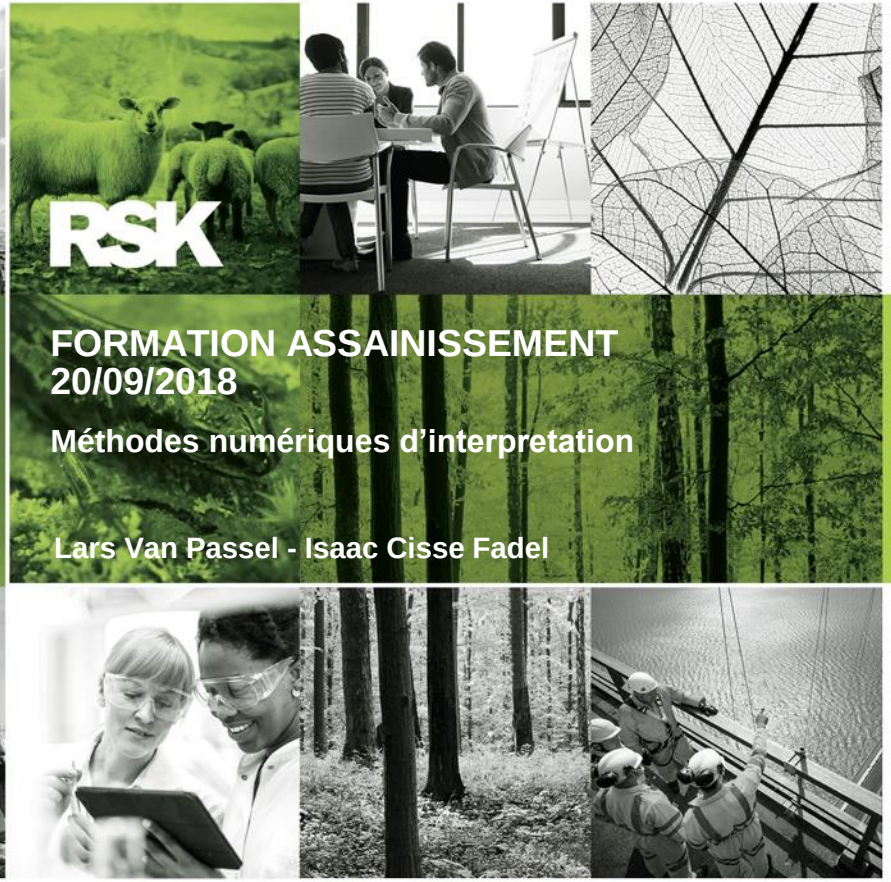


FORMATION ASSAINISSEMENT 20/09/2018

Méthodes numériques d'interpretation

Lars Van Passel - Isaac Cisse Fadel

1. Contenu

Contenu

- Analyse statistique pour les étapes du cycle de vie du projet
- Tests et méthodes statistiques
- Logiciels statistiques

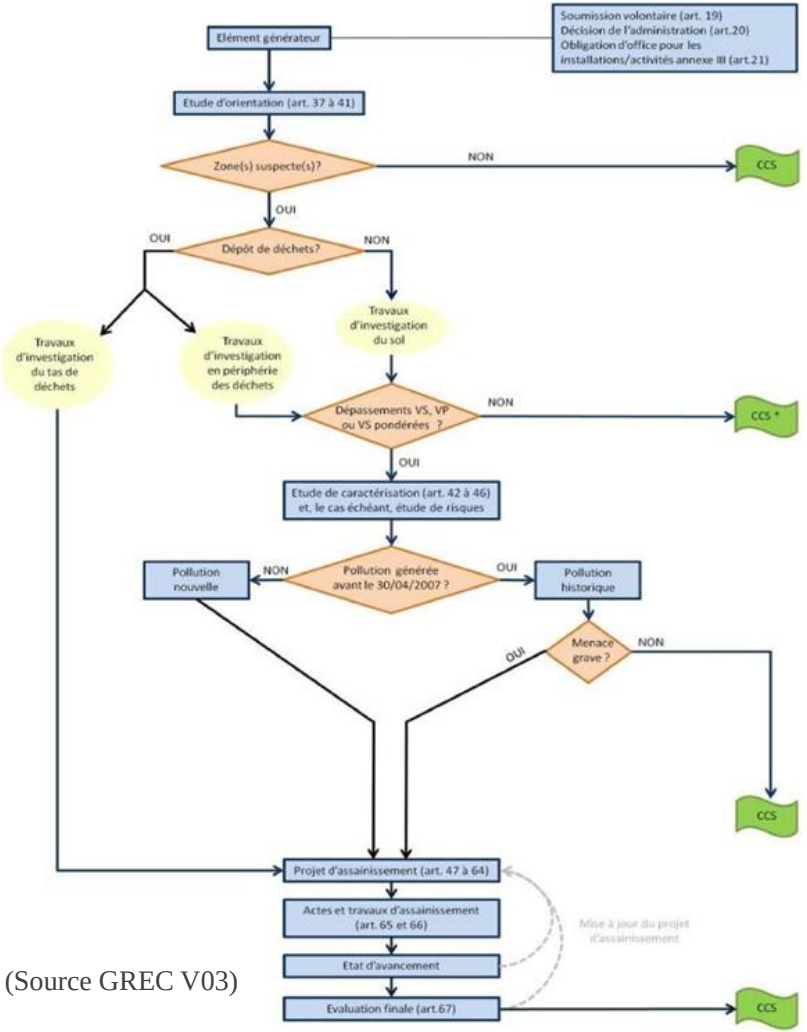
2. Analyse statistique

Pour les étapes du cycle de vie du projet

Les différentes études – Schéma décisionnel

Utilisation des outils de base de statistiques élémentaires

- Etude d'orientation
- Etude de caractérisation
- Etude de risques
- Projet d'assainissement
- Etat d'advancement / Monitoring
- Evaluation finale



(Source GREC V03)

Considérations pour l'analyse statistique

- Analyse exploratoire des données
- Vérification des distributions
- Nombre des échantillons
- Indépendance statistique des données
- Comment traiter non-detects
- …

Etude d’orientation

La statistique peut servir pour:

- Calcul des concentrations de fond
  - Metaux lourds
  - Geochimie de l’eau souterraine
  - Naturel ou anthropique d’origine
- Interprétation des résultats d’analyse suite aux stratégies appliquées
- Définition du modèle conceptuel du site

Methodes applicable:

- Evaluation qualitative des concentrations (Box plot, …)
- Vérification de la stabilité des concentrations de fond (trend tests)
- Interval de confiance de la concentration de fond

Etude de caractérisation (cont.)

Développement continu du modèle conceptuel du site

Questions typique dans cette phase:

- Quelles sont les concentrations de fond?
- Est-ce que les concentrations sont plus grand que les concentrations de fond?
- Est-ce que les concentrations sont au-dessus ou au-dessous une valeur seuil?
- Est-ce que la fréquence d'échantillonage est bonne (optimisation temporelle)?
- Est-ce que le réseau des Pz/F est bonne (optimisation spatiale)?

Etude de caractérisation (cont.)

Methodes statistiques applicable:

- Vérification des trends existants ou absents. (Man-Kendall ou Theil-Sen)
- Comparaison des trend entres Pz ou en différents délais
- Comparaison des données de 2 aquifers
- Vérifiez si des Pz ont une conc plus grand qu'attendu pour un certain pourcentile. Calcul des quartiles supérieurs/limites supérieures pour identifications des zones à investiguer plus.
- Comparaison d'un ensemble de données à un critère (valeur seuil, …) (test de hypothèse)

Etude de caractérisation

(Cadre bien fixé par les différentes stratégies, nombre de sondages, nombre d'échantillons, etc…)

Préambule GREC V03:

- « Bien que fournissant des directives précises, la méthodologie n'a pas la vocation d'enfermer l'expert dans un carcan rigide ou un modèle figé et inflexible. Sur bon nombre d'aspects, elle laisse une place importante au jugement professionnel. Les experts peuvent s'écarter des stratégies d'investigation définies pour autant qu'une justification, étayée par une argumentation de qualité, soit fournie et que la stratégie alternative permette d'obtenir un niveau équivalent dans la qualité de l'information (). »

Caractérisation des remblais pollués (2.2.2.A. Stratégie Car 1. b.1 – Sondages et échantillonnages, Annexe III)

Etude de caractérisation – Extraits du GREC V03 (cont.)

