

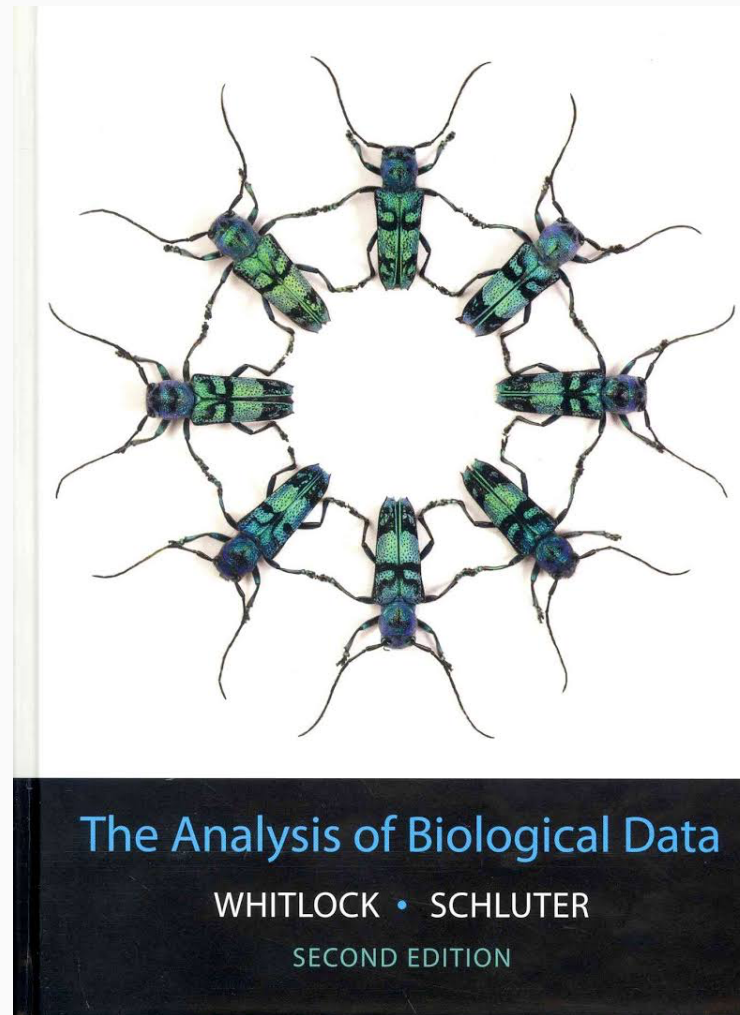
Notions et Concepts Clés en Statistiques

Benoît Simon-Bouhet
La Rochelle Université

Préambule...

Ouvrage de référence

Très bon ouvrage pour démarrer les statistiques en écologie :



Nombreux exercices corrigés et beaucoup de méthodes abordées

Inférence et estimation

Population vs. échantillon

Outils pour quantifier l'incertitude

Définition

L'**estimation** est le processus permettant de *déduire* (inférer) une quantité inconnue d'une **population** grâce aux données d'un **échantillon**

Définition

Un **paramètre** est une quantité permettant de décrire une **population**.

Un **estimateur** (ou une statistique) est une quantité équivalente calculée grâce à un échantillon.

Tests d'hypothèses

Outre **l'estimation**, la deuxième utilisation la plus importante des statistiques est le **test d'hypothèses**.

- **Hypothèse** statistique = **affirmation** spécifique au sujet d'un **paramètre de la population**.
- Exemples

