

Chapitre II: Statistiques avec SPSS

Caroline Verhoeven

1 Tests non paramétriques

- Normalité ?
- Exercice
- Test de rangs de Wilcoxon pour 2 échantillons appariés
- Exercices

2 Régression et ANOVA

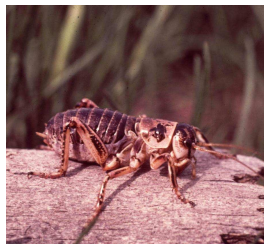
- Introduction
- Régression
- ANOVA
- Exercices

Tester la normalité : Exemple I

Exemple 1

Chez le grillon des sauges (*Cyphoderris strepitans*), durant l'accouplement, la femelle grignote les extrémités des ailes du mâle.

En 1999, Johnson et al. se sont demandé si une femelle affamée aura plus facilement tendance à s'accoupler.



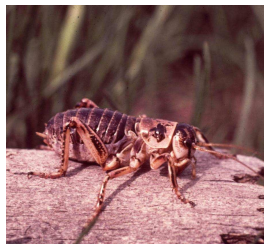
Tester la normalité : Exemple I

Exemple 1

Chez le grillon des sauges (*Cyphoderris strepitans*), durant l'accouplement, la femelle grignote les extrémités des ailes du mâle.

En 1999, Johnson et al. se sont demandé si une femelle affamée aura plus facilement tendance à s'accoupler.

Ils ont pris 24 grillons et ont choisi un groupe de $N_1 = 11$ au hasard qu'ils ont affamé, l'autre groupe de $N_2 = 13$ a été nourri. Après quoi chaque femelle a été mise dans une cage avec 1 mâle, et on a enregistré le temps d'attente pour l'accouplement



Tester la normalité : Exemple I

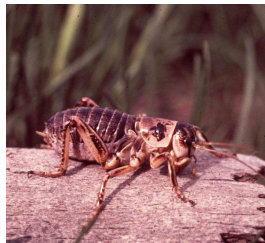
Exemple 1

Chez le grillon des sauges (*Cyphoderris strepitans*), durant l'accouplement, la femelle grignote les extrémités des ailes du mâle.

En 1999, Johnson et al. se sont demandé si une femelle affamée aura plus facilement tendance à s'accoupler.

Ils ont pris 24 grillons et ont choisi un groupe de $N_1 = 11$ au hasard qu'ils ont affamé, l'autre groupe de $N_2 = 13$ a été nourri. Après quoi chaque femelle a été mise dans une cage avec 1 mâle, et on a enregistré le temps d'attente pour l'accouplement

Les mesures se trouvent sur le slide suivant



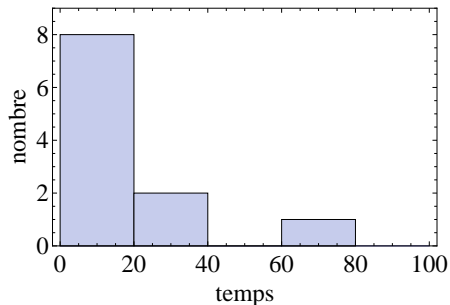
Tester la normalité : Exemple II

Exemple 1

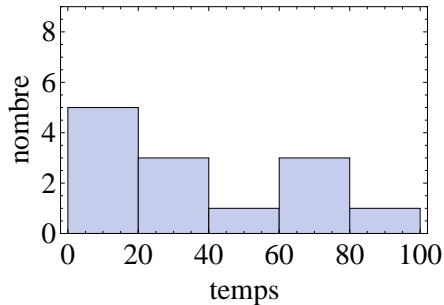
faim	nourri
1,9	1,5
2,1	1,7
3,8	2,4
9,0	3,6
9,6	5,7
13,0	22,6
14,7	22,8
17,9	39,0
21,7	54,4
29,0	72,1
72,3	73,6
	79,5
	88,9

Tester la normalité : Exemple III

Femelles affamées



Femelles nourries



Tester la normalité avec SPSS : Le Q-Q plot I

Le Q-Q plot :

- Compare les quantiles de nos données (centrées et réduites) avec les quantiles (théoriques) de la normale standard.