Caractérisation des remblais pollués (2.2.2.A. Stratégie Car 1. b.1 – Sondages et échantillonnages, Annexe III) :

- Définir de manière optimale le plan d'échantillonnage de la ou des phases d'investigation à travers une analyse exploratoire des données
- Infirmer ou confirmer l'absence de taches de pollution en étudiant la répartition spatiale des données
- Etudier la présence d'éventuelles concentrations atypiques qui pourraient provenir d'une source de pollution non détectée durant l'EO ou d'un(e) horizon/zone de remblais pollués de nature différente particulièrement pollué(e)
- Diminuer le nombre de variables (donc d'analyses) en recherchant des relations qui lient certains polluants à d'autres par des analyses statistiques multivariées (analyses en composantes principales, matrices d'autocorrélations, ...).

Etude de caractérisation – Extraits du GREC V03 (cont.)

Caractérisation des remblais pollués (2.2.2.A. Stratégie Car 1. b.1 – Sondages et échantillonnages, Annexe III) :

- Comparer les niveaux de concentrations mesurés dans les remblais étudiés avec ceux issus de remblais analogues (de même nature) et ayant déjà fait l'objet d'investigations similaires. La cohérence entre les niveaux de concentrations contribue, d'une part, à valider le modèle conceptuel du site et la fiabilité de l'historique et d'autre part, à rendre l'évaluation des risques plus robuste
- Intégrer les résultats analytiques de l'étude d'orientation et d'éventuelles autres études antérieures. !!! Important de justifier au préalable si ces campagnes peuvent être raisonnablement comparées : similarité des niveaux échantillonnés et des protocoles d'échantillonnage et d'analyse. Des protocoles d'investigations, de prélèvements et d'analyses différents, ainsi que des supports d'échantillonnage variables peuvent conduire à des résultats d'investigations peu comparables

Etude de caractérisation – Extraits du GREC V03 (cont.)

Caractérisation des remblais pollués (2.2.2.A. Stratégie Car 1. b.1 – Sondages et échantillonnages, Annexe III) :

- Obtenir un jeu de données plus complet qui peut être soumis aux tests et calculs statistiques opérés en phase exploratoire. Cette synthèse statistique est, selon les cas, réalisée sur l'ensemble de la zone de remblais pollués ou séparément sur tout horizon ou sous-zone particulier que l'analyse exploratoire a mis en évidence.
- Caractériser la population des concentrations du polluant dans le remblai par quelques grandeurs statistiques (concentration moyenne ou médiane, dispersion autour de la moyenne (écart-type), concentrations maximales, quantiles extrêmes (quantiles à 5 % et 95 % (intervalle de confiance à 90 %) voire à 2,5 % et 97,5 % (intervalle de confiance à 95 %) !!! Jeu de données supérieur à 25.
- Définir en fonction du nombre d'analyses disponibles, une valeur représentative du "centre de la distribution" (le "niveau moyen de pollution") ainsi qu'une valeur représentative des "concentrations extrêmes" (le "niveau de pollution maximal")

Etude de caractérisation – Extraits du GREC V03 (cont.)

Caractérisation des remblais pollués (2.2.2.A. Stratégie Car 1. b.1 – Sondages et échantillonnages, Annexe III) :

- Prise en compte de l'incertitude associée à la concentration retenue lors de la comparaison aux normes
- Réalisation de tests statistiques afin de déterminer un nombre d'échantillons minimum pour atteindre un objectif fixé
- Valider l'estimation d'une valeur moyenne ou d'un quantile
- Comparer plusieurs distributions
- En cas d'étude de risques, réalisation d'une analyse de sensibilité de l'évaluation des risques au choix de concentration (scénario optimiste / pessimiste);
- etc.…

Etude de caractérisation – Extraits du GREC V03 (cont.)

Caractérisation des tâches de pollution (2.2.2.B. Stratégie Car 2 – délimitation – Méthodes géostatistiques):

- Sous certaines conditions (très grande tache de pollution, set de données analytique dense, terrain homogène, pollution distribuée uniformément,...), la géostatistique peut constituer un outil intéressant de cartographie des taches de pollution (traçage des courbes d'isoconcentration). L'expert peut en outre utiliser la géostatistique, en fonction des besoins du demandeur, pour affiner les estimations et en particulier les agrémenter de mesures de l'incertitude. La géostatistique peut notamment fournir :
  - la probabilité de dépasser les valeurs seuils pour chaque polluant, en tout point 3D du terrain, ;
  - une classification multi-polluants des zones à dépolluer, délimitées par rapport aux valeurs seuils choisies (ex. métaux lourds, HAP, benzène, etc.) ;
  - une estimation des volumes de sols à dépolluer, avec une incertitude sur cette estimation, permettant ainsi de définir des scénarios de coût de dépollution optimistes et pessimistes ;
  - une cartographie donnant les positions les plus probables des volumes de sols à dépolluer, mais aussi des volumes de sols les plus probablement non pollués.

Assainissement - Monitoring

Sélection des variantes

- Seulement pour atténuation naturelle – analyse des trends

Effectivité de l'assainissement

- Développement des contours (géostatistique)
- Concentrations par rapport à le temps (analyse des trends temporelles)
- Concentrations par rapport à la distance de la source (analyse des trends spatiales)

Assainissement – Monitoring (cont.)

Methodes statistiques applicable:

- Vérification des tendances existantes ou absentes. (analyse de tendance temporelle Man-Kendall ou Theil-Sen)
- Estimation du taux d'atténuation. L’interval de confiance aide pour évaluer l’incertitude
- Evaluation de la surface impactée par l’assainissement par identification des Pz avec taux d'atténuation élevés
- Estimation des concentrations futures en utilisant les concentrations actuelles et les taux d'atténuation estimés
- Tests statistiques de comparaison pour évaluer les différences entres groups de Pz selon leur endroit

Evaluation finale

Modèle conceptuel du site est complète

Methodes statistiques applicable:

- Comparaison de la concentration avec un critère fixe défini dans le PA: comparaison de l'interval de confiance avec le critère fixe
- Comparaison de la limite supérieure de l'interval de confiance (UCL) avec un critère fixe
- Comparaison des concentrations avec concentrations de fond
- Si les concentrations changent encore, vérification des trends existants avec une bande de confiance autour le trend. (Man-Kendall ou Theil-Sen)
- Evaluation de la surface impactée par l'assainissement par identification des Pz avec taux d'atténuation élevés
- Estimation des concentrations futures en utilisant les concentrations actuelles et les taux d'atténuation estimés

Autres aspects

Les outils statistiques peuvent aussi être envisagés pour des cas ponctuels :

- Traitement des non-detect
- Traitement des outliers
- Comparaison des résultats de laboratoires
- Mise en évidence de l'effet d'un traitement ou de la modification des paramètres d'une installation de traitement sur les résultats d'analyses obtenus
- etc …

L'outil statistique peut participer à tous les niveaux à affiner la compréhension, l'analyse et l'interprétation qu'on a des données

3. Tests et méthodes statistiques

Statistiques élémentaires

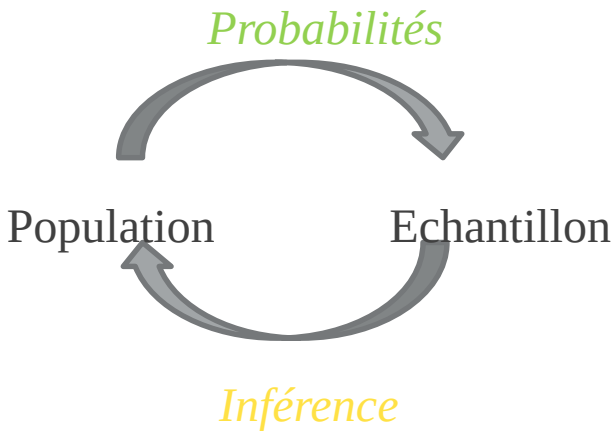
Objectifs

Population → Paramètres

Echantillon → Statistiques

But : Inférer paramètres population sur base statistiques échantillon(s)

Outils → Lois de probabilités



(Source: Probability & statistics for engineers & scientists, 2012)

Echantillon

Population : Ensemble d'individus ou objets qui réalisent un événement

Echantillon : Sous – ensemble de la population → sous - ensemble d'individus qui réalisent un événement
← Echantillon aléatoire

Variable aléatoire : Nombre qui associe à chaque réalisation d'un évènement une valeur numérique

Individu de l'échantillon → Variable aléatoire

Echantillon (ensemble de variables aléatoires !)