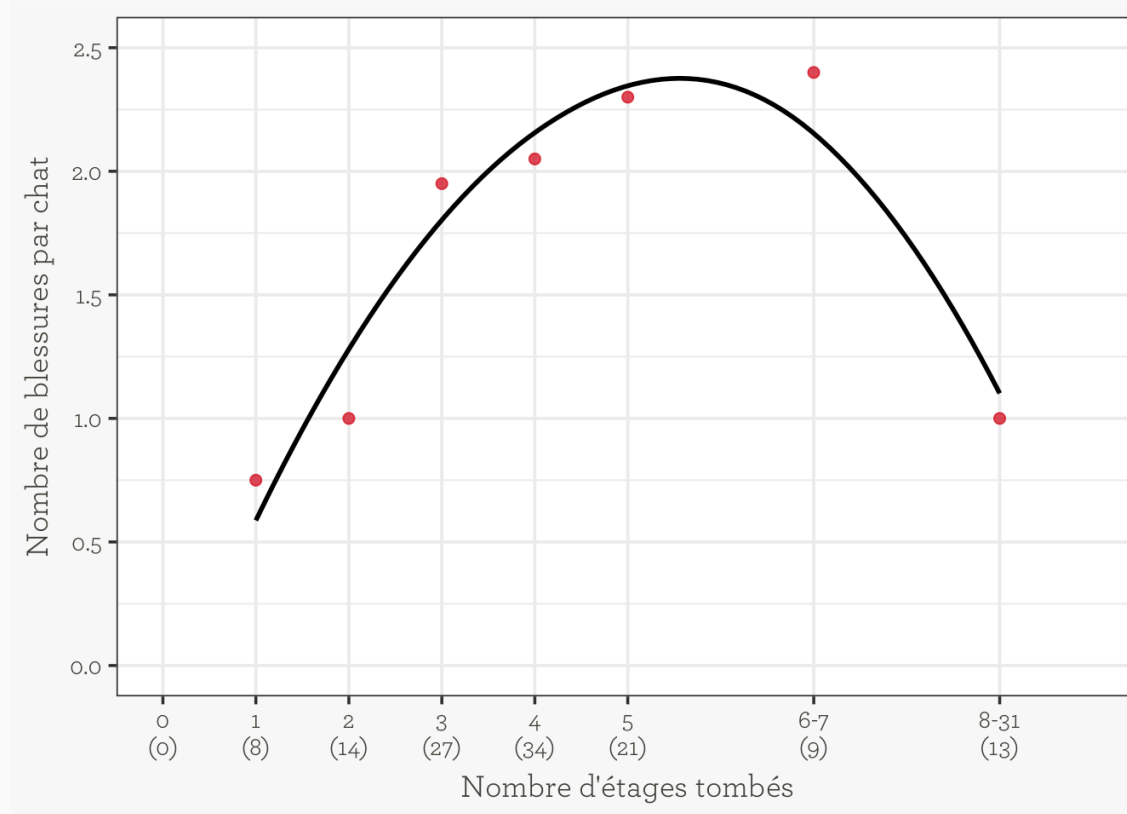
Dans cette section, nous verrons

- comment les **échantillons** doivent être **recueillis**
- quelles **conclusions** peuvent en être tirées
- les différents **types de variables** que nous pouvons mesurer à partir des **échantillons**
- divers termes et concepts que nous utiliserons tout au long de la formation

1. Échantillonner des populations naturelles

Exemple : chutes de 🐱...

Taux de blessures des chats qui tombent des bâtiments de NYC (Whitney & Mehlhaff 1987)



➤ Feline High Rise Syndrome (FHRS)

Exemple : chutes de ...

Comment expliquer cet effet ?

→ Les chats ont besoin de **moins d'un étage** pour retomber sur leurs pattes

Comment les données ont-elles été collectées ?

- > **Aucun** chat n'est tombé du **RDC**
- > Le nombre de chat tombé **augmente** à chaque étage entre le 1^{er} et le 4^e
- > **Tous les chats** qui tombent ne sont pas emmenés chez le vétérinaire : seul **un échantillon** est étudié

L'échantillon est **biaisé** !

 **Danger**

Échantillon **biaisé** ► données **déformées** ► conclusions **erronées**

Propriétés d'un bon échantillon

Définition

L'**erreur d'échantillonnage** (sampling error) est l'écart entre un **estimateur** et la vraie valeur du **paramètre de la population**, causée par le **hasard de l'échantillonnage**.

La **dispersion** des estimateurs en raison de l'erreur d'échantillonnage renseigne sur la **précision** d'un estimateur

Plus l'erreur d'échantillonnage est faible ▼, plus la précision est grande ▲.

Outre la précision, un **estimateur** doit aussi être **juste**, ou **non biaisé**

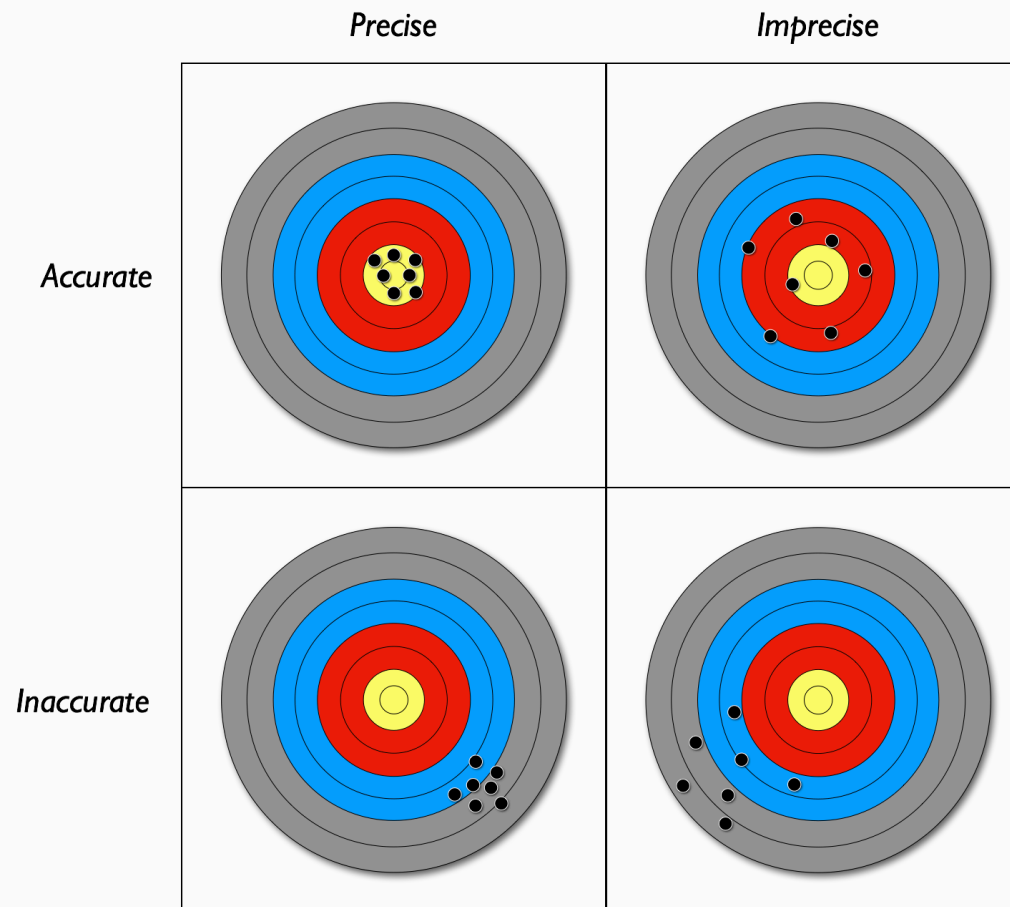
Définition

Le **biais** est un **écart systématique** entre l'**estimation** que nous obtiendrions si nous pouvions échantillonner une population encore et encore, et **la véritable caractéristique de la population**.

Propriétés d'un bon échantillon

Précision vs. justesse

L'objectif principal de toute méthode d'échantillonnage est de minimiser à la fois l'**erreur d'échantillonnage** et les **biais** d'estimation.



