



# DataBird

**Statistiques** : Les estimateurs statistiques et  
l'inférence



# La notion de population et d' échantillon

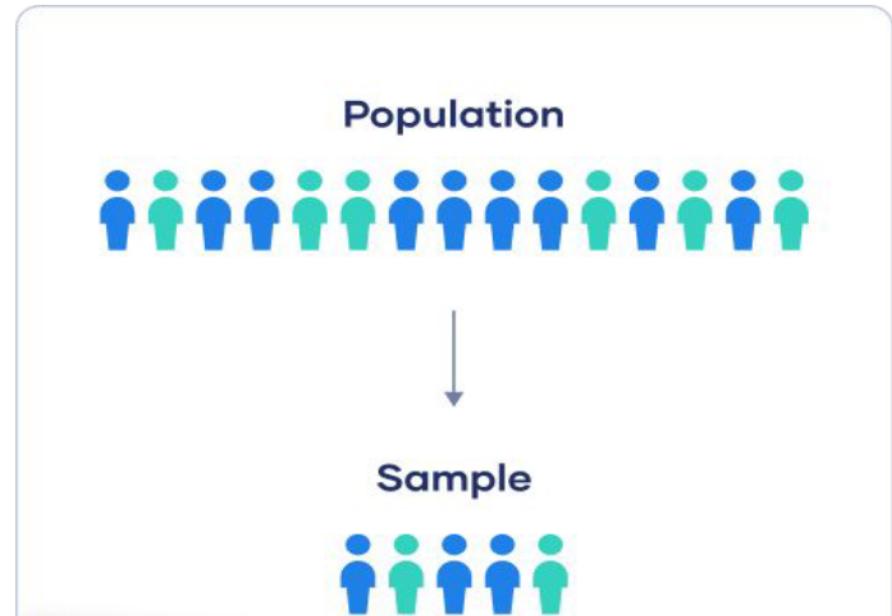
# Population et Échantillon

- La population fait référence à l'ensemble complet de toutes les entités ou individus qui sont étudiés.
- L'échantillon, quant à lui, est un sous-ensemble **représentatif** de la population qui est sélectionné pour l'étude.

Objectifs: prédire/comprendre le comportement d'une population en étudiant un échantillon

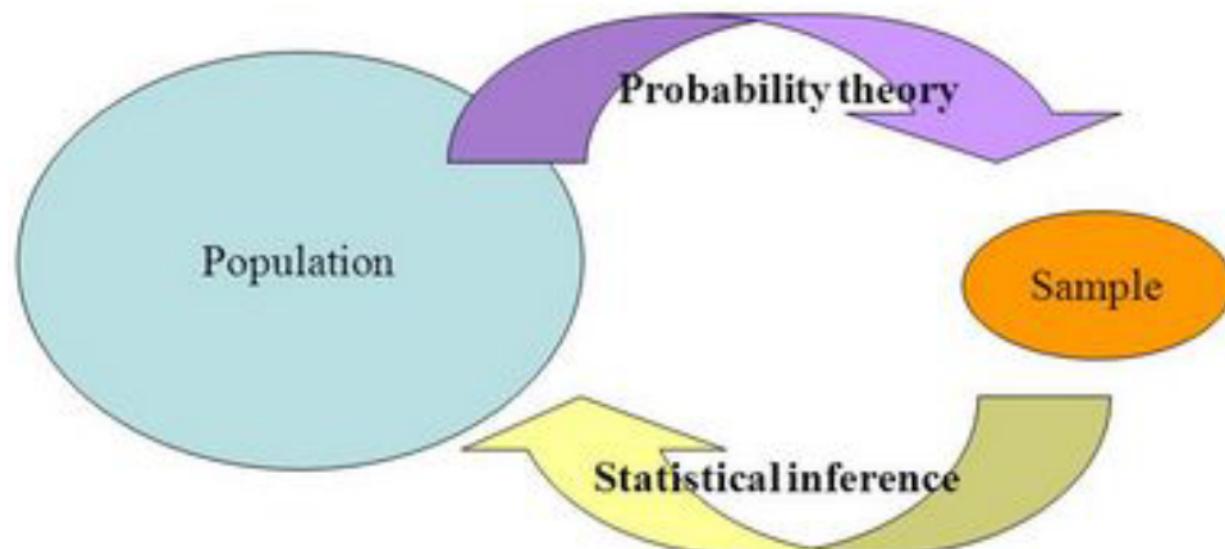
Avantages: Moins coûteux

Limites: Avoir un échantillon qui représente bien la population



# Inférence statistique

Objectif général: tirer des conclusions sur la population étudiée à partir d'observations (un échantillon) obtenues de la population.