Tester la normalité avec SPSS : Le Q-Q plot I

Le Q-Q plot :

- Compare les quantiles de nos données (centrées et réduites) avec les quantiles (théoriques) de la normale standard.
- Si les données sont distribuées normalement, les quantiles observés et théoriques sont (approximativement) égaux

Tester la normalité avec SPSS : Le Q-Q plot I

Le Q-Q plot :

- Compare les quantiles de nos données (centrées et réduites) avec les quantiles (théoriques) de la normale standard.
- Si les données sont distribuées normalement, les quantiles observés et théoriques sont (approximativement) égaux
- **SPSS** : Analyze → Descriptive Statistics → Q-Q plots

Tester la normalité avec SPSS : Le Q-Q plot II

Petit film explicatif youtube

Tester la normalité avec SPSS : Le test de Shapiro-Wilk I

Le test de Shapiro-Wilk :

- Test statistique qui détermine si la variable est distribuée normalement
 - H_0 : La variable est distribuée normalement
 - H_a : La variable n'est pas distribuée normalement

Tester la normalité avec SPSS : Le test de Shapiro-Wilk I

Le test de Shapiro-Wilk :

- Test statistique qui détermine si la variable est distribuée normalement
 - H_0 : La variable est distribuée normalement
 - H_a : La variable n'est pas distribuée normalement
- Si $p \leq 0,05$ RH_0 , si $p > 0,05$ NRH_0

Tester la normalité avec SPSS : Le test de Shapiro-Wilk I

Le test de Shapiro-Wilk :

- Test statistique qui détermine si la variable est distribuée normalement
 - H_0 : La variable est distribuée normalement
 - H_a : La variable n'est pas distribuée normalement
- Si $p \leq 0,05$ RH_0 , si $p > 0,05$ NRH_0
- Il faut minimum 7 données

Tester la normalité avec SPSS : Le test de Shapiro-Wilk I

Le test de Shapiro-Wilk :

- Test statistique qui détermine si la variable est distribuée normalement
 - H_0 : La variable est distribuée normalement
 - H_a : La variable n'est pas distribuée normalement
- Si $p \leq 0,05$ RH_0 , si $p > 0,05$ NRH_0
- Il faut minimum 7 données
- En SPSS : Analyze → Descriptive Statistics → Explore

Tester la normalité avec SPSS : Le test de Shapiro-Wilk I

Le test de Shapiro-Wilk :

- Test statistique qui détermine si la variable est distribuée normalement
 - H_0 : La variable est distribuée normalement
 - H_a : La variable n'est pas distribuée normalement
- Si $p \leq 0,05$ RH_0 , si $p > 0,05$ NRH_0
- Il faut minimum 7 données
- En SPSS : Analyze → Descriptive Statistics → Explore
- On peut également y trouver le test de Kolmogorov-Smirnov

Tester la normalité avec SPSS : Le test de Shapiro-Wilk I

Le test de Shapiro-Wilk :

- Test statistique qui détermine si la variable est distribuée normalement
 - H_0 : La variable est distribuée normalement
 - H_a : La variable n'est pas distribuée normalement
- Si $p \leq 0,05$ RH_0 , si $p > 0,05$ NRH_0
- Il faut minimum 7 données
- En SPSS : Analyze → Descriptive Statistics → Explore
- On peut également y trouver le test de Kolmogorov-Smirnov (moins bien)

Tester la normalité avec SPSS : Le test de Shapiro-Wilk II

Petit film explicatif youtube

Exercice 1

- Ouvrir le fichier `grillon.xls` en SPSS
- Faire un Q-Q plot pour la variable “nourrie” et un pour la variable “affamée”
- Faire le test de Shapiro Wilk pour la variable “nourrie” et un pour la variable “affamée”

Ce petit film peut être utile

Test de Mann-Whitney : Principes

- Egalement appelé test de rangs de Wilcoxon
- Equivalent non-paramétrique du test t à 2 échantillons indépendants

Test de Mann-Whitney : Principes

- Egalement appelé test de rangs de Wilcoxon
- Equivalent non-paramétrique du test t à 2 échantillons indépendants
- Formulation des hypothèses
 - $H_0 : \tilde{\mu}_1 = \tilde{\mu}_2$ médianes !
 - $H_a : \tilde{\mu}_1 \neq \tilde{\mu}_2$

