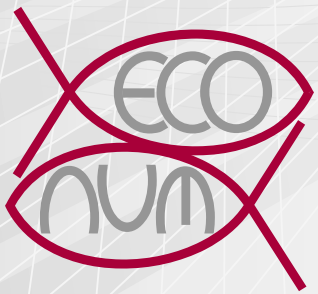
Biostatistique et probabilités

2ème Bachelier en Biologie

Prof. Philippe Grosjean <Philippe.Grosjean@umons.ac.be>
Assist. Guyliann Engels <Guyliann.Engels@umons.ac.be>

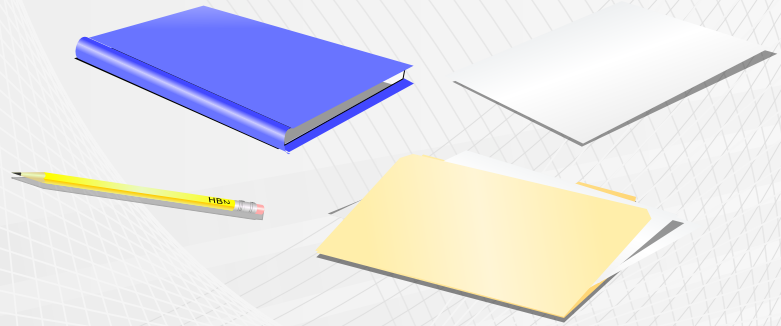
Université de Mons
Laboratoire d'Écologie numérique des Milieux aquatiques

Cours 7



UMONS

Objectifs de ce cours

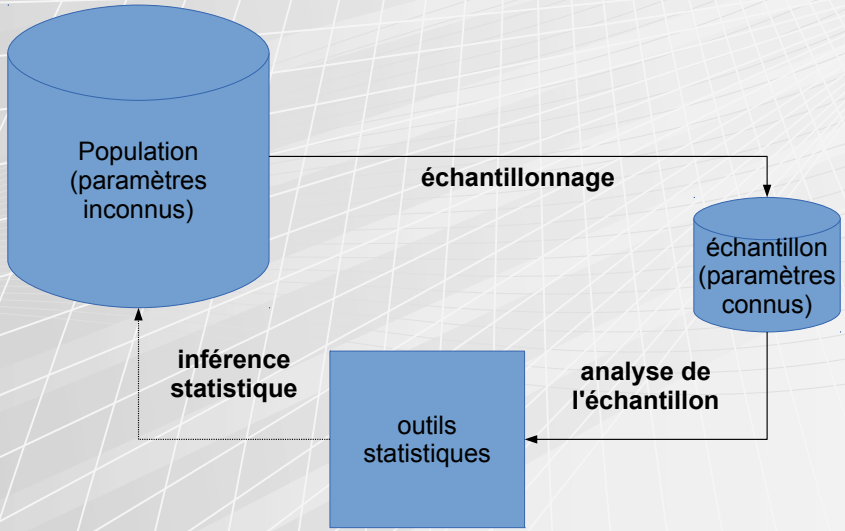


- Découvrir la distribution d'un échantillon à l'aide de métaexpérience
- Maitriser la distribution de Student et son utilisation
- Calculer un intervalle de confiance autour d'une moyenne
- Découvrir le test d'hypothèse et en particulier, le test de Student

UMONS

Échantillonnage et inférence

- L'inférence (inference) est l'estimation des caractéristiques de la population sur base de l'analyse statistique d'un échantillon représentatif.
- Un échantillon peut être qualifié de représentatif s'il est issu d'un processus d'échantillonnage **correct** (ex. : échantillonnage aléatoire).
- L'échantillonnage aléatoire fait intervenir le **hasard** dans la sélection des individus, tel que chaque individu de la population a autant de chance que les autres d'être pris (*question : comment faire en pratique ?*).



Expérience, observation et causalité

- Deux façons d'étudier un phénomène biologique : l'expérience (experimentation) ou l'observation (monitoring or survey).

Les concepts et méthodes de la biostatistique visent à mettre en évidence un signal en présence de bruit, et ce, aussi bien au niveau de la planification d'une **expérience** ou d'une **observation** qu'à l'interprétation des données obtenues.

- Corrélation (correlation) : variation proportionnelle ou inversement proportionnelle de deux variables observées.
- Relation (relationship) : idem, mais association des deux variables (un mécanisme sous-jacent est rendu implicitement responsable de cette relation).
- Causalité (causality) : la variation de la variable X est la cause de la variation de la variable Y. Hypothèse la plus forte, mais difficile à montrer par l'observation ; l'expérimentation est nécessaire !

