du nombre d'observations.
- En effet, pour ces différentes situations, c'est la variance estimée commune des différentes populations qui est utilisée.
  - Ceci suppose que ces variances doivent être significativement égales pour garantir une bonne précision de calcul de la variance commune.
  - Il en est de même en analyse discriminante linéaire où les matrices de variances-covariances doivent être significativement égales pour une estimation non biaisée de la matrice de variances-covariances groupée des populations multivariées considérées.

# 4. Conséquences pratiques du non-respect des conditions d'application

## *4.1. Non-normalité associée à une homoscédasticité*

- Lire pages 24-27 du document reçu

# 4. Conséquences pratiques du non-respect des conditions d'application

***4.2. Hétéroscédasticité associée à une normalité***

- Lire pages 29-30 du document reçu
- Conséquences de la non-homogénéité des variances en inférence statistique sont abordées par Dehlert (2000). Résultats de son étude, pour le cas du test *t* de Student :
- En cas d'égalité des tailles de deux échantillons,
  - si le rapport des variances est inférieur à 5, le risque réel est de 50 % supérieur au risque nominal (5 %). ➔ valeurs de probabilités sont sous-estimées conduisant à un *test libéral* (signification plus facile).
- En cas d'inégalité des tailles avec un rapport des variances inférieur à 5 et si de plus, les plus grandes variances sont relatives aux plus grands échantillons,
  - la valeur F de Fisher-Snedecor diminue et le risque réel est inférieur au risque nominal.
  - ➔valeurs de probabilité son surestimées : on parle de *test conservateur* car il serait difficile de rejeter l'hypothèse nulle.
- En cas d'inégalité des tailles avec un rapport des variances inférieur à 5 et si de plus,
  - les plus grandes variances sont relatives aux plus petits échantillons, le risque réel est supérieur de 400 % au risque nominal.
  - Il y a augmentation de la valeur de F et les valeurs de probabilité sont largement sous-estimées.
  - Le test est alors très *libéral* : il serait facile de rejeter l'hypothèse nulle.

# 4. Conséquences pratiques du non-respect des conditions d'application

## *4.3. Non-normalité et Hétéroscédasticité*

- Lire pages 30-32 du document reçu

# 5. Tests d'hypothèses pour la vérification des conditions d'application

**5.1. Test du caractère aléatoire et simple d'une série de données**

***Principe***

- Caractère aléatoire et simple d'un échantillon ➔
  - tous les individus de la population-parent ont une même probabilité de faire partie de l'échantillon
  - les choix successifs des différents individus qui doivent constituer l'échantillon sont réalisés indépendamment les uns des autres (Dagnelie, 1998).
- Cette condition dépend du matériel observé et de la méthode de collecte de données utilisée. Lorsque les données sont prises suivant un protocole d'échantillonnage adéquat, la condition d'échantillonnage aléatoire et simple est supposée vérifiée. Toutefois, en raison des éventuelles erreurs de mesure ou de prises de données erronées sur le terrain, on peut tester le caractère aléatoire et simple des échantillons à soumettre à une méthode statistique paramétrique. Le non-respect d'une telle condition n'affecte pas l'exécution de la méthode statistique à utiliser mais donne souvent lieu à des conclusions statistiques erronées du fait des résultats biaisés qu'il entraîne.

- Dans Stata, vous pouvez tester la normalité par des méthodes graphiques ou numériques.
- Méthodes graphiques
  - stem-and-leaf plot,
  - scatterplot,
  - box-plot,
  - histogram,
  - probability-probability (P-P) plot,
  - quantile-quantile (Q-Q) plot
- Les premiers comprennent le dessin d'un diagramme à tiges et feuilles, d'un nuage de points, d'un diagramme en boîte, d'un histogramme, d'un diagramme probabilité-probabilité (P-P) et d'un diagramme quantile-quantile (Q-Q). Ces derniers impliquent le calcul des tests Shapiro-Wilk, Shapiro-Francia et Skewness/Kurtosis.