Test de Mann-Whitney : Principes

- Egalement appelé test de rangs de Wilcoxon
- Equivalent non-paramétrique du test t à 2 échantillons indépendants
- Formulation des hypothèses
 - $H_0 : \tilde{\mu}_1 = \tilde{\mu}_2$ médianes !
 - $H_a : \tilde{\mu}_1 \neq \tilde{\mu}_2$
- SPSS : Analyze → Nonparametric Tests → Independent Samples

Test de Mann-Whitney : conditions

- Il n'y a pas de conditions sur la distribution de la population
- Les distributions de 2 populations doivent avoir la même forme
- Les 2 échantillons sont aléatoires simples
- Les 2 échantillons sont indépendants

Wilcoxon pour 2 échantillons appariés : Principe

- Également appelé test de Wilcoxon des rangs signés
- Equivalent non-paramétrique du test t pour 2 échantillons appariés

Wilcoxon pour 2 échantillons appariés : Principe

- Également appelé test de Wilcoxon des rangs signés
- Équivalent non-paramétrique du test t pour 2 échantillons appariés
- Hypothèse sur la médiane $\tilde{\delta}$ des différence entre les 2 mesures d'1 paire

Wilcoxon pour 2 échantillons appariés : Principe

- Egalement appelé test de Wilcoxon des rangs signés
- Equivalent non-paramétrique du test t pour 2 échantillons appariés
- Hypothèse sur la médiane $\tilde{\delta}$ des différence entre les 2 mesures d'1 paire
- Formulation des hypothèses
 - $H_0 : \tilde{\delta} = 0$
 - $H_a : \tilde{\delta} \neq 0$

Wilcoxon pour 2 échantillons appariés : Principe

- Egalement appelé test de Wilcoxon des rangs signés
- Equivalent non-paramétrique du test t pour 2 échantillons appariés
- Hypothèse sur la médiane $\tilde{\delta}$ des différence entre les 2 mesures d'1 paire
- Formulation des hypothèses
 - $H_0 : \tilde{\delta} = 0$
 - $H_a : \tilde{\delta} \neq 0$
- SPSS : Analyze → Nonparametric Tests → Related Samples

Exercice 1

Exercice 2

- Reprendre le fichier `grillon.xls`
- Déterminer s'il y a une différence significative entre le temps d'accouplements de femelles nourries et affamées

Exercice 3

En 1994, Régis et Millot ont étudié l'impact de la pratique du poney sur des enfants handicapés mentaux.

Ils ont mesuré le nombre de comportements de stéréotypie (répétition des mêmes geste ou mots) de 6 enfants lors de travaux manuels à leur institut médico-pédagogique et lors de la pratique du poney.

- Ouvrir le fichier `hippother.xls` en SPSS
- Déterminer s'il y a une différence de comportement significative entre l'activité manuelle et l'activité avec les poneys

Exercice II

Exercice 4

Les gens faisant beaucoup de sport ont la réputation d'être plus attirants d'un point de vue sexuel. Pour vérifier cela, on a compté le nombre de partenaires sexuelles qu'on eu des étudiants en sport et des étudiants en bio, sur une année.

- Ouvrir le fichier `sport_vs_bio.xls`
- Tester si le nombre de partenaires sexuelles est une variable distribuée normalement pour les étudiants en bio et les étudiants en sport
- Déterminer s'il y a une différence significative entre le nombre de partenaires sexuelles des étudiants en sport et en bio.