Exemple : Ensemble des concentrations mesurées en différents points ou différents jours en un ou plusieurs points.

Statistiques descriptives

Statistiques descriptives

Caractériser l’échantillon → définir ses statistiques (résumé)

Position

- Moyenne

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

- Médiane

$$0{,}5(n+1) \rightarrow M$$

- Variance – Ecart type

$$s^2 = \frac{1}{n-1}\sum_{i}^{n} (x_i - \bar{x})^2$$

- Quartiles

$$0{,}25(n+1) \rightarrow 1^{\text{er}}\ \text{Quartile}$$
$$0{,}75(n+1) \rightarrow 3^{\text{ème}}\ \text{Quartile}$$

- Percentiles (p)

$$0, p(n+1) \rightarrow \text{percentile p}$$

RSK

Exemple

Statistique	Urea Content 1 (mg/l)	Ammonium 1 NH4 (mg/l)	Nitrite NO2- 1 (mg/l)	Nitrate NO3 1 (mg/l)	Total Nitrogen 1 (mg/l)	Mineral oil C10-C40 1 (mg/l)
Nb. d'observations	54	54	54	54	54	54
Minimum	0.1	0.69	0.25	3.1	2.7	0.37
Maximum	54	94	24	230	110	87
Amplitude	53.9	93.31	23.75	226.9	107.3	86.63
1er Quartile	1.95	8	2.9	27.25	18.25	1.1
Médiane	4.65	18	6.95	67	34	2.85
3ème Quartile	12	30.75	10.75	107.5	51.5	5.65
Moyenne	9.403	23.336	7.518	71.996	37.706	8.153
Variance (n-1)	146.72	425.882	34.525	2618.705	575.918	315.338
Ecart-type (n-1)	12.113	20.637	5.876	51.173	23.998	17.758
Ecart-type de la moyenne	1.648	2.808	0.8	6.964	3.266	2.417

Visualiser les données

- Boxplot

  (permet de visualiser graphiquement la répartition des données, les outliers, la position de la médiane, de la moyenne, permet de comparer plusieurs échantillons, ….)

- Steam-leaf plots

  (Permet à la fois de visualiser les données mais aussi d'avoir une idée sur la forme de leur distribution, de visualiser les modes, permet de reconstruire l'échantillon …)

- Histogrammes

  (résume l'information sur le jeu de données, permet de visualiser la forme et les paramètres de position du jeu de données, la dispersion, l'assymétrie, la présence d'outliers, les modes. Elle permet de se faire une idée de la distribution que suive les données de l'échantillon, variantes)

- Scatter plots (données multivariées)

  (révèle la présence de relations ou d'associations structurées entre plusieurs variables (linéaire, quadratique, exponentielle etc). Ces relations se manifestent par des tendances ou des motifs dans la représentation, …)! Association et non forcément causalité

- Diagramme Q-Q

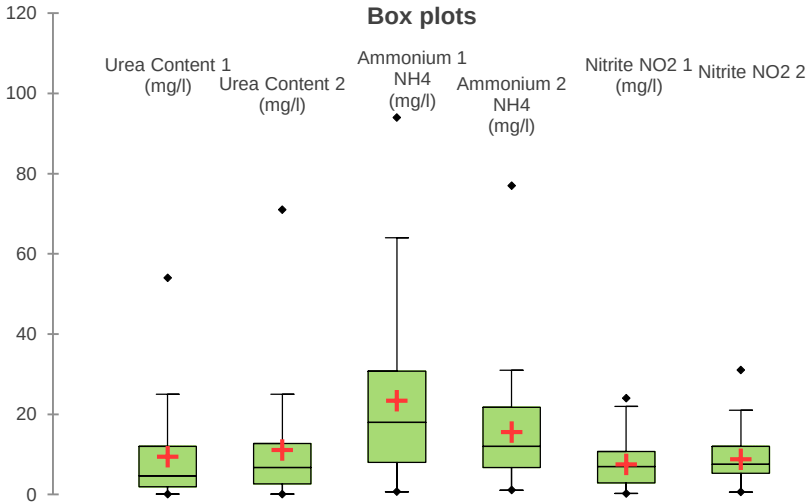
  (permet de savoir si une distribution suit une loi normale – Normal probability plot, ou si deux jeux de données proviennent d'une même population ou du moins de populations qui suivent la même loi de distribution …)

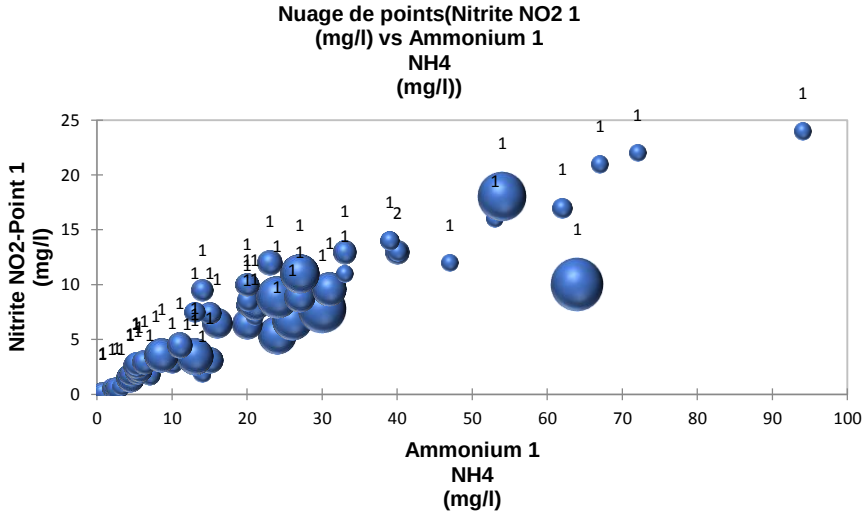
- Lags plots (séries temporelles)

  (permet de voir si un jeu de données, notamment une série temporelle est aléatoire ou non. Permet de se faire une idée du type de model que l'on pourrait utiliser (lag plot linéaire suggère série autoregressive)

  Autres outils de visualisations des données (standard deviation plot, Bi-histogramme, youden plot, …)

  (Voir EDA analysis - Nist)

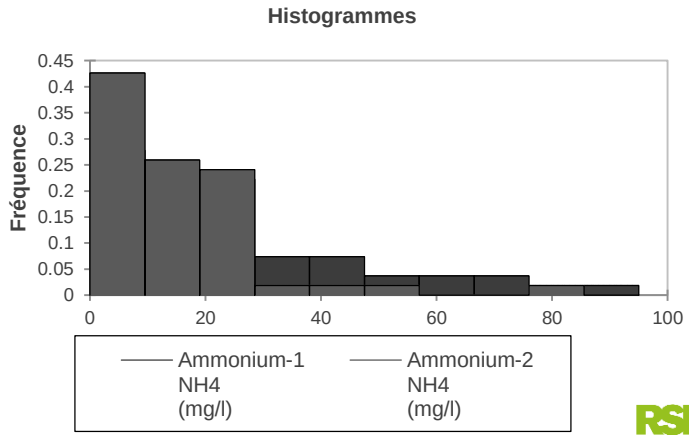




Représentation Stem-and-leaf (Urea Content 1) (mg/l)) :