UMONS

Distribution d'échantillon

- Quelle est la distribution de la moyenne d'un échantillon ? Processus : métaexpérience / variation d'échantillonnage => de combien $\bar{y}$ peut-il différer de μ ?
- Moyenne de la distribution d'échantillonnage : $\mu_{\bar{Y}} = \mu$
- Ecart type de la distribution d'échantillonnage : $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$ = erreur standard de la population initiale.
- Théorème central limite : « quelle que soit la distribution de Y, la distribution d'échantillons (distribution des moyennes des échantillons) issus de Y est approximativement normale si n est suffisamment grand ». Illustration pratique : http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Donc, si la distribution de population est inconnue, mais que n est grand, la distribution d'échantillon tendra toujours vers une distribution Normale.

Distribution t de Student

- Distribution typique de la moyenne d'échantillons lorsque n est petit et la distribution de la population d'où sont issus ces échantillons est Normale. Elle a 3 paramètres : μ, σ et ddl (degrés de liberté) = $n - 1$.
- Tend vers une distribution normale lorsque n tend vers l'infini (cf, théorème central limite).
- Comme pour la distribution Normale réduite, on utilise très souvent la distribution t réduite, de paramètres μ=0, σ=1 et $ddl = n$ -1. Notation : pour indiquer le quantile d'une distribution t réduite de $n - 1$ ddl et correspondant à une aire à droite (probabilité) de p, on note : t_p^{n-1}
- Visualiser et calculer des quantiles/probabilités dans R Studio pour la distribution de Student t.

UMONS

Intervalle de confiance

Rappel

Notation moyenne et écart type de la population ou de l'échantillon ; introduction de l'estimateur de μ ou σ (dit mu chapeau ou sigma chapeau)

- Concept d'intervalle de confiance : principe de l'homme invisible qui promène son chien (voir aussi : http://onlinestatbook.com/stat_sim/conf_interval/index.html).
- Si la distribution de la moyenne d'échantillonnage est connue, alors, on peut indiquer l'intervalle dans lequel la moyenne de la population se trouve avec une probabilité donnée (notée $1 - \alpha$) :

$$IC(1-\alpha) = \mu_{\bar{Y}} \pm t^{n-1}_{\alpha/2} \cdot \sigma_{\bar{Y}}$$

- Cet intervalle, c'est l'intervalle de confiance au seuil α. En remplaçant par les valeurs connues, on a :

$$IC(1-\alpha) = \hat{\mu} \pm t^{n-1}_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}} = \bar{Y} \pm t^{n-1}_{\alpha/2} \cdot \frac{s_Y}{\sqrt{n}}$$

Conditions de validité

- Échantillonnage aléatoire,
- Observations indépendantes les unes des autres,
- Distribution de la population :
  - Normale => intervalle de confiance utilisant la distribution t de Student est exacte,
  - Approximativement normale => cet intervalle de confiance est approximativement exact,
  - Non normale => cet intervalle est approximativement exact si n est grand,

En pratique

Test à l'aide d'un graphique quantile-quantile de comparaison de l'échantillon à une distribution t de Student ou Normale (à condition, bien sûr, que l'échantillon soit suffisamment grand).

UMONS

Utilisation

- De même que l'écart type, la variance ou l'erreur standard, l'intervalle de confiance peut servir à représenter la dispersion des moyennes d'un échantillon, soit dans un tableau numérique, soit sur un graphique (barres d'erreurs).
- L'écart type sert à représenter la dispersion des données ; l'erreur standard ou l'intervalle de confiance sont plus adaptés pour représenter la dispersion de la moyenne de l'échantillon dans une optique de comparaison de moyennes. De plus, l'I.C. permet de comparer 2 ou plusieurs moyennes (superposition ou non des I.C., ou de l'I.C. par rapport à une valeur cible).

Attention

Graphe avec barres d'erreurs : il faut toujours préciser ce que sont les barres d'erreur dans la légende du graphe ou de l'axe correspondant !

UMONS

Test d'hypothèse utilisant l'IC

- Représentation d'une question posée par le biologiste sous la forme d'une hypothèse et de son contraire concernant la moyenne.

Exemple

Est-ce qu'il y a eu croissance massique d'un organisme entre deux mesures => utilisation de la différence de masse entre *t0* et *t1* = M_{t1}- M_{t0}= ΔM.

