Tests de normalité et de Student avec SAS®

Nicolas Poulin

Institut de Recherche Mathématique Avancée

UFR Mathématique Informatique Strasbourg

Université de Strasbourg

nicolas.poulin@math.unistra.fr

1. Tests d'hypothèses

Lors d'un test statistique, on teste une hypothèse contre une autre.

Hypothéses

- H_0 (hypothèse nulle) : pas de différence significative.
- H_1 (hypothèse alternative) : différence significative.

Remarque :

- si H_1 oriente la différence : test unilatéral,
- sinon : test bilatéral.

nicolas.poulin@math.unistra.fr

Comment ça marche ?

Les tests sont construits grâce à une modélisation probabiliste et des techniques mathématiques.

Statistique de décision :

- une formule en fonction de différents paramétres (moyenne, écart-type,$\cdots$),
- c'est une variable aléatoire,
- lorsqu'on se place sous H_0 la statistique de décision suit une loi de probabilité $\mathscr{L}$ donnée,
- la valeur observée sur l'échantillon peut être calculée,
- si H_0 est vérifiée, la valeur observée doit être une réalisation probable de $\mathscr{L}$.

nicolas.poulin@math.unistra.fr

Seuil α d'un test statistique

α :

- $0 < \alpha < 1$
- est la probabilité se tromper dans son choix. FAUX
- c'est l'erreur de type I, il existe une erreur de type II : β

	accepter H_0	accepter H_1
H_0 vraie	$1-\alpha$	α
H_1 vraie	β	$1-\beta$

$1-\beta$:

- puissance du test,
- plus α est grand, plus $1-\beta$ est grand,
- permet de calculer la taille de l'échantillon à collecter.

Déroulement d'un test

- définir les hypothèses à tester,
- sélectionner le test correspondant,
- choisir le seuil α,
- vérifier les conditions d'application du test,
- calculer la valeur de la statistique de décision,
- comparer cette valeur de référence à la valeur critique au seuil α.

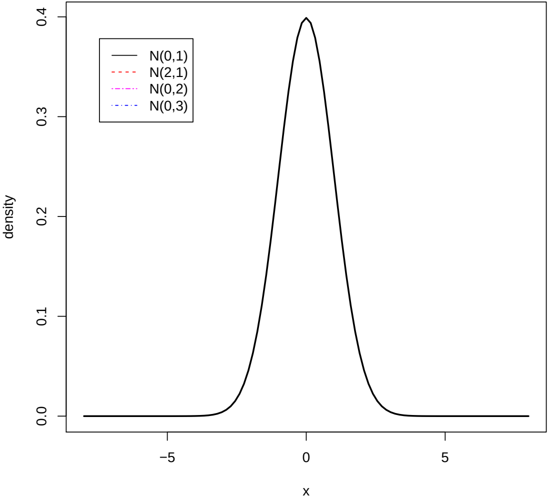
La valeur critique est lue dans des tables.

p-value

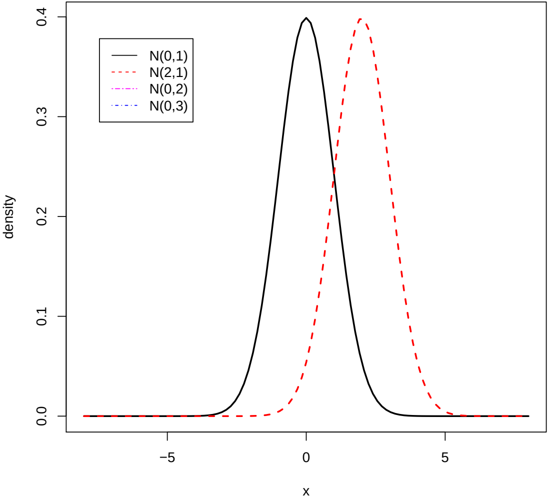
- Valeur critique : obsolète avec les ordinateurs.
- La décision est maintenant prise grâce à la p-value.
- p-value $< \alpha \Longrightarrow H_1$.
- p-value $> \alpha \Longrightarrow H_0$.