Utilisation de la régression et ANOVA

Dans les études on a

- des données imposées par les chercheurs (prédicteurs)
- des données qui peuvent être observées comme réactions (réponses)

Quand utiliser l' ANOVA et la régression

Modèle	Réponse	Prédicteur
ANOVA à 1 facteur	1 quantitative	1 qualitative
ANOVA à 2 facteur	1 quantitative	2 qualitative
Régression simple	1 quantitative	1 quantitative
Régression multiple	1 quantitative	2 (ou plus) quantitatives
Régression logistique	1 qualitative	1 (ou plus) quantitative

Régression logistique I

- x_j : données quantitatives, prédicteur
- Réponse : données qualitatives avec 2 valeurs possibles
- Pour chaque x_j une proportion p_j de succès

Régression logistique I

- x_j : données quantitatives, prédicteur
- Réponse : données qualitatives avec 2 valeurs possibles
- Pour chaque x_j une proportion p_j de succès

$$\text{logit}(p_j) = \ln \left(\frac{p_j}{1 - p_j} \right)$$

Si il y a un lien linéaire

Régression logistique I

- x_j : données quantitatives, prédicteur
- Réponse : données qualitatives avec 2 valeurs possibles
- Pour chaque x_j une proportion p_j de succès

$$\text{logit}(p_j) = \ln \left(\frac{p_j}{1 - p_j} \right)$$

Si il y a un lien linéaire

Question : comment déterminer la droite

$$\text{logit}(p) = \ln \left(\frac{p}{1 - p} \right) = b_0 + b_1 x$$

Régression logistique I

- x_j : données quantitatives, prédicteur
- Réponse : données qualitatives avec 2 valeurs possibles
- Pour chaque x_j une proportion p_j de succès

$$\text{logit}(p_j) = \ln \left(\frac{p_j}{1 - p_j} \right)$$

Si il y a un lien linéaire

Question : comment déterminer la droite

$$\text{logit}(p) = \ln \left(\frac{p}{1 - p} \right) = b_0 + b_1 x$$

Minimisation au sens des moindres carrés : $\sum_{i=1}^n d_i^2$

d_i : distance entre les mesures et la droite

Régression logistique I

- x_j : données quantitatives, prédicteur
- Réponse : données qualitatives avec 2 valeurs possibles
- Pour chaque x_j une proportion p_j de succès

$$\text{logit}(p_j) = \ln \left(\frac{p_j}{1 - p_j} \right)$$

Si il y a un lien linéaire

Question : comment déterminer la droite

$$\text{logit}(p) = \ln \left(\frac{p}{1 - p} \right) = b_0 + b_1 x$$

Minimisation au sens des moindres carrés : $\sum_{i=1}^n d_i^2$

d_i : distance entre les mesures et la droite

En SPSS : Analyze → Regression → Binary Logistic

Régression logistique II

Petit film explicatif youtube

Régression multiple I

- $X_{j1}, X_{j2}, \dots, X_{jk}$: données quantitatives, prédicteurs
- y_j : données quantitatives, réponse

Régression multiple I

- $X_{j1}, X_{j2}, \dots, X_{jk}$: données quantitatives, prédicteurs
- y_j : données quantitatives, réponse

Régression multiple I

- $x_{j1}, x_{j2}, \dots, x_{jk}$: données quantitatives, prédicteurs
- y_j : données quantitatives, réponse

Question : comment déterminer la droite

$$y = b_0 + b_1x + b_2x_2 + \dots + b_kx_k, \quad b_0? b_1? \dots b_k?$$

Régression multiple I

- $x_{j1}, x_{j2}, \dots, x_{jk}$: données quantitatives, prédicteurs
- y_j : données quantitatives, réponse

Question : comment déterminer la droite

$$y = b_0 + b_1x + b_2x_2 + \dots + b_kx_k, \quad b_0? b_1? \dots b_k?$$

En SPSS : Analyze → Regression → linear

Régression multiple II

Petit film explicatif youtube

ANOVA à 1 facteur

On veut comparer la moyenne de plus de 2 groupes

Exemple : On veut comparer l'effet d'un traitement 1, un traitement 2 et un placebo

- $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$
- H_a : Au moins une des moyennes est différente des autres

ANOVA à 1 facteur

On veut comparer la moyenne de plus de 2 groupes

Exemple : On veut comparer l'effet d'un traitement 1, un traitement 2 et un placebo

- $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$
- H_a : Au moins une des moyennes est différente des autres

Principe :

Comparer le variation des moyennes par groupes au variation dans chaque groupe

ANOVA à 1 facteur : conditions

- Echantillons indépendants
- Les sujets doivent être indépendants
- Distribution normale de la population pour chaque niveau du facteur
- Variances identiques pour chaque niveau du facteur