# Quelle est la taille moyenne des étudiants aux Pays-Bas?

1. **Définir notre population** : les personnes qui sont "undergraduates" dans des universités aux Pays-Bas
2. **Choisir la taille de son échantillon** : sachant que notre population est d'environ 1% de la population des Pays-Bas, j'estime ma population à 170.000.  
J'utilise par exemple le 'ideal sample size calculator' de qualtrics (<https://www.qualtrics.com/blog/calculating-sample-size/>) qui m'estime une population idéale de 384 personnes.  
Dans un soucis de coûts je choisis de demander à 300 répondants volontaires dans des universités différentes.
3. **Choisir son échantillon aléatoirement** : Le fait de ne pas sélectionner des personnes autrement que par le fait qu'ils soient étudiants, quelque soit leur université, leur sexe, leur âge, etc. implique que je sélectionne des personnes de ma population de manière (plus ou moins) aléatoire.
4. **Conduire son analyse**: conduire des analyses statistiques après qu'ils aient répondu à un questionnaire.



# Sélectionner un échantillon

# Population et Échantillon

1. Etablir les objectifs de l'enquête
2. Définir la population cible
3. Déterminer les données à obtenir
4. Fixer le degré de précision
5. Définir la taille de l'échantillon
6. Puis la méthode de sélection de l'échantillon

# La taille de l'échantillon

La taille de l'échantillon joue un rôle essentiel dans la précision et la représentativité des résultats :

- Une taille d'échantillon insuffisante peut entraîner des résultats non fiables ou biaisés.
- Une taille d'échantillon excessive peut entraîner un gaspillage de ressources et de temps.

La formule généralement utilisée pour calculer la taille de l'échantillon est la formule de l'échantillon aléatoire simple.

- La formule est la suivante :  $n = (Z^2 * p * q) / E^2$   
(où n représente la taille de l'échantillon, Z est le score de la distribution normale associé au niveau de confiance, p est l'estimation de la proportion de la population, q est  $1 - p$ , et E est la marge d'erreur acceptable).

Pour plus d'infos vous pouvez consulter [cette page](#)

## Exemple

Supposons que vous souhaitez étudier la satisfaction des clients dans une entreprise et que vous voulez obtenir une estimation avec une marge d'erreur maximale de 5% et un niveau de confiance de 95%.

- Si vous estimez que la proportion de clients satisfaits est de 50% ( $p = 0,5$ ), alors la formule devient :  
$$n = (1,96^2 * 0,5 * 0,5) / (0,05^2)$$
- En résolvant cette formule, vous obtiendrez la taille de l'échantillon nécessaire pour votre étude



# L'échantillonnage probabiliste VS l'échantillonnage non probabiliste

# L'échantillonage probabiliste

Les 5 concepts fondamentaux

- Représentativité: L'échantillonage probabiliste vise à obtenir un échantillon représentatif de la population étudiée.
- Minimisation des biais: Il réduit les biais de sélection potentiels qui peuvent affecter les résultats.
- Inférence statistique: Il permet de tirer des conclusions sur la population dans son ensemble grâce à des méthodes statistiques appropriées.
- Mesure de précision: Il fournit des estimations précises et permet d'évaluer la marge d'erreur associée.
- Techniques variées: Il comprend différentes techniques, comme l'échantillonage aléatoire simple, stratifié, par grappes et systématique, adaptées aux besoins de l'étude.