nicolas.poulin@math.unistra.fr

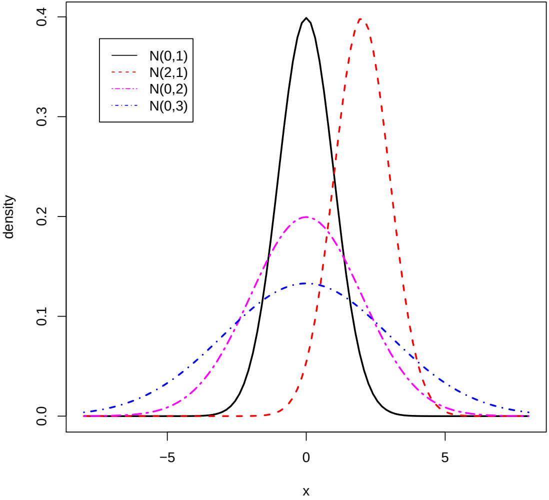
(Simulation de) Lois Normales



(Simulation de) Lois Normales



(Simulation de) Lois Normales



2. Le test de Shapiro-Wilk

Parmi les conditions d'application possibles pour un test d'hypothèse, la normalité des données (ou des résidus) est courante.

Le test de Shapiro-Wilk est un test d'ajustement à une loi normale.

Hypothèses

- H_0 : l'échantillon est issu d'une population normalement distribuée.
- H_1 : l'échantillon n'est pas issu d'une population normalement distribuée.

nicolas.poulin@math.unistra.fr

Unique condition d'application : l'échantillon doit être composé d'observations indépendantes d'une variable quantitative continue.

La décision de rejetter ou non H_0 est basée sur une valeur observée de la statistique de décision.

La décision de rejetter ou non H_0 est prise en comparant le seuil α à la p-value.

nicolas.poulin@math.unistra.fr

Le test de Shapiro-Wilk avec SAS®.

La procédure permettant de réaliser le test de Shapiro-Wilk est PROC UNIVARIATE.

Cette procédure permet de réaliser différents tests d'ajustement :

- test de Shapiro-Wilk
- test de Kolmogorov-Smirnov
- test d'Anderson-Darling
- test de Cramér-von Mises

On peut faire un test de normalité avec chacun de ces tests mais le plus adapté à la normalité est le test de Shapiro-Wilk (qui ne permet de tester que la normalité).

nicolas.poulin@math.unistra.fr

Importation des données

- Télécharger le jeu de données mais depuis le site de Laurent Gardes
- Télécharger le fichier sur le serveur SAS
- Créer une table SAS avec les données

```
PROC IMPORT OUT=mais
            DATAFILE="folders/myfolders/mais.xlsx"
            DBMS= xlsx ;
     GETNAMES=YES ;
RUN ;
```


nicolas.poulin@math.unistra.fr

Le test de Shapiro-Wilk avec SAS®.

```
PROC UNIVARIATE DATA=mais1 NORMAL ;
     var Hauteur ;
RUN ;
```


nicolas.poulin@math.unistra.fr

Le Q-Q plot comme évaluation graphique de la normalité.

Le Q-Q plot est une méthode graphique pour comparer 2 distributions de probabilité en traçant les quantiles de ces distributions.

Si les 2 distributions comparées sont similaires, les points du Q-Q plot seront approximativement sur la droite $y = x$.

Il ne s'agit pas d'un test mais d'une méthode qui permet de voir, notamment si le test de Shapiro-Wilk rejette l'hypothèse de normalité, si la distribution de l'échantillon est éloignée ou non de la normalité.

La majorité des tests qui requièrent la normalité sont relativement robustes à la non-normalité et leurs résultats resteront valables tant que la distribution n'est "pas trop éloignée" de la normalité.

nicolas.poulin@math.unistra.fr

Le Q-Q plot comme évaluation graphique de la normalité.

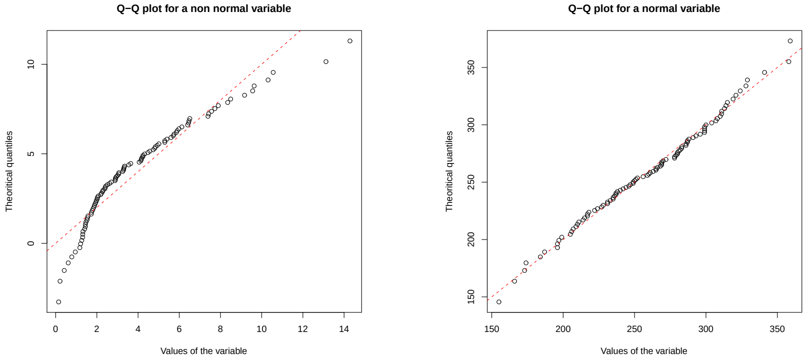


Figure – Exemples typiques de Q-Q plots.

Q-Q plot avec SAS®.