Échantillonnage aléatoire

C'est une **condition nécessaire** pour **toutes les méthodes** statistiques que nous aborderons.

Un **échantillon** aléatoire issu d'une **population** doit vérifier **deux critères** :

- Toutes les unités de la **population** ont la **même chance** d'intégrer l'**échantillon**
- La sélection des unités est **indépendante**

! Important

Dans un **échantillon aléatoire**, chaque membre de la **population** a une **probabilité égale et indépendante** d'être sélectionné

! Important

L'**échantillonnage aléatoire** minimise les **biais** et rend possible la quantification de l'**erreur d'échantillonnage**

Comment prélever un échantillon aléatoire ?

- Créer une **liste** de chaque unité de la **population d'intérêt**, et attribuer à chacun un **numéro** compris entre 1 et la taille de la population
- Décider du nombre d'unités à **échantillonner** (en général, nommé n)
- À l'aide d'un **générateur de nombre aléatoire**, tirer au hasard n nombres entre 1 et le nombre total d'unités de la **population**
- **Collecter** les unités dont les numéros ont été tirés au sort

C'est évidemment impossible pour les **populations naturelles**.

Une solution : **grouper** (parcelles, champs, lots, transects, cadrats...)

L'échantillon de convenance

À éviter à tout prix !

Définition

Un **échantillon de convenance** est une collection d'unités d'échantillonnage **facilement accessible** à l'expérimentateur

Ça n'est pas un **échantillon aléatoire** et son principal problème est le **biais** :

- Si seuls les chats qui arrivent jusqu'au cabinet vétérinaire sont examinés, le taux de blessures est probablement **sous-estimé**
- L'**effondrement** de la pêcherie de morue de l'Atlantique Nord au siècle dernier a pour cause principale une **surestimation** des stocks disponible (Walters & Maguire 1996)
- L'ajout de **renforts inefficaces** aux niveaux des impacts de balles des avions de la British Royal Air Force pendant la seconde guerre mondiale

L'échantillon de convenance

À éviter à tout prix !

✍ Définition

Un **échantillon de convenance** est une collection d'unités d'échantillonnage **facilement accessible** à l'expérimentateur



Un **échantillon de convenance** peut aussi ne pas respecter la condition d'**indépendance**.

2. Tests d'hypothèses : les grand **principes**

Principe des tests statistiques

Les hypothèses

Notion d'inférence statistique :

- On étudie une population à partir d'un échantillon
- On tente de déterminer la valeur d'un paramètre à partir d'un estimateur

1. On formule une hypothèse nulle concernant la valeur d'un paramètre d'intérêt (e.g. la moyenne μ de la population vaut 32.4)
2. On cherche à prendre une décision concernant cette hypothèse : sommes nous en mesure de la rejeter ou non, compte tenu des données dont nous disposons et de l'incertitude inhérente au processus d'inférence statistique ?

Principe des tests statistiques

Les hypothèses

Définition : hypothèse nulle

On la note H_0 . Elle porte **toujours** sur la valeur d'un **paramètre de la population** d'intérêt. L'hypothèse nulle est l'hypothèse "**inintéressante**", celle qu'on aimerait bien rejeter car cela indiquerait que quelque chose d'intéressant, original, inattendu se passe dans la population étudiée.

Important

- Outre H_0 , on formule aussi une **hypothèse alternative** opposée, notée H_A ou H_1 .
- La **décision** du test est toujours prise **par rapport à H_0** .
- Les hypothèses sont des **affirmations**, pas des questions !

Principe des tests statistiques

La statistique du test

Après les hypothèses, la **statistique du test** est le second ingrédient indispensable aux tests statistiques.

Statistique d'un test

C'est un **nombre calculé** à partir des **données disponibles** et dont on connaît la **distribution** théorique sous H_0 .