**Ce sont des échantillonnages plus coûteux à mettre en place mais plus performant**

# L'échantillonnage aléatoire simple

C'est la méthode d'échantillonnage probabiliste la plus simple

On vient sélectionner aléatoirement certains membres de notre population (sachant que chaque membre de notre population ont la même probabilité d'être sélectionné)

**Prenons un exemple parlant, chez Databird nous voulons évaluer le niveau moyen de nos alumnis pour se faire nous souhaitons envoyer ces tests à un échantillon (une partie de nos alumnis) de la population (tous nos alumnis)**

Dans notre exemple pour les étudiants aux pays bas on viendrait sélectionner aléatoirement un certain nombre de nos alumnis sans regarder la formation qu'ils ont suivis ou autres

Avantages: C'est une méthode très simple à mettre en place et moins coûteuse que certaines méthodes probabilistes.

Inconvénients: Par contre elle peut être sujet à des biais et pourrait ne pas représenter correctement une partie de la population

# L'échantillonnage systématique

Contrairement à l'échantillonnage aléatoire simple, cette méthode suit une approche systématique basée sur un intervalle prédéfini.

Dans un échantillonnage systématique, la population est ordonnée selon une variable pertinente (par exemple, l'âge, le revenu, etc.).

Un point de départ aléatoire est choisi, puis chaque élément est sélectionné à intervalles réguliers à partir de là.

Avantages: Moins coûteuse que certaines méthodes,

Risques: Il peut y avoir des risques de biais si une structure périodique est présente dans la population, qui coïncide avec l'intervalle de sélection.

# Exemple de l'échantillonnage systématique

Supposons que vous souhaitiez réaliser une enquête sur la satisfaction des clients dans un grand centre commercial.

- Si le centre commercial compte 500 boutiques, vous pouvez choisir un point de départ aléatoire, par exemple, en sélectionnant une boutique au hasard entre 1 et 10.
- Ensuite, vous pouvez sélectionner les boutiques suivantes toutes les 10 boutiques (par exemple, 11, 21, 31, etc.) jusqu'à atteindre la taille d'échantillon souhaitée.

# L'échantillonnage avec probabilités proportionnelles à la taille

Prend en compte la taille de chaque élément de la population lors de la sélection de l'échantillon. accorde une probabilité plus élevée aux éléments les plus importants → chaque élément se voit attribuer une probabilité par rapport à sa taille qui est évalué par rapport à une métrique spécifique (entreprise: nombre de personnes, chiffres d'affaires)

Les probabilités sont calculées en divisant la taille de chaque élément par la somme des tailles de tous les éléments de la population.

Avantages: Utile lorsque la taille des éléments de la population varie considérablement, Assure une meilleure représentativité en tenant compte de la taille des éléments de la population.

Inconvénients: plus complexe, repose sur l'hypothèse que la variable utilisée pour mesurer la taille est corrélée à l'intérêt de l'étude.

# Exemple

Vous menez une enquête sur les habitudes de consommation de différents groupes d'âge dans une ville. La population est composée de trois groupes d'âge : les jeunes de 18 à 25 ans, les adultes de 26 à 40 ans et les personnes âgées de plus de 40 ans.

La taille estimée de chaque groupe est de 500, 800 et 300 respectivement. Vous utilisez l'échantillonnage avec probabilités proportionnelles à la taille.

- Jeunes (18-25 ans) :  $500/1600 = 0,3125$
- Adultes (26-40 ans) :  $800/1600 = 0,5$
- Personnes âgées (plus de 40 ans) :  $300/1600 = 0,1875$

En utilisant ces probabilités, vous pouvez sélectionner des individus aléatoirement dans chaque groupe proportionnellement à leur taille respective. Cela garantit que chaque groupe d'âge est représenté de manière adéquate dans votre échantillon, reflétant ainsi la répartition de la population dans la ville.

# L'échantillonnage stratifié (par grappes)

On divise l'échantillon en groupes ou en grappes **homogènes**, et une sélection aléatoire de grappes est effectuée pour constituer l'échantillon.

Avantages :

- Pratique et économique
- Facilite l'organisation de l'échantillonnage lorsque la population est vaste et dispersée.

Inconvénients :

- Le degré de similitude des individus à l'intérieur des grappes peut entraîner une surestimation ou une sous-estimation de certaines caractéristiques de la population.
- Nécessite une connaissance préalable de la structure de la population pour former les grappes.

## Exemple

Exemple : Supposons qu'une étude soit menée sur l'opinion politique des citoyens dans un pays. On peut diviser le pays en régions géographiques (grappes) et sélectionner un échantillon aléatoire de régions. Ensuite, on peut effectuer un sondage auprès des citoyens de ces régions sélectionnées pour obtenir les données nécessaires à l'étude.