- Observation ou réalisation d'une expérience pour quantifier le problème à travers la mesure d'une variable quantitative pertinente.
- Expression de l'hypothèse et de son contraire sous forme d'une relation mathématique, notation :
  - Hypothèse de base (pas de croissance) H_0: $\Delta M = 0$
  - Son contraire (il y a eu croissance) H_1 ou H_A: $\Delta M > 0$

UMONS

Test d'hypothèse utilisant l'IC - Exemple

Question

Dans une étude visant à soigner l'anorexie, est-ce que les patientes qui n'ont pas subi de traitement (contrôle) ont gardé un poids constant ?

- Observation : 72 patientes anorexiques dont 26 ont été placées au hasard dans le groupe contrôle. => mesure de leur masse corporelle au début (*Prewt*) et à la fin (*Postwt*) en livres.
- Expression de l'hypothèse de son contraire sous forme d'une relation mathématique :
  - Calcul de *DeltaM* pour l'échantillon = *PostWt* - *Prewt*
  - Hypothèse de base (pas de changement de masse) : H_0: $\Delta M = 0$
  - Son contraire (changement de masse) : H_1: $\Delta M \neq 0$

Problème

Comment discerner les vraies différences des erreurs de mesure, de la présence de quelques cas extrêmes introduits au hasard de l'échantillonnage, etc. ?

Test d'hypothèse utilisant l'IC - Exemple (2)

Cela revient à inférer la valeur de ΔM pour l'ensemble des jeunes filles anorexiques, dont on a extrait un échantillon aléatoirement.

Qu'est-ce qu'on possède comme outil pour cela ?

- La moyenne de l'échantillon pour la variable *DeltaM*, considérée comme représentative de ΔM (moyenne pour la population),
- On connaît la distribution statistique de la moyenne de *DeltaM* :

$$\bar{DeltaM} \sim t(\Delta M, \sigma/\sqrt{n}, n-1)$$

- On peut donc calculer un intervalle de confiance autour de $\bar{DeltaM}$.

Raisonnement

Si $\Delta M = 0$ (valeur sous H_0) appartient à l'IC, on ne pourra pas rejeter H_0 (puisqu'elle fait partie des valeurs possibles). Dans le cas contraire, on pourra la rejeter. La probabilité α associée à l'IC indique la possibilité que ce dernier n'inclue pas la vraie moyenne alors que H_0 est vraie et quantifie donc le risque de se tromper dans ce cas.

Nous avons maintenant construit ce qu'on appelle un test d'hypothèse ; ici, c'est le test t de Student univarié.

Test d'hypothèse

Donc, pour élaborer un test d'hypothèse (cas général), il faut :

- Construire un descripteur statistique pertinent dont on peut calculer la probabilité d'occurrence sous H_0 (par exemple, parce qu'on connaît sa distribution statistique),
- Si la probabilité d'occurrence sous H_0 de la valeur de la statistique est trop faible, on en conclura que H_0 a une probabilité trop faible d'être vraie et on la rejettera au bénéfice de H_1.
- La probabilité à la limite de l'acceptation ou du rejet est la valeur p associée au test. Si $p < \alpha$, on rejette H_0, sinon, on ne rejette pas H_0.

IC ou valeur p ?

Dans le cas du test de Student, le calcul peut se faire soit par rapport à l'IC, soit en utilisant la valeur p. C'est un cas particulier. En général, les tests d'hypothèses ne renvoient que la valeur p.

Test d'hypothèse - Important!

Attention

Afin de construire un jeu de données propice pour faire de l'inférence, il faut toujours réaliser un échantillonnage aléatoire (ou autre type d'échantillonnage valide, stratifié par exemple) et avoir des observations indépendantes entre elles.

- Echantillonnage aléatoire et observations indépendantes sont donc deux contraintes de départ *indispensables* avant de pouvoir réaliser un test d'hypothèse sur les données.
- Autre élément important : la réplication. Plus n est grand, plus on a d'information à disposition, et plus le test est efficace.
  Montrez à l'aide d'un exemple fictif comment l'IC diminue lorsque n augmente.

UMONS

Interprétation graphique

- Représenter la distribution de la statistique étudiée *telle qu'elle se présente si H_0 est vraie*. Par exemple, H_0: $\Delta M = 0$ => la distribution de la moyenne d'échantillon attendue sous H_0 est :

$$\bar{Delta}M \sim t(0, \sigma/\sqrt{n}, n-1)$$

- Y reporter le seuil α considéré, et préciser la zone de rejet de H_0 (quantiles égaux ou plus extrêmes que celui délimitant la zone de rejet par rapport à α) et celle de non rejet de H_0 (aire complémentaire).
- Y reporter ensuite le quantile observé de la statistique pour notre échantillon, $\bar{Delta}M$.
- Est-il situé dans la zone de rejet ou non de H_0 ?