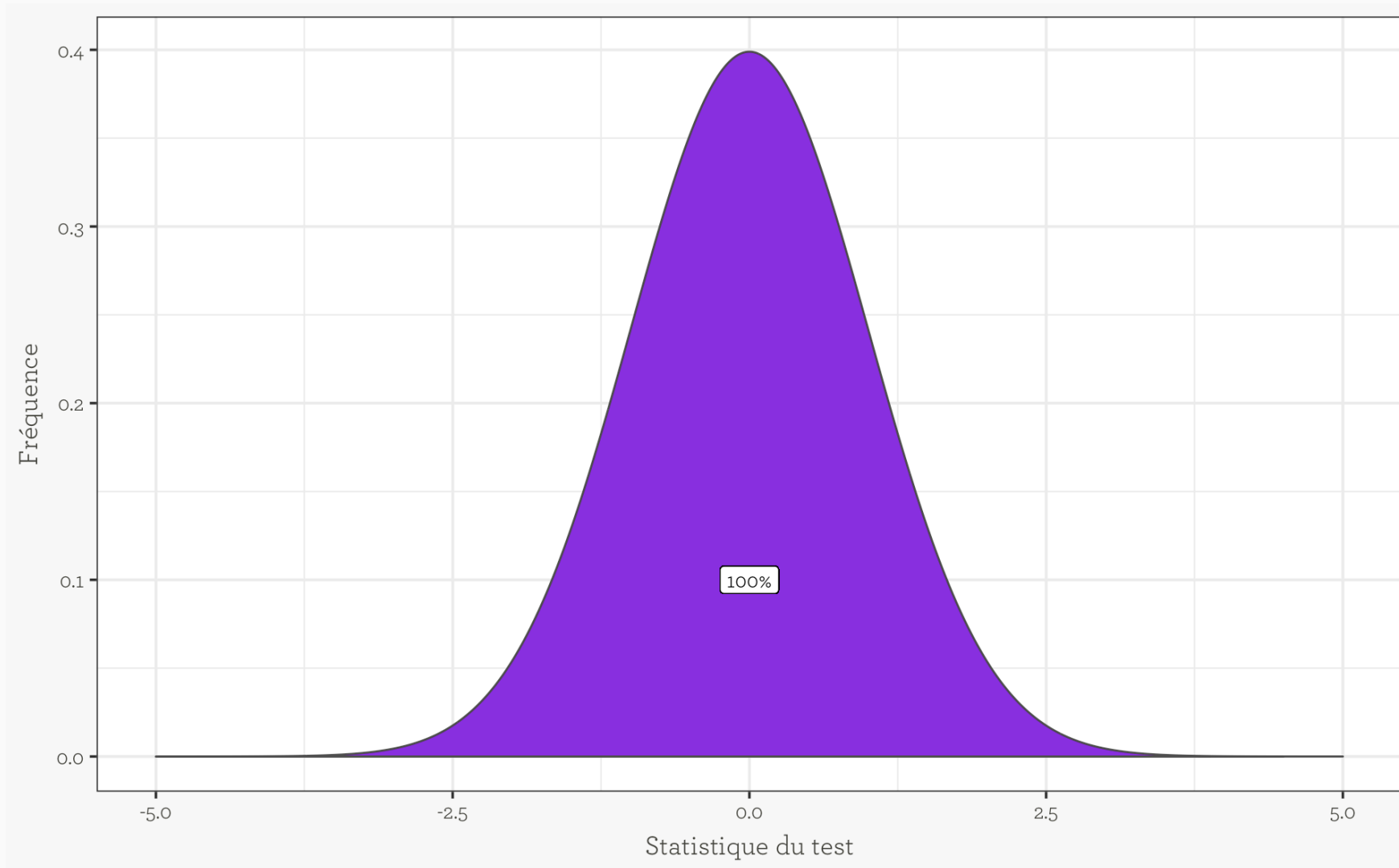
C'est ce nombre qui, indirectement, nous permettra de prendre une décision concernant le test statistique (et en particulier, le rejet ou non de H_0).

Tout test statistique possède une statistique du test. En voici quelques exemples :

- Test de normalité de Shapiro-Whilk : W ou W_{calc}
- Test d'homoscédasticité de Levene : F ou F_{calc}
- Test d'homoscédasticité de Bartlett : χ^2 ou χ^2_{calc}
- Test de comparaison de moyennes de Student : t ou t_{calc}

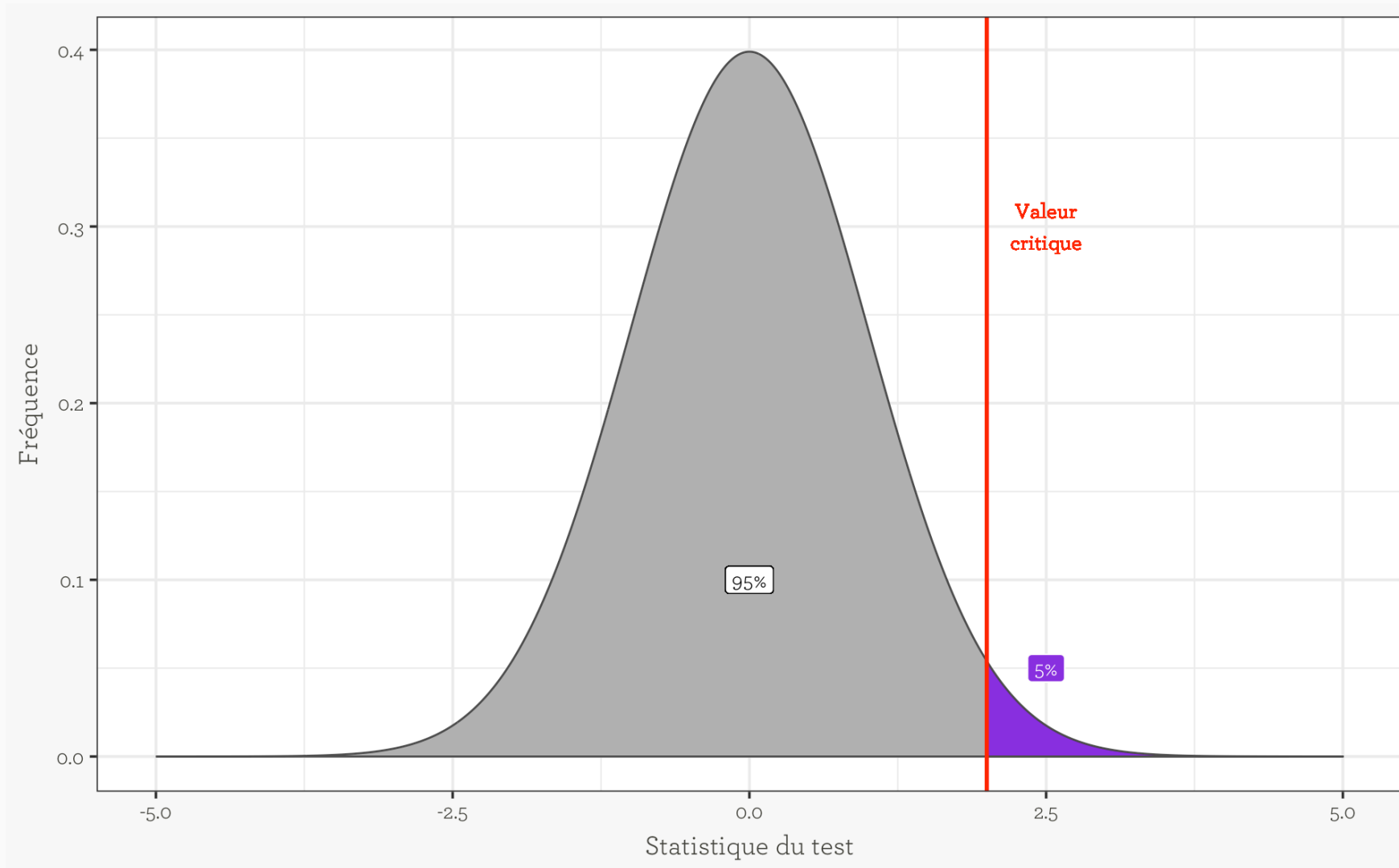
Principe des tests statistiques

La valeur critique de la statistique du test



Principe des tests statistiques

La valeur critique de la statistique du test



Principe des tests statistiques

La décision

On compare la statistique du test à la valeur critique :