Pour tracer un Q-Q plot avec SAS® on peut utiliser le paramètre QQPLOT de PROC UNIVARIATE.

```
PROC UNIVARIATE DATA=mais1 ;
    VAR Hauteur ;
    QQPLOT Hauteur / NORMAL (MU=EST SIGMA=EST) ;
RUN ;
```

3. Test de Student.

- comparaison de moyennes d'une variable aléatoire quantitative dans différents groupes (variable qualitative).
- 2 groupes $\Longrightarrow$ test de Student.
- $>$2 groupes $\Longrightarrow$ ANOVA.

nicolas.poulin@math.unistra.fr

Contexte du test de Student.

- X variable quantitative continue.
- Y variable qualitative à 2 niveaux (marqueur de groupe).
- On note X_1 et X_2 la variable X pour chacun des groupes.
- μ_1 : moyenne de X dans le groupe 1 (ie de X_1).
- μ_2 : moyenne de X dans le groupe 2 (ie de X_2).
- σ_1 : écart-type de X_1.
- σ_2 : écart-type de X_2.

nicolas.poulin@math.unistra.fr

Test de Student.

Hypothèses :

- $H_0 : \mu_1 = \mu_2$.
- $H_1 : \mu_1 \neq \mu_2$.

Conditions d'application :

1. Chaque échantillon est composé d'observations indépendantes.
2. Les 2 échantillons sont indépendants.
3. $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$.
4. $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$.
5. σ_1 et σ_2 inconnus.
6. $\sigma_1 = \sigma_2$.

Vérification des conditions d'application.

1. Protocole de récolte des données.
2. Protocole de récolte des données.
3. Test de Shapiro-Wilk sur X_1.
4. Test de Shapiro-Wilk sur X_2.
5. En pratique, c'est le cas.
6. Test de Fisher-Snedecor d'égalité de 2 variances.

nicolas.poulin@math.unistra.fr

Que faire si les conditions d'application ne sont pas vérifiées.

1. Pas évident : modèles mixtes ?
2. Si dépendance du type "avant/après" (ie observations sur les mêmes individus) : test de Student apparié.
3. Q-Q plot : test relativement robuste à la non-normalité.
4. Q-Q plot : test relativement robuste à la non-normalité. Si petit échantillon : test non-paramétrique : Mann-Withney.
5. Utiliser le test pour écart-types connus.
6. Utiliser le test de Welch-Satterwaite.

nicolas.poulin@math.unistra.fr

Test de Fisher-Snedecor.

Hypothèses :

- H_0 : $\sigma_1^2 = \sigma_2^2$.
- H_1 : $\sigma_1^2 \neq \sigma_2^2$.

Conditions d'application :

1. Chaque échantillon est composé d'observations indépendantes.
2. Les 2 échantillons sont indépendants.
3. $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$.
4. $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$.

Statistique de décision : $F = \dfrac{S_{X_1}^2}{S_{X_2}^2}$.

Sous H_0, F suit la loi de Fisher de paramètres $n_1 - 1$, $n_2 - 1$.

nicolas.poulin@math.unistra.fr

Test de Fisher-Snedecor sous SAS®.

Ce test est fait automatiquement lorsqu'on fait le test de Student.

Test de Student.

Hypothèses :

- $H_0 : \mu_1 = \mu_2$.
- $H_1 : \mu_1 \neq \mu_2$.

Statistique de décision :

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S^2_{X_1}}{n_1} + \frac{S^2_{X_2}}{n_2}}}$$

Sous H_0, T suit une loi de Student à $n_1 + n_2 - 2$ ddl.

nicolas.poulin@math.unistra.fr

Test sous SAS®.

La procédure permettant de réaliser le test de Student est PROC TTEST.

```
PROC TTEST DATA=mais1 ;
    VAR Hauteur ;
    CLASS Verse_Traitement ;
RUN ;
```


nicolas.poulin@math.unistra.fr

Test sous SAS®.

La PROC TTEST produit automatiquement :

Remarque

- *le test de Fisher d'égalité des variances,*
- *les QQ plots pour les deux groupes.*
- *mais pas les tests de Shapiro-Wilk.*

Test sous SAS®.

Pour faire les tests de Shapiro-Wilk dans chacun des deux groupes :

```
PROC UNIVARIATE DATA=mais1 NORMAL ;
    VAR Hauteur ;
    CLASS Verse_Traitement ;
RUN ;
```

Test pour des données appariées sous SAS®.

Si les données ne sont pas indépendantes mais qu'il s'agit d'une dépendance de type avant/aprés : test de Student apparié.

Les autres conditions d'application du test de Student doivent être vérifiées.

```
PROC TTEST DATA=mais1;
    PAIRED Hauteur*Hauteur_J7;
RUN;
```


nicolas.poulin@math.unistra.fr