# L'échantillonnage stratifié (aléatoire)

On divise la population en sous-groupes **homogènes** appelés strates, puis on sélectionne aléatoirement des individus à partir de **chaque strate** pour constituer l'échantillon.

Avantages :

- Garantit une représentativité plus précise des sous-groupes de la population.
- Réduit la variabilité de l'échantillon (chaque strate est représentée.)
- Fournit des estimations plus précises pour chaque strate, en comparaison avec d'autres méthodes d'échantillonnage.

Inconvénients :

- Nécessite une connaissance préalable des caractéristiques de la population pour former les strates.
- Plus complexe à mettre en œuvre que d'autres méthodes d'échantillonnage
- Peut être coûteux si les strates sont difficiles à atteindre ou à identifier.

## Exemple

Supposons qu'une entreprise souhaite évaluer la satisfaction de ses clients. Elle divise sa base de clients en strates en fonction de critères tels que l'âge, le sexe, et le type de produit acheté. Ensuite, elle sélectionne aléatoirement un échantillon de clients à partir de chaque strate et les interroge sur leur satisfaction. Cela permettra à l'entreprise d'obtenir des informations précises sur la satisfaction des clients dans chaque groupe spécifique et d'identifier les domaines d'amélioration.

# L'échantillonnage non-probabiliste

5 principes fondamentaux:

- Non-représentativité : L'échantillonnage non probabiliste ne garantit pas la représentativité de l'échantillon par rapport à la population étudiée.
- Facilité et praticité : Il est choisi pour sa facilité et sa praticité dans le recrutement des participants.
- Résultats non généralisables : Les résultats ne peuvent pas être généralisés à toute la population.
- Biais de sélection : Il peut introduire des biais de sélection importants.
- Utilisation spécifique : Il est utilisé dans des situations où l'objectif est plus exploratoire, la population est difficile à atteindre ou des contraintes pratiques existent.

Ce sont des échantillonnages moins coûteux à mettre en place mais moins performant

# L'échantillonnage de commodité

C'est une méthode où les individus sont choisis en fonction de leur disponibilité et de leur accessibilité pour le chercheur. Les individus inclus dans l'échantillon sont souvent ceux qui sont facilement disponibles et pratiques à recruter.

Avantages :

- Facilité et praticité dans le recrutement des participants.
- Gain de temps et d'efforts pour le chercheur.
- Peut être utilisé lorsque les ressources sont limitées ou lorsque le recrutement de participants est difficile.

Inconvénients :

- Biais de sélection important, car les individus inclus dans l'échantillon ne sont pas sélectionnés de manière aléatoire.
- Les résultats ne peuvent pas être généralisés à la population → sélection

## Exemple

Un chercheur qui étudie l'efficacité d'une nouvelle application mobile pour la gestion du temps recrute des participants parmi ses collègues de travail et ses amis, simplement parce qu'ils sont facilement accessibles. Les résultats obtenus ne pourront pas être généralisés à l'ensemble de la population qui utiliserait réellement l'application et risquent de ne pas refléter précisément l'impact de l'application sur les utilisateurs.

# L'échantillonnage à participation volontaire

C'est une méthode d'échantillonnage où les individus choisissent de participer ou non à l'étude. Les individus s'auto-sélectionnent en fonction de leur volonté de participer.

Avantages :

- Facile à mettre en place et à réaliser.
- Approprié pour des études exploratoires ou lorsqu'il est difficile d'atteindre la population cible
- Permet de recueillir des données sur des sujets sensibles ou difficiles à aborder.

Inconvénients :

- Biais de sélection important, car les individus auto-sélectionnés peuvent différer significativement de la population générale.
- Les participants volontaires peuvent être plus motivés, engagés ou avoir des caractéristiques différentes de ceux qui choisissent de ne pas participer.

## Exemple

Une enquête en ligne est diffusée auprès du public pour recueillir des opinions sur un nouveau produit. Les personnes intéressées par le produit ou désireuses de donner leur avis volontairement répondent à l'enquête. Les résultats obtenus refléteront uniquement les opinions des personnes ayant choisi de participer, et ne seront pas représentatifs de l'ensemble de la population cible.