- Si la statistique du test est **supérieure ou égale** à la valeur critique, on **rejette H_0** (et on valide donc H_A).
- Si la statistique du test est **inférieure** à la valeur critique, on **ne peut pas rejeter H_0** .

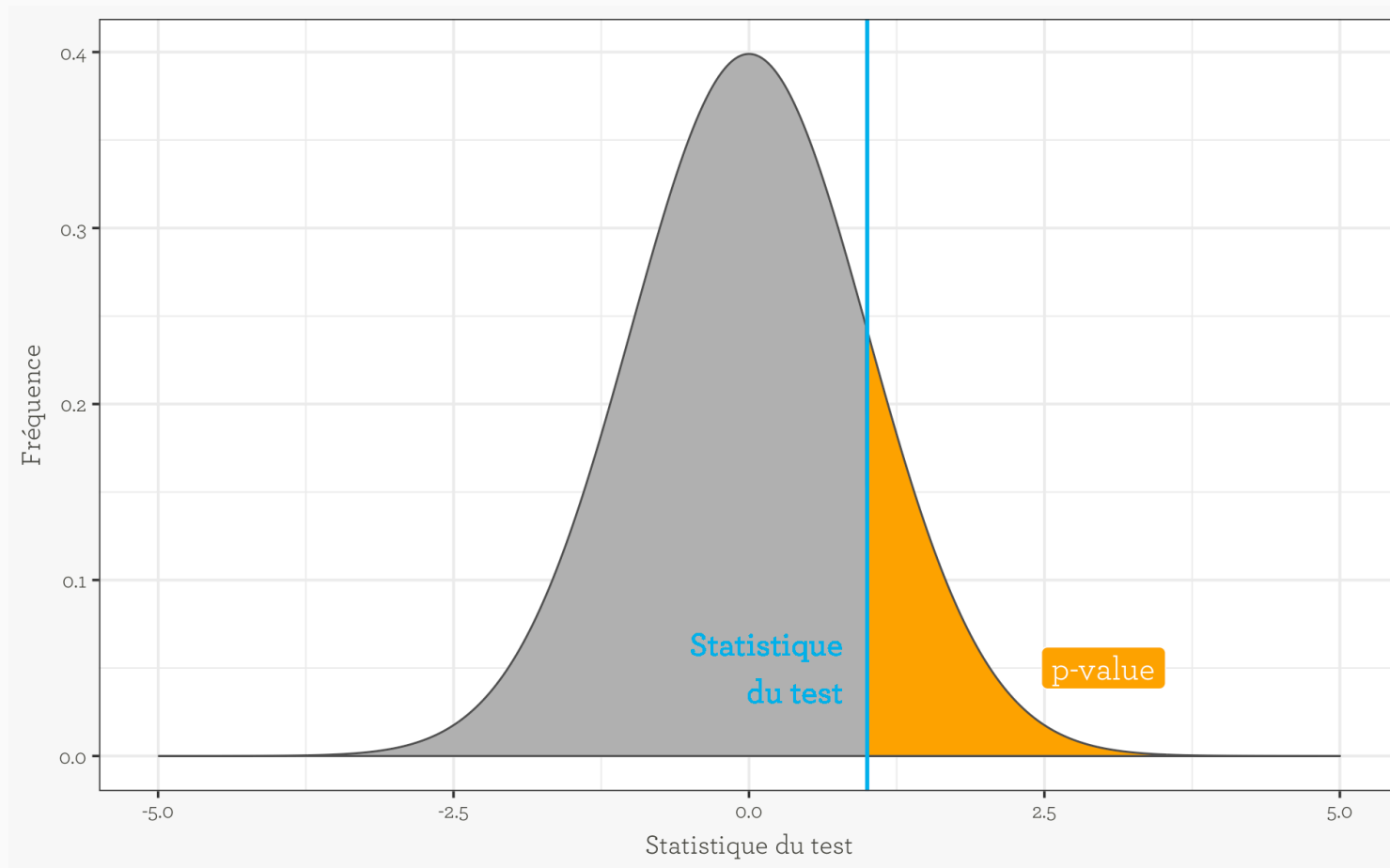
! Important

- La décision du test se prend **toujours** par rapport à H_0 .
- On ne dit jamais que H_0 est vraie. Au mieux, on dit qu'**on ne peut pas rejeter H_0** .