=> *tirer les conclusions correspondantes à propos de la question biologique posée au départ.*

UMONS

Test unilatéral et bilatéral

Beaucoup de tests d'hypothèses, tel le test de Student, admettent des variantes unilatérales et bilatérales selon la manière dont H_1 est posée :

- Test bilatéral (correspondant à la résolution à l'aide de l'IC) : l'aire est divisée en deux parties égales réparties à gauche et à droite de la distribution.
  H_1: $Y \neq cste$,
- Test unilatéral à gauche : l'aire de rejet est répartie intégralement à gauche.
  H_1: $Y < cste$,
- Test unilatéral à droite : l'aire de rejet est répartie intégralement à droite.
  H_1: $Y > cste$.

Importance de la représentation graphique pour s'y retrouver !

L'aire de rejet correspondant à α est positionnée différemment.

UMONS

Résolution de l'exemple

- Vérification des conditions d'application :
  - Échantillon aléatoire et observations indépendantes ?
  - Distribution normale ou quasi-normale des données et/ou n grand ?
- Choix du seuil α du test avant de le réaliser ! Valeurs souvent utilisées : $\alpha = 5\%$, 1%, 0.1%, ...
- Formulation des hypothèses H_0 et H_1 (et donc, choix entre un test uni- ou bilatéral).
- Calcul de la statistique t et de la valeur p associée au test.
- Réalisation du graphique de la zone de rejet et non rejet du test (facultatif, mais utile pour visualiser ce qui est fait).
- Rejet ou non de H_0 ?
- Interprétation biologique : la *masse moyenne* des sujets anorexiques témoins est-elle restée stable entre le début et la fin de l'expérience ? Il s'agit bien sûr d'une condition nécessaire pour que les autres comparaisons aient un sens.

UMONS

Présentation des résultats

- Expression du résultat : attention à la formulation adéquate !
  - « Telle différence ou telle valeur est/n'est pas significative (significant) au seuil α de 5% (test t univarié bilatéral, n=..., t=..., p=...) ».
  - « Nous pouvons/ne pouvons pas considérer que la valeur observée soit significativement différente de ... au seuil α de 5% (test t univarié unilatéral à gauche, n=..., t=..., p=...) ».
  - « Nous observons/n'observons pas de différence significative entre ... au seuil α de 1% (test t univarié unilatéral à droite, n=..., t=..., p=...) ».
- Test complexe, ou beaucoup de résultats : adjoindre un tableau résumant l'analyse statistique, sa probabilité et éventuellement, l'intervalle de confiance correspondant.
  - Utiliser une représentation à l'aide d'astérisques pour indiquer la *significativité*. Ex. : * = moyennement significatif (5%), **= significatif (1%), *** = très significatif (0.1%).
  - Représentation graphique de la *significativité* en utilisant le même principe (astérisques reportées sur les graphes en barres verticales ou sur les boites de dispersion parallèles, par exemple).

UMONS

Exemple de présentation complexe

TABLE II

Statistics on the four batches of Fc in the beginning of the experiment (initial) and after 4 months (final). ** = does not fit a normal curve (P < 0.01).

batch	Fc1	Fc2	Fc3	Fc4
treatment	size sorted batches of mean-sized individuals only, no "head" already differentiated.			
initial mean size (mm)	6.0	6.0	6.0	6.0
final mean size (mm)	14.3	15.1	14.0	14.9
increase in size (mm)	8.3	9.1	8.0	8.9
Kolmogorov-Smirnov / Lilliefors on final sizes (mm)				
maximum difference	0.161	0.148	0.106	0.124
probability	0.003**	0.008**	0.175	0.061

UMONS

Seuil de probabilité critique

- Le seuil de probabilité critique (ou valeur p associée au test, p-value en anglais) définit la probabilité qui se situe à la limite entre acceptation et rejet de H_0 (explication sur un graphique).
- La décision du rejet ou de l'acceptation d'un test d'hypothèse sur base du seuil de probabilité critique est simple. Elle suit la règle suivante :
  - Si $p \geqslant \alpha$, on ne rejette pas H_0(écriture: $\not{R}H_0$)
  - Si $p < \alpha$, on rejette H_0 (écriture : RH_0)

Notez ceci

Évitez de dire "acceptation de H_0". En effet, comme l'interprétation basée sur les IC le montre clairement, ne pas rejeter H_0 n'est pas équivalent à l'accepter. Il se peut toujours que H_0 soit quand même fausse, mais que les données soient insuffisantes (par exemple, n trop faible) pour le montrer. Évitez également de dire "acceptation de H_1" pour une raison similaire.