Unité : 10

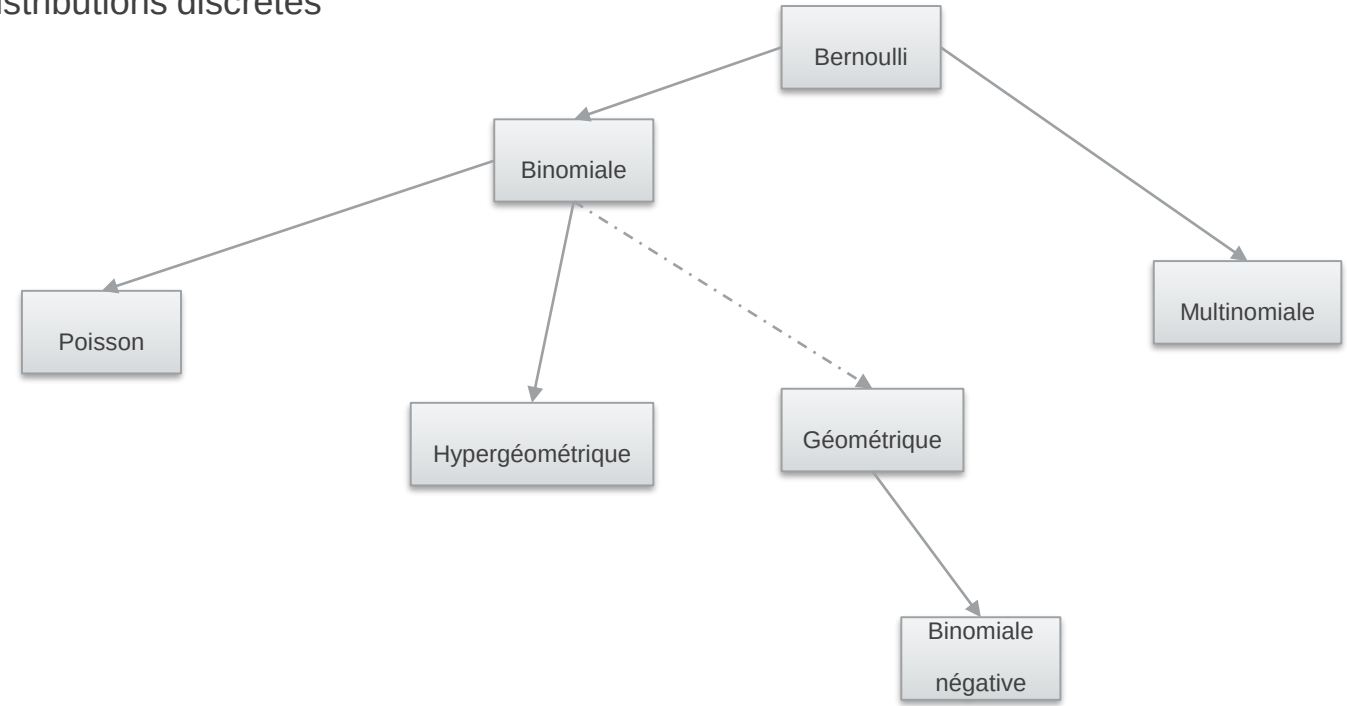
```
0 0000000000012223333333444455556667777 8
1 12234588
2 2558
3 1
4 35
5 4
```



Distributions

Quelques distributions remarquables

Distributions discrètes



Quelques distributions continues

Uniforme

U(a, b)

T-student

t_{n-1}

Normale

N(μ,σ)

Log – Normale

LogN(μ,σ)

Exponentielle

Exp(λ)

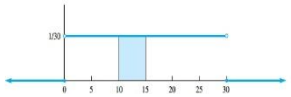
Weibull

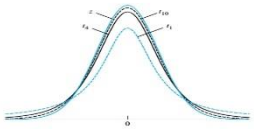
Gamma

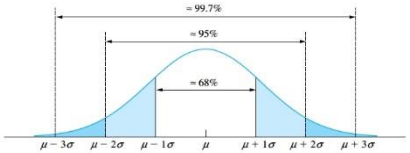
Γ(r,λ)

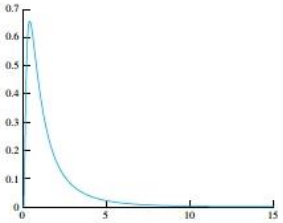
Chi – 2

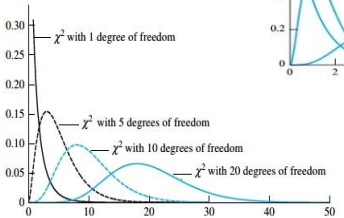
χ^2_{n-1}

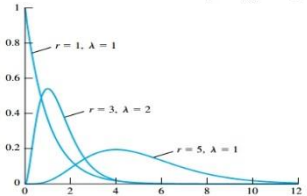


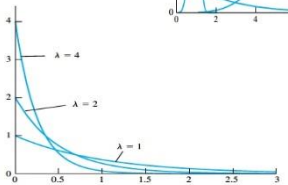


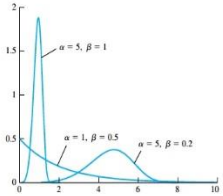












Intervalles de confiance et Tests d’hypothèses

Statistiques → IC / Tests d'hypothèses

Statistique z - Z-score

$$z = \frac{(\overline{X} - \mu_x)}{\sigma_x} \sim N(0,1)$$

Statistique t-student

$$t = \frac{(\overline{X} - \mu_x)}{\sigma_x} \sim t_{n-1}$$

Statistique χ^2

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Exactitude et précision d'un estimateur

Collecte des données → estimer les paramètres de la population dont l'échantillon est issu :

Moyenne – Variance / ecart type

Exemple :

- $\bar{X}$ est un estimateur de la moyenne μ de la population
- S^2 est un estimateur de la variance σ^2 de la population

2 grandeurs

- Exactitude → biais (Ecart au carré de la moyenne de l'estimateur à celle de la population)
- Précision → incertitude (Ecart type)
- Mean Squarred Error $MSE_{\hat{\theta}} = (\mu_{\hat{\theta}} - \theta)^2 + {\sigma_{\hat{\theta}}}^2$

RSK

Théorème central limite

Si X_1, ….., X_n sont des variables aléatoires issues d'une population de moyenne μ et de variance σ²

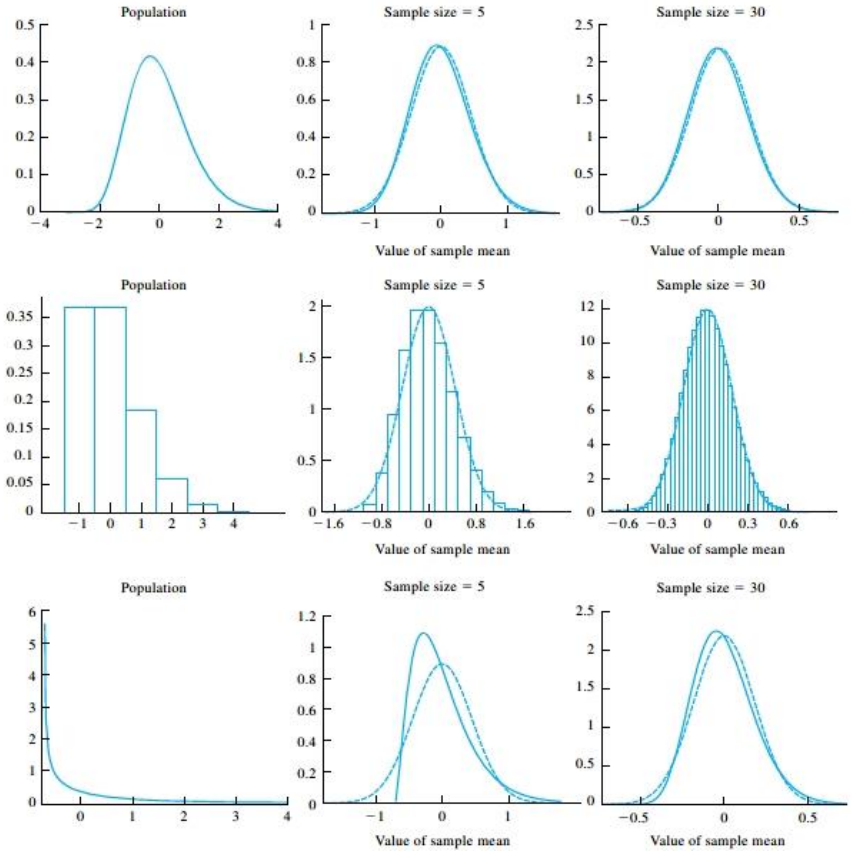
Soit $\bar{X} = \frac{1}{n}(X_1 + X_2 + \ ... + X_n)$

Si n est suffisamment grand, alors :

$\bar{X}$ suit une loi $N\left(\mu, \frac{\sigma^2}{n}\right)$

S_n suit une loi $N(n\mu, n\sigma^2)$

Pour la plupart des populations si n > 25, l'approximation normale de la distribution de probabilités de la moyenne peut être utilisée.


Population
Sample size = 5
Sample size = 30
Value of sample mean
Value of sample mean
Population
Sample size = 5
Sample size = 30
Value of sample mean
Value of sample mean
Population
Sample size = 5
Sample size = 30
Value of sample mean
Value of sample mean

Intervalles de confiance (méthodes paramétriques)

But : définir à un niveau de confiance que l'on se donne, un intervalle susceptible de contenir un paramètre de la population comme par exemple, la moyenne, la variance, une proportion, ….