Principe des tests statistiques

La p -value

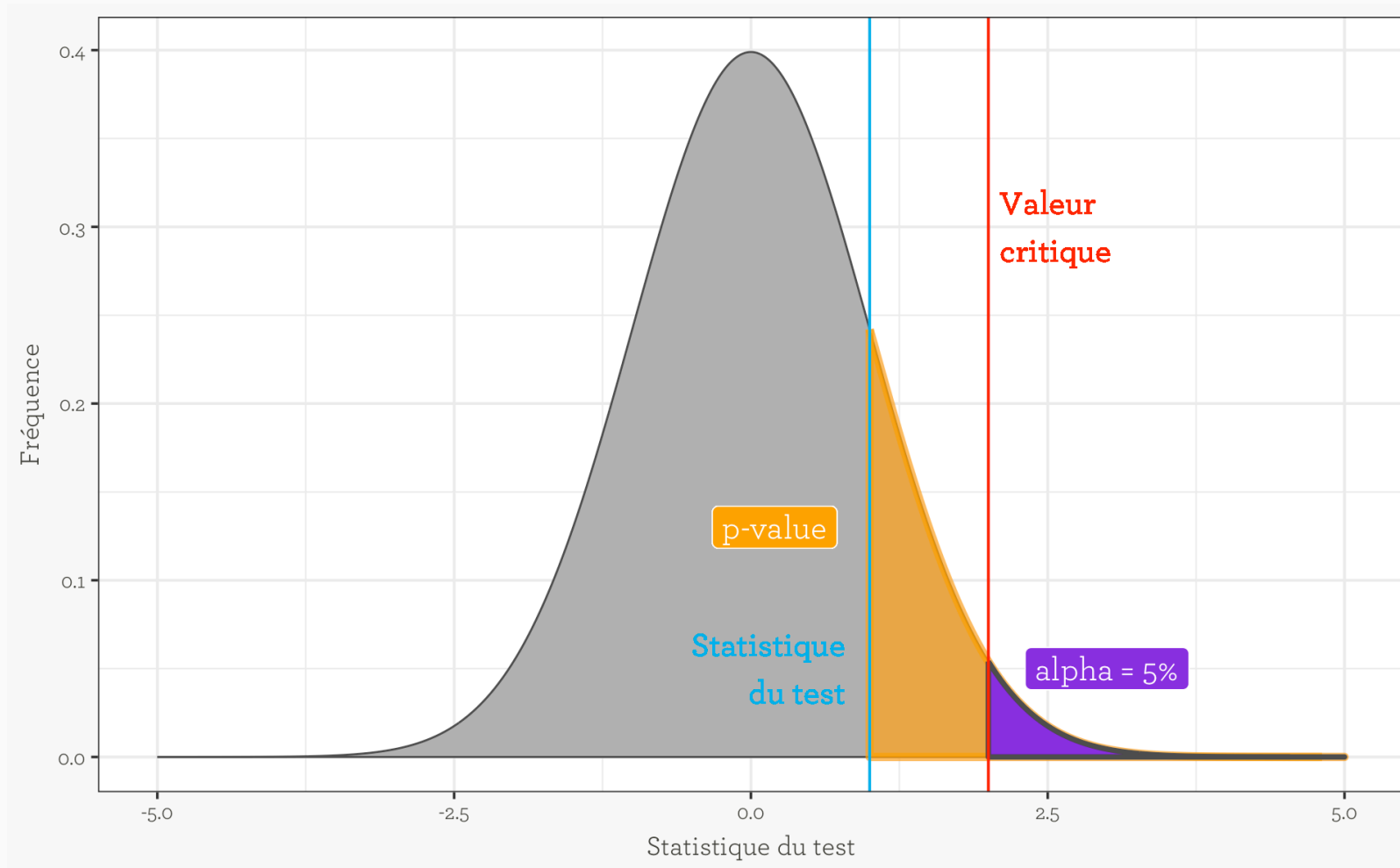
Aujourd'hui, on n'a plus besoin des valeurs critiques. On n'a plus besoin des tables statistiques : on utilise la p -value.



Principe des tests statistiques

La p -value

Mettons tout ça sur le même graphique :



Principe des tests statistiques

La p -value

Il y a donc équivalence entre les éléments suivants :

- Statistique du test $<$ valeur critique $\Leftrightarrow p\text{-value} > \alpha$
- Statistique du test \geq valeur critique $\Leftrightarrow p\text{-value} \leq \alpha$

! Important

Quand la p -value est inférieure ou égale au seuil α (généralement fixé à 5% ou 0.05 dans le domaine des Sciences du Vivant), on rejette l'hypothèse nulle H_0 du test statistique.

Sinon, on ne peut pas rejeter H_0 .

Dans **R**, toutes les fonctions permettant de réaliser des tests statistiques renvoient, au minimum, la valeur de la statistique du test, et la p -value associée.

À vous de choisir H_0 et H_A et de fixer α **avant** de réaliser le test.

Principe des tests statistiques

La p -value

Définition : la p -value

- > Ça n'est pas la probabilité que H_0 soit vraie ou fausse.
- > C'est la probabilité, si H_0 est vraie, d'obtenir par hasard, un effet au moins aussi grand que celui qu'on a observé.

Exemple

Échantillon	Moyenne
A	$\bar{x}_A = 10$
B	$\bar{x}_B = 11$

H_0 : les moyennes des populations A et B sont égales, $\mu_A = \mu_B$

H_A : les moyennes des populations A et B sont différentes, $\mu_A \neq \mu_B$

- > Si p -value = 0.021, on rejette H_0
- > Si p -value = 0.49, on ne peut pas rejeter H_0 .

3. Tests d'hypothèses : les notions importantes

Notions importantes

Le seuil de significativité

C'est le seuil α . On le choisit une fois pour toutes avant de réaliser les tests.

À moins d'avoir une bonne raison de faire autrement, on fixe $\alpha = 0.05$.

α est également appelé l'**erreur de type I**.

Définition : erreur de type I

C'est la probabilité de rejeter à tort H_0 .

Autrement dit, puisque que tout accusé est présumé innocent (l'innocence de l'accusé est l'hypothèse nulle), α est la probabilité de **condamner un innocent**.

Notions importantes

Les 2 types d'erreurs

Il existe un autre type d'erreur : l'erreur de type II, ou **erreur β** .

Définition : erreur de type II

C'est la probabilité d'accepter à tort H_0 .

Autrement dit, c'est la probabilité de **relâcher un coupable**.