Risque α et β, puissance d'un test

- Quatre cas de figure possibles dans l'interprétation d'un test :
  - Ne pas rejeter H_0 lorsque H_0 est vraie (correct !),
  - Rejeter H_0 lorsque H_0 est fausse (correct !),
  - Rejeter H_0 lorsque H_0 est vraie (erreur de $1^{\text{ère}}$ espèce),
  - Ne pas rejeter H_0 lorsque H_0 est fausse (erreur de $2^{\text{ème}}$ espèce),.
- L'erreur de première espèce est associée au risque α, équivalent au seuil α (cf. applet simulation sur les IC, étudiée plus haut).
- L'erreur de seconde espèce est associée au risque β
- Plus le risque β est petit, plus le test est puissant. β dépend à la fois du test et de n (β diminue quand n augmente) => puissance d'un test = $1 - \beta$. Voir : http://onlinestatbook.com/stat_sim/robustness/index.html

UMONS

Test t de Student pour échantillons appariés

- Reprenons notre exemple de la stabilité massique de sujets anorexiques dans le lot contrôle.
- Le même problème peut être spécifié sous la forme d'une comparaison de deux populations de mesures : *Prewt* et *Postwt* dont les observations sont appariées (i.e., mesurées sur les mêmes individus).
- Calcul sous forme d'un test t de Student pour deux échantillons appariés (test t apparié).
- Comparaison avec le calcul sur ΔM. Il s'agit bien du même test, mais formulé légèrement différemment.
- Mêmes conditions d'application que le test t de Student univarié. La distribution des données se teste également par un graphe quantile-quantile comparant la différence des observations par rapport à une distribution t de Student.

UMONS

Test *t* de Student non apparié

Peut-on comparer les moyennes d'une même variable mesurées sur des sous-populations, et donc, des individus, différentes ?

Exemple

Peut-on comparer *DeltaW* entre le contrôle *Cont* et le groupe traité en famille *FT* pour le jeu de données *anorexia* ?

- Moyennes des groupes A et B sont-elles égales (H_0: $\mu_A = \mu_B$) ? Ce qui revient parfois à considérer qu'elles sont issues de la même population.
- H_1: moyennes différentes (test bilatéral) ou moyenne de A supérieure (ou inférieure) à moyenne de B (test unilatéral).
- Test paramétrique : test *t* de Student, distribution de même nom pour les moyennes de A et de B.
- Pas d'appariement des données.
- n_A n'est pas nécessairement égal à n_B.

Dans un tel cas, quelle statistique utiliser ? Quelle est sa distribution ? Comment construire un test de comparaison des deux moyennes ?

Comparaison de deux moyennes

- Étude de la distribution de $\mu_A - \mu_B$, sur base des données de l'échantillon $\bar{Y}_A - \bar{Y}_B$
- Calcul de l'intervalle de confiance :

$$\text{IC}_{95\%} = (\bar{Y}_A - \bar{Y}_B) \pm t_{0.025}^{\min(n_A-1, n_B-1)} \cdot \mathbf{SE}_{(\bar{Y}_A - \bar{Y}_B)}$$

avec

$$\mathbf{SE}_{(\bar{Y}_A - \bar{Y}_B)} = \sqrt{\text{SE}_A^2 + \text{SE}_B^2} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

Dans le logiciel

Une variable facteur à deux niveaux, et une variable quantitative. Il existe deux variantes : variances égales ou différentes, avec calcul légèrement différent (optimisé) dans les deux cas (exemple sur `anorexia`).

Test *t* de Student non apparié - hypothèses de base

- Échantillon représentatif (ex. : échantillonnage aléatoire).
- Observations indépendantes.
- La distribution des écarts aux moyennes respectives (résidus ou residuals en anglais) doit suivre une distribution *t* de Student. Utilisation de la fonction **ave()** pour calculer ces écarts.

Note

Le test *t* de Student pour échantillons non appariés se dit aussi *test de Student pour échantillons indépendants*.

UMONS