Intervalle de confiance pour la moyenne

Hypothèse : Taille n > 25 → TCL permet de construire l'intervalle

Statistique : Z –score

Hypothèse : Taille n < 25 et V. a suit une distribution normale

Statistique : T-student

Intervalle de confiance pour une proportion (basé sur l'approximation normale de la loi binomiale)

Hypothèses : Nombre d'échantillons succès > 10 et nombre d'échantillons échec > 10

Statistique : Z-score !!! X → X+2 et n → n + 4

Idem pour l'intervalle de confiance pour la différence entre 2 moyennes / différence entre deux proportions

!!! Comparaison de deux moyennes → Hypothèses sur les variances → Test de Welsh

Intervalle de confiance pour des données appariées

- Hypothèse : $n > 25$
- Statistique : Z – score
- Hypothèse : $n < 25$ et différence entre V.a suit une loi normale
- Statistique : t – student

Intervalles de confiance pour la variance /écart type

- Hypothèse : population suivent une loi normale
- Statistique : X^2

- Intervalle de confiance distribution Log-Normal ou présentant assymétrie droite (Methode de Land)

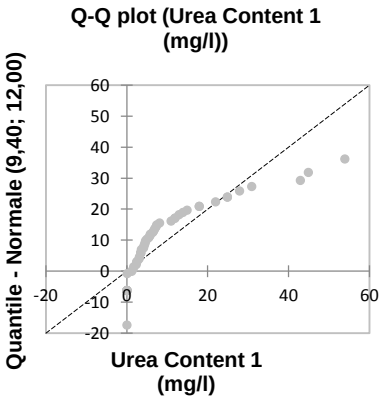
Intervalles de confiance (méthodes non paramétriques)

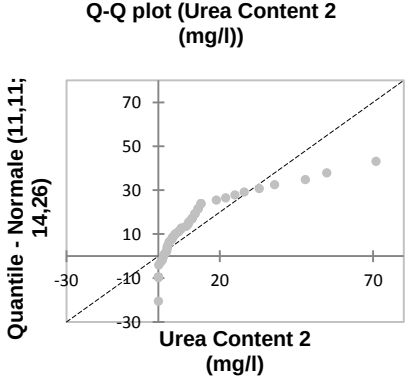
Si les hypothèses précédentes ne sont pas vérifiées :

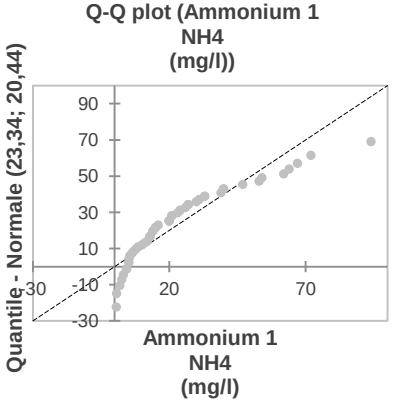
- Distribution sous jacente inconnue, transformation des données inefficace
  - calcul des intervalles de confiance autour de la médiane, proportions, etc….
  - Nécessite cependant un nombre plus important d'échantillons pour un même niveau de précision
- Plus grande flexibilité à définir des intervalles de confiances sans nécessité de transformer les données ou de faire des hypothèses sur la distribution de probabilités de la population !!!!.

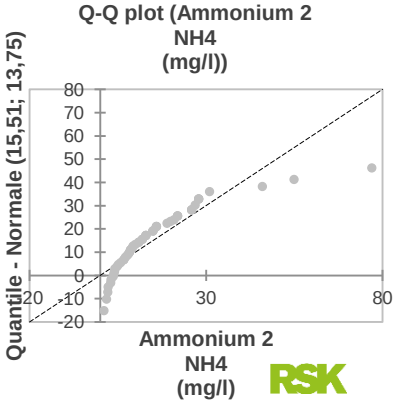
Exemple

Statistique	Urea Content 1 (mg/l)	Urea Content 2 (mg/l)	Ammonium 1 NH4 (mg/l)	Ammonium 2 NH4 (mg/l)
Nb. d'observations	54	54	54	54
Moyenne	9.403	11.114	23.336	15.515
Variance (n-1)	146.720	207.063	425.882	192.716
Ecart-type (n-1)	12.113	14.390	20.637	13.882
Ecart-type de la moyenne	1.648	1.958	2.808	1.889
Borne inf. de la moyenne (95%)	6.097	7.186	17.703	11.726
Borne sup. de la moyenne (95%)	12.709	15.042	28.969	19.304









Tests d'hypothèses

But : Déterminer à quel point l'on peut être certain d'une hypothèse concernant la valeur d'un paramètre d'une population (moyenne, variance etc…) ou la comparaison des paramètres de plusieurs populations

Démarche générale :

Hypothèse nulle : H0 (exemple $\mu \geq \mu_0$)

Hypothèse alternative : Ha ($\mu < \mu_0$)

On construit la distribution de probabilité à partir de l'hypothèse nulle et on calcule la probabilité que l'hypothèse alternative soit avérée. Cette probabilité mesure le degré de plausibilité de l'hypothèse nulle (P-Value)

Si P-Value < α (niveau de signification) →, on rejette l'hypothèse nulle

Si P-value > α → on ne peut pas rejeter l'hypothèse H0

RSK

Familles de tests d'hypothèses

Tests paramétriques

Raisonnement similaires aux intervalles de confiance

TCL si n> 25 ou hypothèse de normalité de la population → (Z-score ou t-student)

- Test sur la valeur de la moyenne d'une population
- Test sur la différence des moyennes

Tests sur la différence de moyenne de plusieurs populations k > 2 (Tests de Dunet)

Test sur la proportion d'une population

- Hypothèse (np_0> 10, $n(1-p_0)$> 10)
- Statistique : z –score

Test sur la différence entre deux proportions !! (calcul d'une proportion commune)

Test sur la variance

- Hypothèse : Populations suivent une loi de distribution normale
- Statistique : $F_{m-1,n-1}$ rapport des variances suit un loi de Fischer

Test de Bartlett (Hypothèse de normalité des populations), **Test de Levene** (Hypothèse de normalité aussi mais moins sensible aux écarts à la normalité).

Test sur la moyenne des différences de données appariées

- Si échantillon n > 25 → Z score
- Si n < 25 et différences suivent une loi normale → t-student

RSK

Exemple

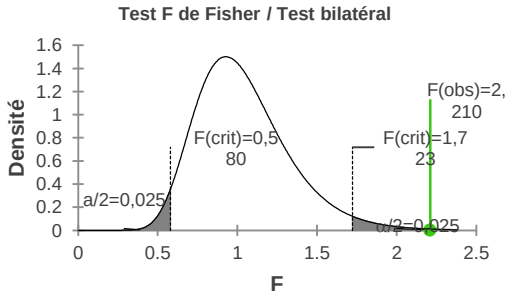
Variable	Observations	Obs. avec données manquantes	Obs. sans données manquantes	Minimum	Maximum	Moyenne	Ecart-type
Ammonium 1 NH4 (mg/l)	54	0	54	0.690	94.000	23.336	20.637
Ammonium 2 NH4 (mg/l)	54	0	54	1.100	77.000	15.515	13.882

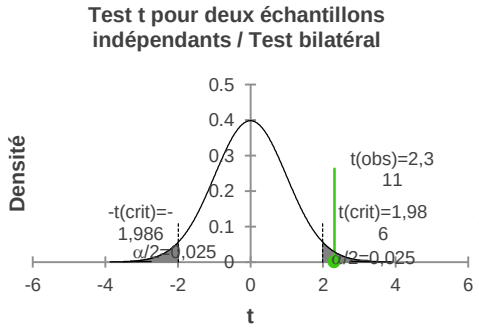
Test F de Fisher / Test bilatéral :

Intervalle de confiance à 95% autour du rapport des variances :

[1.282;3.808]

Rapport	2.210
F (Valeur observée)	2.210
|F| (Valeur critique)	1.723
DDL1	53
DDL2	53
p-value (bilatérale)	0.005
alpha	0.05





Test t pour deux échantillons indépendants / Test bilatéral :

Intervalle de confiance à 95% autour de la différence des moyennes :

[1.100;14.542]

Différence	7.821
t (Valeur observée)	2.311
|t| (Valeur critique)	1.986
DDL	92.814
p-value (bilatérale)	0.023
alpha	0.05

Le nombre de degrés de liberté est calculé en utilisant la formule de Welch-Satterthwaite

Tests non paramétriques

Si hypothèse requis pour appliquer les tests paramétriques ne sont pas respectées, on a la possibilité d'utiliser des tests non paramétriques. Ils sont en général basés sur les rangs.