L'erreur de type II **n'est pas sous notre contrôle** ! Elle dépend notamment :

- De la variabilité des données
- De la taille de l'échantillon
- Du type de test statistique réalisé

Notions importantes

La puissance statistique

Il s'agit d'une autre **probabilité**, qui dépend directement de l'erreur de type II.

On la note $1 - \beta$.

Définition : puissance statistique ($1 - \beta$)

C'est la probabilité de **détecter un effet** lorsqu'il y en a réellement un.
Autrement dit, c'est la probabilité de **condamner un coupable**.

On aimerait pouvoir maximiser la puissance. Augmenter la puissance revient à diminuer l'erreur de type II (erreur β).

Problème : diminuer β , c'est augmenter α .

Globalement, si on libère moins d'accusés, on libère moins de coupables (on baisse l'erreur de type II), mais on condamne aussi plus de monde, y compris des innocents (on augmente l'erreur de type I).

Notions importantes

La puissance statistique

Les leviers dont nous disposons pour **augmenter la puissance** sont les mêmes que ceux dont nous disposons pour diminuer l'erreur de type II :

- Augmenter la **taille** de l'échantillon
- Faire des tests **unilatéraux**
- Faire des tests **paramétriques**

Notions importantes

Tests unilatéraux et bilatéraux

Les tests *unilatéraux* sont plus **puissants** que les tests *bilatéraux*.

Toutefois, à moins d'avoir une bonne raison de faire le contraire, on choisit de préférence les **tests bilatéraux**.

Reprenons l'exemple examiné plus tôt :

Exemple

Échantillon	Moyenne
A	$\bar{x}_A = 10$
B	$\bar{x}_B = 11$


L'hypothèse nulle est toujours H_0 : les moyennes sont égales, $\mu_A = \mu_B$

Pour l'hypothèse alternative, nous avons **3 choix possibles** :

1. Hypothèse bilatérale, $H_A : \mu_A \neq \mu_B$
2. Hypothèse unilatérale, $H_A : \mu_A > \mu_B$
3. Hypothèse unilatérale, $H_A : \mu_A < \mu_B$

Notions importantes

Tests unilatéraux et bilatéraux

Pour chaque essai, un participant examine la photo d'une fille  et de deux hommes adultes, dont l'un est le père de la fille.

Le participant doit deviner quel homme est le père.

► H_0 : Il n'y a **aucune ressemblance** entre les pères et les filles : les participants identifient le père correctement **la moitié du temps** ($p = \frac{1}{2}$).

► H_A : Il y a **une ressemblance** entre les pères et les filles : les participants identifient le père correctement **plus souvent que la moitié du temps** ($p > \frac{1}{2}$).

Définition

Dans un **test unilatéral**, l'hypothèse alternative inclut des valeurs de paramètre uniquement **d'un côté** de la valeur spécifiée par l'hypothèse nulle.

H_0 est rejetée uniquement si les données s'écartent de celle-ci **dans la direction** spécifiée par H_A .

Notions importantes

Tests paramétriques et non paramétriques

Définition : test paramétrique

Un test paramétrique est un test qui suppose que **les données respectent** un certain nombre de **conditions** qu'il conviendra de vérifier avant de réaliser le test (ou parfois après avoir réalisé le test, comme pour la régression linéaire et l'analyse de variance).

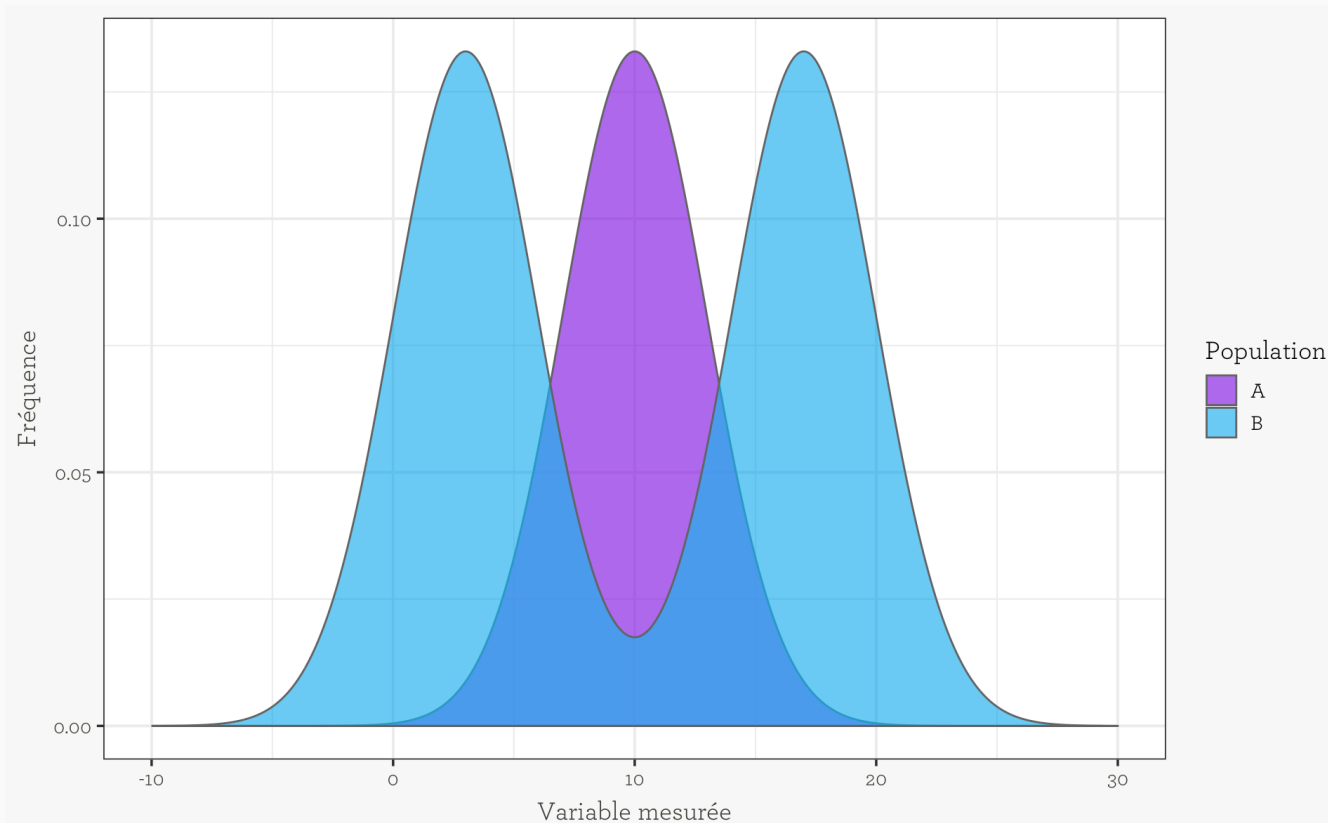
L'une de ces conditions est très souvent la **normalité des données**.