ANOVA à 1 facteur, mesures répétées

On veut comparer les moyennes pour un même échantillon dans différentes conditions

- $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$
- H_a : Au moins une des moyennes est différente des autres

ANOVA à 1 facteur, mesures répétées

On veut comparer les moyennes pour un même échantillon dans différentes conditions

- $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$
- H_a : Au moins une des moyennes est différente des autres

Conditions :

- Les sujets dans l'échantillon doivent être indépendants
- Sphéricité : la variance doit être la même pour les différences entre 2 groupes, quels que soient les groupes qu'on considère (Test de Mauchly)
- Si on n'a pas la sphéricité, il faut adapter le test

ANOVA à 1 facteur, mesures répétées

On veut comparer les moyennes pour un même échantillon dans différentes conditions

- $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$
- H_a : Au moins une des moyennes est différente des autres

Conditions :

- Les sujets dans l'échantillon doivent être indépendants
- Sphéricité : la variance doit être la même pour les différences entre 2 groupes, quels que soient les groupes qu'on considère (Test de Mauchly)
- Si on n'a pas la sphéricité, il faut adapter le test

En SPSS : Analyze → General Linear Model → Repeated Measures

ANOVA à 2 facteurs

On veut évaluer les effets de 2 facteurs : on compare les moyennes des populations correspondant à toutes les combinaisons de tous les niveaux de chacun de ces 2 facteurs.

Exemple : Etude de la sensibilité aux effets de l'insuline chez des femmes en fonction de leur poids (normal ou surpoids) et selon qu'elles sont hyperthyroïdiennes ou non.

ANOVA à 2 facteurs

On veut évaluer les effets de 2 facteurs : on compare les moyennes des populations correspondant à toutes les combinaisons de tous les niveaux de chacun de ces 2 facteurs.

Exemple : Etude de la sensibilité aux effets de l'insuline chez des femmes en fonction de leur poids (normal ou surpoids) et selon qu'elles sont hyperthyroïdiennes ou non.

Plusieurs facteurs \Rightarrow possibilité d'interaction entre ces facteurs

ANOVA à 2 facteurs

On veut évaluer les effets de 2 facteurs : on compare les moyennes des populations correspondant à toutes les combinaisons de tous les niveaux de chacun de ces 2 facteurs.

Exemple : Etude de la sensibilité aux effets de l'insuline chez des femmes en fonction de leur poids (normal ou surpoids) et selon qu'elles sont hyperthyroïdiennes ou non.

Plusieurs facteurs \Rightarrow possibilité d'interaction entre ces facteurs

Formulation des hypothèses nulles :

- H_0 : Le poids n'a pas d'influence sur la sensibilité à l'insuline
- H_0 : La thyroïde n'a pas d'influence sur la sensibilité à l'insuline
- H_0 : L'effet du poids sur la sensibilité à l'insuline ne dépend pas de la thyroïde

ANOVA à 2 facteurs

On veut évaluer les effets de 2 facteurs : on compare les moyennes des populations correspondant à toutes les combinaisons de tous les niveaux de chacun de ces 2 facteurs.

Exemple : Etude de la sensibilité aux effets de l'insuline chez des femmes en fonction de leur poids (normal ou surpoids) et selon qu'elles sont hyperthyroïdiennes ou non.

Plusieurs facteurs \Rightarrow possibilité d'interaction entre ces facteurs

Formulation des hypothèses nulles :

- H_0 : Le poids n'a pas d'influence sur la sensibilité à l'insuline
- H_0 : La thyroïde n'a pas d'influence sur la sensibilité à l'insuline
- H_0 : L'effet du poids sur la sensibilité à l'insuline ne dépend pas de la thyroïde

En SPSS : Analyze \rightarrow General Linear Model \rightarrow Univariate

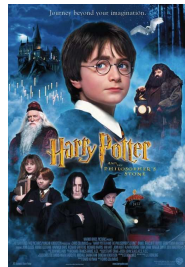
Exercice 5

Ouvrir le fichier `film.xls`.

Déterminer l'équation permettant des recettes d'un film hollywoodien tiré d'un livre sur base de :

- Coûts de production
- Coûts publicitaires
- Recettes du livres

On considère 10 films.