Test des rangs signés de Wilcoxon

- Permet de test la moyenne d'une population
- Hypothèse : Distribution symétrique, Variables aléatoires continues
- Statistique : $S_+ ou S_-$
- Tables disponibles
- Si n > 25 TCL la statistique S_+ suit une loi normale

Test de la somme des rangs de Wilcoxon ou test de Wilcoxon-Mann

- Permet de tester la différence de moyenne entre deux populations
- Hypothèse : Distributions ont même forme → même répartition et même étendue
- Statistique : W = Somme des rangs après classement des valeurs

Tests sur la différence de moyenne ou de médianes de plusieurs populations (Tests de Fligner Wolfe, test de Kruskal –Wallis, …)

Exemple

Variable	Observations	Obs. avec données manquantes	Obs. sans données manquantes	Minimum	Maximum	Moyenne	Ecart-type
Ammonium 1 NH4 (mg/l)	54	0	54	0.690	94.000	23.336	20.637
Ammonium 2 NH4 (mg/l)	54	0	54	1.100	77.000	15.515	13.882

Test de Mann-Whitney / Test bilatéral :

U	1777
U (normalisé)	1.958
Espérance	1458.000
Variance (U)	26467.318
p-value (bilatérale)	0.050
alpha	0.05

Transformations de variables

Transformations des variables

Beaucoup de ces tests, excepté les tests non paramétriques requièrent que la distribution sous jacente sois normale ou au moins connue. Dans le cas où la distribution ne suit pas une loi normale, il est possible de procéder à une transformation des données.

Standardiser les données (centrer – réduire)

Si on connait la loi de distribution qui régit la fonction de densité de la V.a, il existe des méthodes qui permettent de normaliser la distribution : Exemple (**transformation log** → Log-Normal**, transformation racine carrée** pour une loi Gamma**, transformations de Johnson**, etc.…)

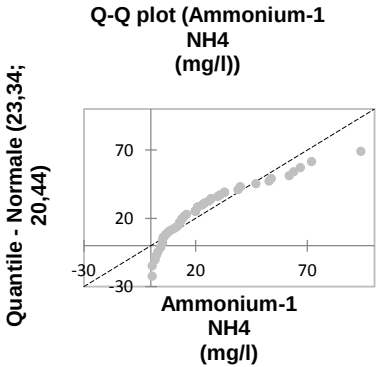
Transformations puissance : Une des plus connues **Box-Cox**

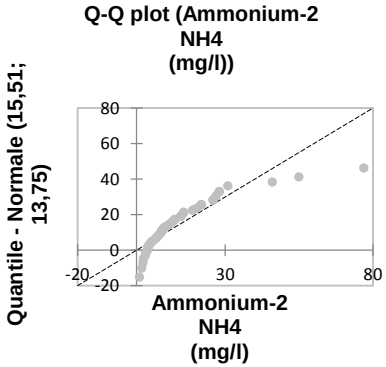
$$g(C_k) = \begin{cases} \log(C_k + k_2) & si\ k_1 = 0 \\ \dfrac{(C_k + k_2)^{k_1} - 1}{k_1} & si\ \ k_1 \neq 0 \end{cases}$$

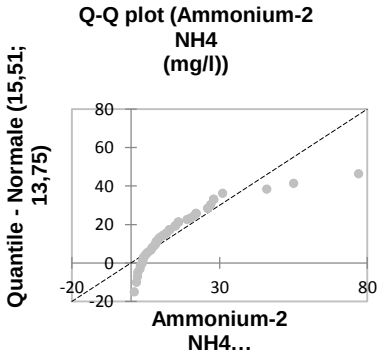
C_k , concentrations mesurées aux points 1B et 17

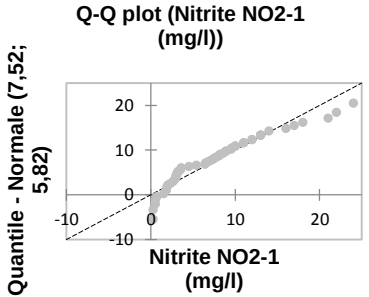
k_1,k_2, paramètres de transformations ; $k_2 = 0$ (les mesures de concentrations sont strictement positives). Le paramètre k_1, est optimisé afin de maximiser la normalité des données transformées

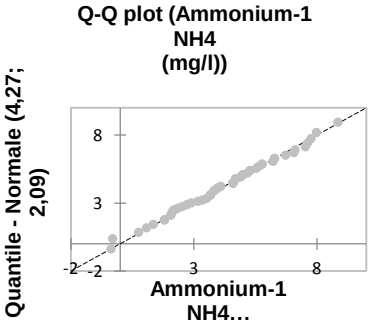
Exemple

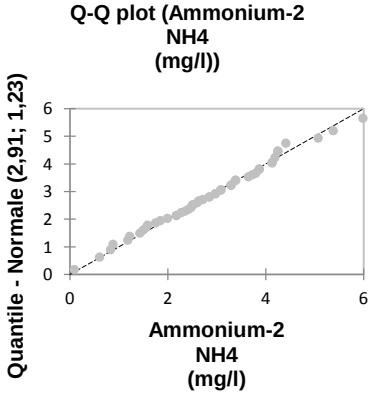

Q-Q plot (Ammonium-1 NH4 (mg/l))
Quantile - Normale (23,34; 20,44)
70
20
-30
-30
20
70
Ammonium-1 NH4 (mg/l)

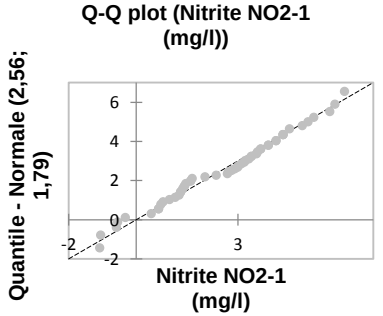

Q-Q plot (Ammonium-2 NH4 (mg/l))
Quantile - Normale (15,51; 13,75)
80
60
40
20
0
-20
-20
30
80
Ammonium-2 NH4 (mg/l)

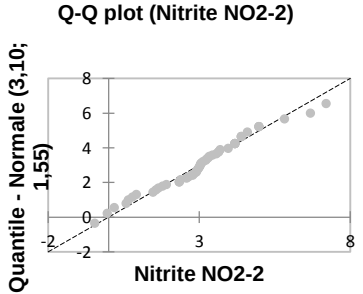

Q-Q plot (Ammonium-2 NH4 (mg/l))
Quantile - Normale (15,51; 13,75)
80
60
40
20
0
-20
-20
30
80
Ammonium-2 NH4…


Q-Q plot (Nitrite NO2-1 (mg/l))
Quantile - Normale (7,52; 5,82)
20
10
0
-10
-10
0
10
20
Nitrite NO2-1 (mg/l)