Les tests paramétriques sont **plus puissants** que les tests non paramétriques car les statistiques de ces tests sont calculées à partir des données observées **non modifiées**.

Notions importantes

Tests paramétriques et non paramétriques

Faire un test paramétrique pour comparer la moyenne de ces 2 populations n'aurait pas de sens :

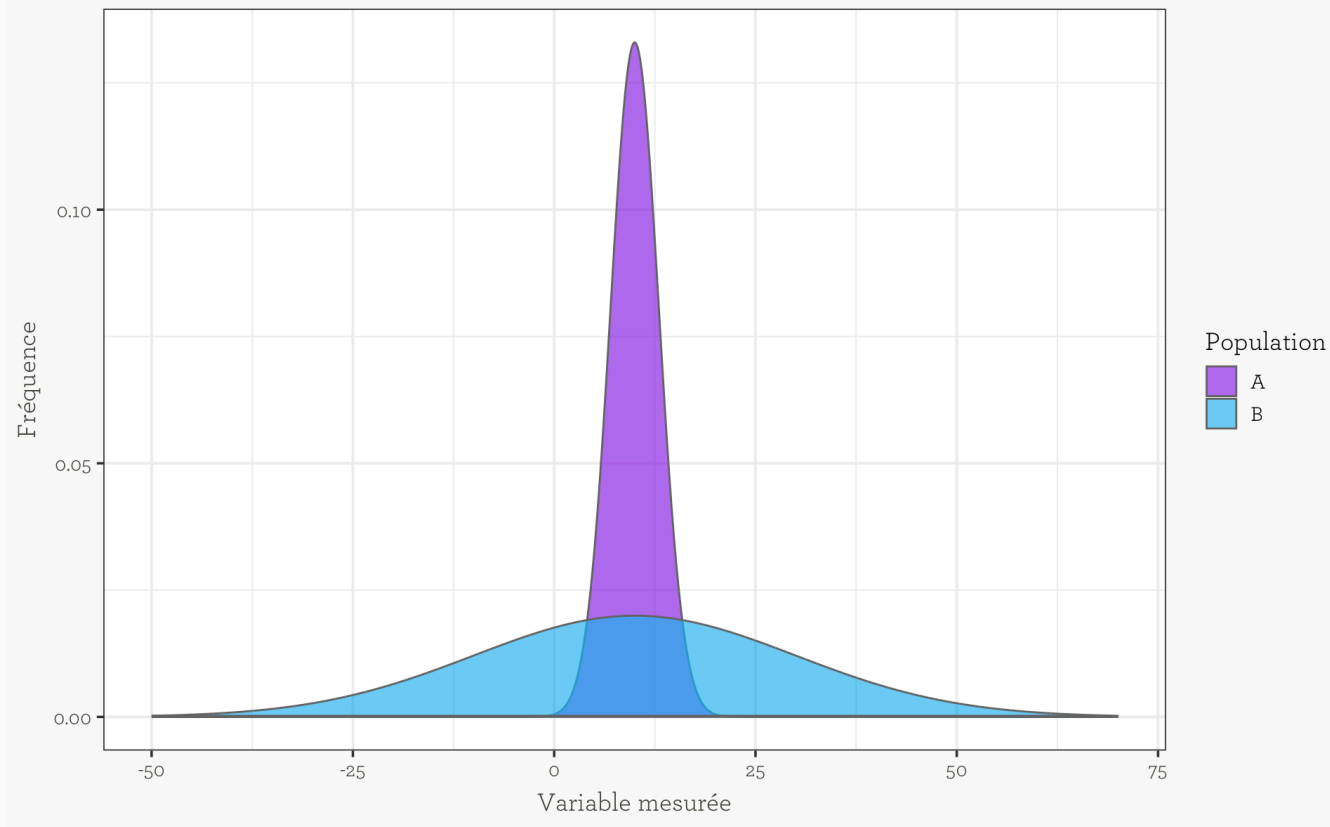


La moyenne de ces 2 populations vaut 10. Pour autant, peut-on dire que ces 2 populations ont les mêmes caractéristiques ?

Notions importantes

Tests paramétriques et non paramétriques

Même chose pour ces 2 populations :



Ces 2 populations ont une distribution normale et même moyenne. Mais leurs variances diffèrent énormément.

Bibliographie

Walters C, Maguire J-J (1996) Lessons for Stock Assessment from the Northern Cod Collapse. Reviews in Fish Biology and Fisheries 6:125-137.

Whitney WO, Mehlhaff CJ (1987) High-Rise Syndrome in Cats. Journal of the American Veterinary Medical Association 191:1399-1403.