Exercice 6

Walker et al ont étudié le stress chez les manchots de Magellan en 2005. Certains se reproduisent dans une région retirée avec peu d'activités humaine. D'autres se reproduisent dans des régions touristiques.



Exercice 6

Walker et al ont étudié le stress chez les manchots de Magellan en 2005. Certains se reproduisent dans une région retirée avec peu d'activités humaine. D'autres se reproduisent dans des régions touristiques.



On veut savoir si les manchots stress plus en grandissant et si le fait de grandir dans une zone touristique ou non influence le stress.

Exercice 6

Walker et al ont étudié le stress chez les manchots de Magellan en 2005. Certains se reproduisent dans une région retirée avec peu d'activités humaine. D'autres se reproduisent dans des régions touristiques.



On veut savoir si les manchots stress plus en grandissant et si le fait de grandir dans une zone touristique ou non influence le stress.

Pour cela, on les capture et on mesure leur concentration de corticostérone 30 minutes après. On fait cela pour 3 catégories de manchots : récemment éclos, de 40 à 50 jours et juste adultes.

Exercice 6

Walker et al ont étudié le stress chez les manchots de Magellan en 2005. Certains se reproduisent dans une région retirée avec peu d'activités humaine. D'autres se reproduisent dans des régions touristiques.



On veut savoir si les manchots stress plus en grandissant et si le fait de grandir dans une zone touristique ou non influence le stress.

Pour cela, on les capture et on mesure leur concentration de corticostérone 30 minutes après. On fait cela pour 3 catégories de manchots : récemment éclos, de 40 à 50 jours et juste adultes.

Ouvrir le fichier `pinguin.xls`. Déterminer si l'âge des jeunes manchots a une influence sur leur niveau de stress, si le fait de grandir dans une région retirée ou touristique a une influence sur leur niveau de stress et si l'effet de l'âge dépend de l'environnement.

Les tests post-hoc pour l'ANOVA

- LSD : Compare tous les groupes 2 à 2. Grand risque de faire une erreur du type I
- Tukey : Correction par rapport à LSD pour diminuer le risque d'erreur du type I. Fonctionne bien quand les différents groupes ont le même nombre de sujets
- Bonferroni : Le plus simple. basse puissance
- Sidak : Même idée que Bonferroni, mais puissance un peu meilleure
- Scheffe : pas très bonne puissance, mais très populaire
- Dunnett : compare les différents groupes avec un groupe contrôle

Les test post-hoc pour l'ANOVA à mesures répétées

SPSS permet uniquement :

- LSD
- Bonferroni
- Sidak

Exercices intima-média

Ouvrir le fichier `intima_media.xls` en SPSS

Exercice 7 Regarder si l'âge a une influence sur la pratique du sport

Exercice 8 Regarder si l'âge et l'IMC ont une influence sur l'épaisseur de l'intima-média

Exercice 9 Regarder si l'épaisseur de l'intima-média dépend de la consommation d'alcool et de tabac

Exercices enseignement enfants trisomiques I

de Graaf et al ont étudié, en 2013 l'importance du choix de l'enseignement (enseignement spécialisé ou enseignement traditionnel) pour des enfants atteints de trisomie. Ils ont pour cela fait remplir un questionnaire à des parents d'enfants trisomiques suivant les deux types d'enseignements. Ils ont interrogé les parents sur les performances des enfants en lecture, écriture et mathématiques. On a également des données sur l'âge, l'IQ des enfants et le niveau d'étude des parents.

Exercices enseignement enfants trisomiques II

Vous trouverez des données basées sur cette étude dans le fichier `down.xls`

Exercice 10 Y a-t-il une différence entre les performances des enfants suivant l'enseignement spécialisé et l'enseignement traditionnel en lecture, écriture et mathématiques.

Exercice 11 Etudier l'effet du type d'enseignement et du niveau d'études des parents sur les performances de l'enfant en lecture.

Exercice 12 Etudier l'effet de l'IQ, l'âge et du niveau d'étude de la mère sur les performances de l'enfant en lecture, écriture et mathématiques.

Exercice 13 Etudier le lien entre niveau d'études des parents sur le choix de l'enseignement chez l'enfant.