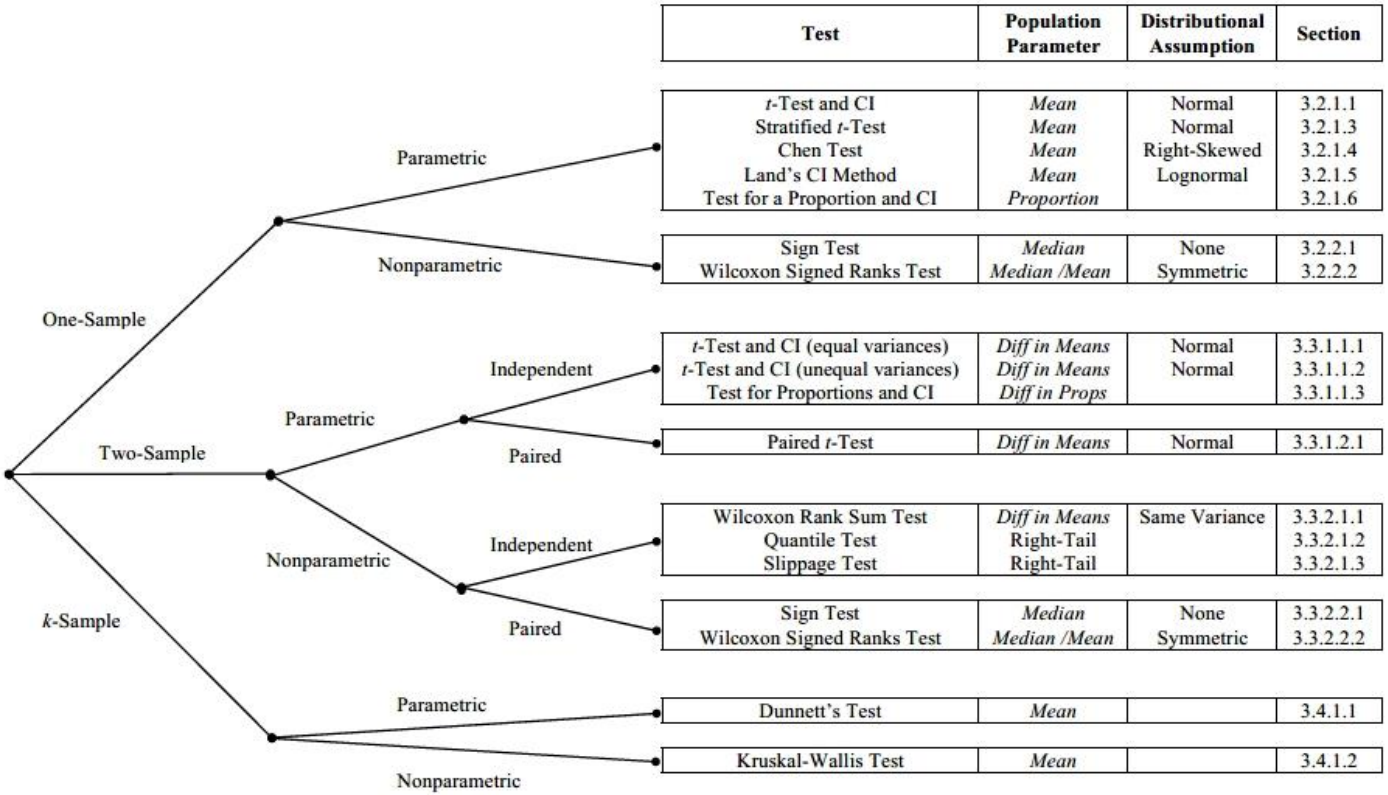

Q-Q plot (Ammonium-1 NH4 (mg/l))
Quantile - Normale (4,27; 2,09)
8
3
-2
-2
3
8
Ammonium-1 NH4…


Q-Q plot (Ammonium-2 NH4 (mg/l))
Quantile - Normale (2,91; 1,23)
6
5
4
3
2
1
0
0
2
4
6
Ammonium-2 NH4 (mg/l)


Q-Q plot (Nitrite NO2-1 (mg/l))
Quantile - Normale (2,56; 1,79)
6
4
2
0
-2
-2
3
Nitrite NO2-1 (mg/l)


Q-Q plot (Nitrite NO2-2)
Quantile - Normale (3,10; 1,55)
8
6
4
2
0
-2
-2
3
8
Nitrite NO2-2

Decision Tree for Selecting the Specific Method



Test	Population Parameter	Distributional Assumption	Section
t-Test and CI	*Mean*	Normal	3.2.1.1
Stratified *t*-Test	*Mean*	Normal	3.2.1.3
Chen Test	*Mean*	Right-Skewed	3.2.1.4
Land's CI Method	*Mean*	Lognormal	3.2.1.5
Test for a Proportion and CI	*Proportion*		3.2.1.6
Sign Test	*Median*	None	3.2.2.1
Wilcoxon Signed Ranks Test	*Median /Mean*	Symmetric	3.2.2.2
t-Test and CI (equal variances)	*Diff in Means*	Normal	3.3.1.1.1
t-Test and CI (unequal variances)	*Diff in Means*	Normal	3.3.1.1.2
Test for Proportions and CI	*Diff in Props*		3.3.1.1.3
Paired *t*-Test	*Diff in Means*	Normal	3.3.1.2.1
Wilcoxon Rank Sum Test	*Diff in Means*	Same Variance	3.3.2.1.1
Quantile Test	Right-Tail		3.3.2.1.2
Slippage Test	Right-Tail		3.3.2.1.3
Sign Test	*Median*	None	3.3.2.2.1
Wilcoxon Signed Ranks Test	*Median /Mean*	Symmetric	3.3.2.2.2
Dunnett's Test	*Mean*		3.4.1.1
Kruskal-Wallis Test	*Mean*		3.4.1.2

Tests de normalité

Tests de normalité

Beaucoup de méthodes statistiques sont construites autour de la distribution normale → Données brutes ou transformées
→Tests de normalité

Histogramme , Diagramme Q-Q, Droite de Henry , etc.…

Test	Section	Sample Size	Recommended Use
Shapiro Wilk *W* Test	4.2.2	≤ 5000	Highly recommended.
Filliben's Statistic	4.2.2	≤ 100	Highly recommended, especially when used in conjunction with a normal probability plot.
Skewness and Kurtosis Tests	4.2.3	> 50	Useful for large sample sizes.
Studentized Range Test	4.2.4	≤ 1000	Highly recommended (with some conditions).
Geary's Test	4.2.4	> 50	Useful when tables for other tests are not available.
Chi-Square Test	4.2.5	Large[a]	Useful for grouped data and when the comparison distribution is known.
Lilliefors Kolmogorov- Smirnoff Test	4.2.5	> 50	Useful when tables for other tests are not available.

[a]

Exemple

Test de Shapiro-Wilk (Ammonium-1 NH4 (mg/l)) :

W	0.990
p-value (bilatérale)	0.936
alpha	0.05

Test de Shapiro-Wilk (Ammonium-2 NH4 (mg/l)) :

W	0.991
p-value (bilatérale)	0.962
alpha	0.05

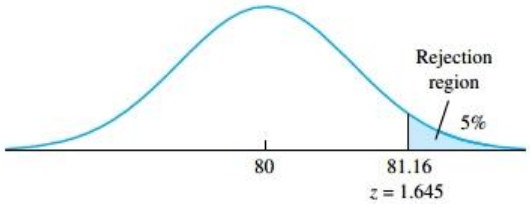
Erreur de 1ère et de 2nde espèce, Puissance d'un test

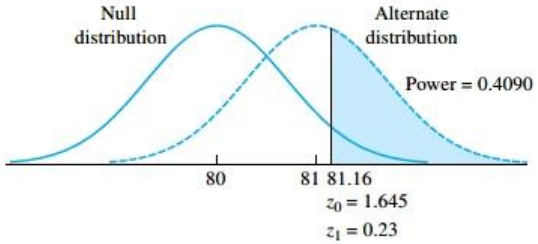
Erreur de première espèce : Probabilité de rejeter l'hypothèse nulle alors qu'elle n'est pas correcte = Niveau de signification α

Erreur de seconde espèce : Probabilité de ne pas rejeter l'hypothèse nulle alors qu'elle n'est pas correcte.

Puissance d'un test : 1 – P(Erreur de seconde espèce)

- Calcul de la région de rejet (z –score ou t-student sur base de l'hypothèse nulle)
- Calculer la probabilité que le test statistique soit dans la région de rejet si l'hypothèse alternative est vraie. (On déplace la valeur de l'hypothèse nulle et on calcule la probabilité si l'échantillon a une moyenne égale à la valeur de l'hypothèse alternative)



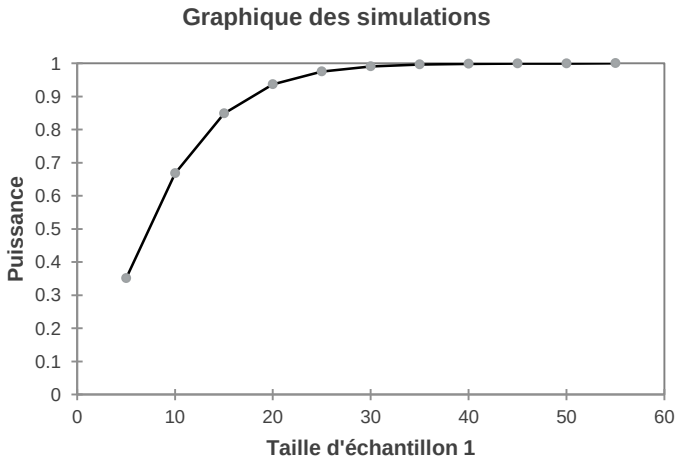


But, minimiser l'erreur de première et de seconde espèce

Quand la puissance du test n'est pas suffisante, elle peut être améliorée en augmentant la taille de l'échantillon

Exemple

Paramètres	Résultats
Taille d'échantillon 1	20
Taille d'échantillon 2	20
alpha	0.05
Moyenne (Groupe 1)	9
Moyenne (Groupe 2)	11
Ecart-type (Groupe 1)	1.5
Ecart-type (Groupe 2)	2
Bêta	**0.064**
Puissance	**0.936**



Traitement des non-detect et Valeurs extrêmes

Traitement des non-detect

Plusieurs possibilités de gérer les non détect :

Remplacer non-detect par $^{DL}/_{2}$

Moyenne ajustée (Trimmed mean, → on se donne un pourcentage et on écarte les valeurs extrêmes qui correspondent à ce pourcentage, valeurs des queues de distributions)

Méthode Aitchinson → Formule pour calculer la moyenne et la variance de la population !!! (hypothèse)

Méthode de Cohen → Démarche proche, formules différentes

La méthode de Cohen suppose que la population contient une distribution normale, celle de Aitchinson suppose que les proportions de la population en tenant compte des non-detect suit une distribution normale. Pour choisir entre les deux méthodes → diagramme Q-Q

D'autres méthodes existent :

Méthode de Kaplan Meier

(Méthode non paramétrique) → EPA Unified guidance (permet de déterminer moyenne et variance)

Méthode ROS (Robust Regression on Order Statistics)

(Permet d'imputer des valeurs aux non-detect)

MLE Maximum Likelihood Estimator (suppose de faire une hypothèse sur la distribution sous-jacente)

Exemple

Urea	D_Urea
0.1	0
6	1
0.1	0
5.6	1
14	1
71	1
78	1
0.1	0
0.1	0
58	1
69	1
0.1	0
390	1
46	1
99	1
9.4	1

	Num Obs	Num Miss	Num Valid	Detects	NDs	% NDs
Raw Statistics	70,00	1,000	69,00	53,00	16,00	23,19%
	Number	Minimum	Maximum	Mean	Median	SD
Statistics (Non-Detects Only)	16,00	0,100	230,0	14,47	0,100	57,48
Statistics (Non-Detects Only)	53,00	0,130	780,0	90,92	54,00	132,4
Statistics (All: NDs treated as DL value)	69,00	0,100	780,0	73,19	34,00	123,3
Statistics (All: NDs treated as DL/2 value)	69,00	0,0500	780,0	71,52	34,00	121,9
Statistics (Normal ROS Imputed Data)	69,00	-298,5	780,0	34,43	32,00	159,3
Statistics (Gamma ROS Imputed Data)	69,00	0,0100	780,0	69,97	32,00	122,0
Statistics (Lognormal ROS Imputed Data)	69,00	0,130	780,0	70,49	32,00	121,7

Valeurs extremes (outliers)

Test des valeurs extrêmes– Test de Dixon(n ≤ 25)

- **Test paramétrique**
  - Le test considère que à l'exception des valeurs extrêmes, la distribution suit une loi normale.
  - Faire un test de normalité avant de procéder au test
  - Si les données ne suivent pas une loi normale, procéder à une transformation

Test de Discordance (n ≥ 25)

- Même principe que le test de Dixon → Test de normalité à faire (1 valeur)

Test de Roesner (n ≥ 25)

- Même principe que les précédents → Test de normalité à faire (10 valeurs)

Test de Walsh (n ≥ 60)

- **Test non paramétrique**
- Permet d'identifier le nombre possible d'outliers – basé sur r (échantillons les plus petits et les plus grands) , taille échantillon, calcul de la distance à X(r) et X(n+1 – r)

Valeurs extrêmes (cont.)

Identifier les valeurs qui sont susceptibles d'être des outliers (Graphiquement, Boxplots, Diagramme Quantiles, Histogrammes, …)

Appliquer les tests statistiques

Se documenter sur la validité effective des conclusions des différents tests

Analyser les données avec et sans outliers

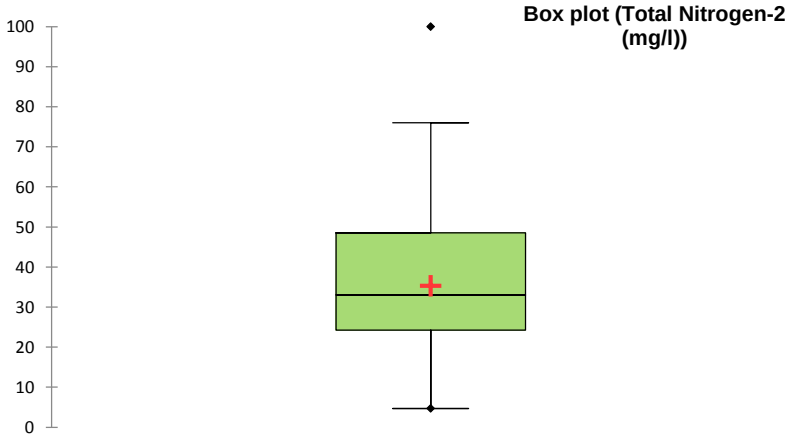
Documenter le processus

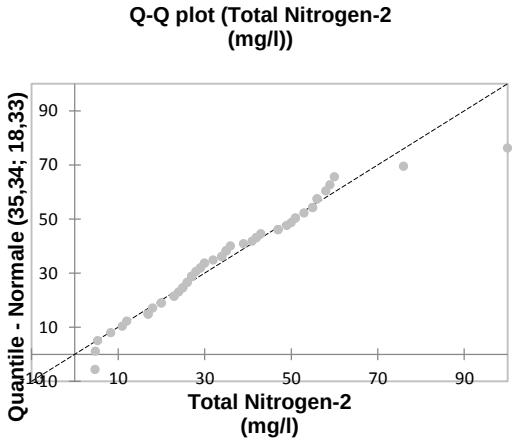
Tests d'identification d'outliers

Table 4-3. Recommendations for Selecting a Statistical Test for Outliers

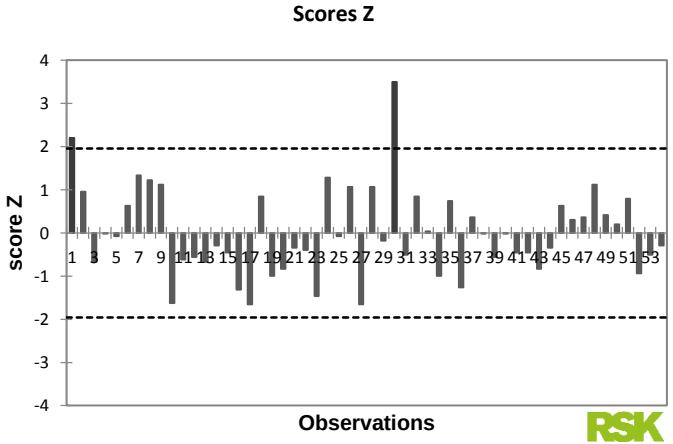
Sample Size	Test	Section	Assumes Normality	Multiple Outliers
$n \leq 25$	Extreme Value Test	4.4.3	Yes	No/Yes
$n \leq 50$	Discordance Test	4.4.4	Yes	No
$n \geq 25$	Rosner's Test	4.4.5	Yes	Yes
$n \geq 50$	Walsh's Test	4.4.6	No	Yes

Exemple





Total Nitrogen-2 (mg/l)	G	G(Valeur critique)	p-value	Pas
100.000	0.252	0.250	0.048	1



4. Logiciels statistiques

Logiciels statistiques

Freeware

- GWSDAT (Groundwater Spatio-Temporal Data Analysis Tool)
  - Source: API www.api.org
- MAROS (Monitoring and remediation optimization system software)
  - Souce: GSI www.gsi-net.com
- Pro-UCL
  - Source: US EPA www.epa.gov
- R for statistics
  - Source: R Project www.r-project.org
- VSP (visual sampling plan)
  - Source: Pacific Northwest National Laboratory http://vsp.pnnl.gov

Questions and answers

Thank you

SAFEGUARDING YOUR
BUSINESS ENVIRONMENT