# L'échantillonnage au jugé

C'est une méthode où les individus sont choisis de manière subjective ou arbitraire par le chercheur, sans suivre un processus de sélection aléatoire formel.

Avantages :

- Simple et facile à mettre en œuvre.
- Peut être utilisé lorsque les ressources, le temps ou l'accès à la population sont limités.
- Peut être approprié pour des études préliminaires ou exploratoires.

Inconvénients :

- Biais de sélection important, car le choix des individus repose sur le jugement subjectif du chercheur.
- Les résultats ne peuvent pas être généralisés à la population dans son ensemble
- Soumis au biais du chercheur → idées préconçues du chercheur

## Exemple

Supposons qu'un chercheur souhaite étudier les habitudes de consommation de café des étudiants d'une université. Pour constituer son échantillon, le chercheur se rend dans une cafétéria sur le campus, il observe une table avec 10 étudiants, **8 étudiants de cette table lui semblent représentative de l'ensemble des étudiants.** Il les interroge ensuite sur leur consommation de café, leur fréquence de consommation, leurs préférences, etc.

Dans cet exemple, le chercheur a utilisé l'échantillonnage au jugé en se basant sur sa propre décision subjective pour choisir les participants à son étude.

Cependant, il est important de noter que l'échantillon obtenu par cette méthode ne sera pas représentatif de la population étudiante dans son ensemble. Les 8 étudiants sélectionnés au jugé ne reflètent pas nécessairement la diversité et les caractéristiques de l'ensemble des étudiants de l'université. Par conséquent, les résultats de l'étude peuvent être limités à ce petit groupe spécifique et ne peuvent pas être généralisés à l'ensemble de la population étudiante.

# L'échantillonnage boule de neige

C'est une méthode où les participants initiaux sont recrutés par le chercheur, puis ils aident à identifier et à recruter d'autres participants qui répondent aux critères de l'étude. Ce processus de recrutement se poursuit de manière itérative

Avantages :

- Peut être utilisé dans des situations où il n'existe pas de liste ou de base de données complète des participants potentiels.
- Peut permettre une diversité et une variété de participants dans l'échantillon.

Inconvénients :

- Biais de sélection important, car les participants sont recrutés de manière non aléatoire et dépendent des relations et des connexions personnelles.
- Le processus de recrutement en boule de neiges peut entraîner une surreprésentation de certains profils ou caractéristiques dans l'échantillon.

## Exemple

Un chercheur souhaite étudier les habitudes de consommation de drogues dans une communauté spécifique. Il recrute initialement quelques participants de cette communauté et leur demande de recommander d'autres personnes qui pourraient être intéressées par l'étude. Les nouveaux participants recommandés fournissent ensuite d'autres recommandations, et ainsi de suite. Les résultats obtenus seront spécifiques à la communauté étudiée et ne pourront pas être généralisés à d'autres populations.



# Les biais de l'échantillonnage

# Les différents types de biais

- Les biais de sous-couverture
- Les biais de réponses volontaires
- Le biais de survie
- Le biais de non réponses

## Comment réduire ces biais?

- Définir la taille de l'échantillon et la population.
- S'assurer que la population cible et le cadre d'échantillonnage sont les mêmes. Limitez au maximum la durée de l'enquête.
- Faites en sorte que les enquêtes soient faciles à remplir.
- Effectuez des suivis.
- L'échantillonnage de commodités n'est pas la meilleure option.
- Établir les objectifs de l'enquête.
- Permettez à tous les répondants de participer sur un pied d'égalité.

# Le biais de sous-couverture

**Reprenons notre exemple précédent où on s'intéresse à la taille moyenne des étudiants aux Pays-Bas**

Pour des contraintes de praticité on envoie un sondage que dans une seule université. Or cette université est une université spécialisée dans le sport et notamment le basketball.

Les étudiants de cette université seront donc sensiblement plus grands que les étudiants d'une autre université? Les résultats calculés sur cette échantillon ne seront donc pas/peu représentatifs de l'ensemble de la population

Même principe si l'on choisit une université composée uniquement d'hommes ou de femmes.

**Lorsqu'on choisit notre échantillon il faut s'assurer qu'il soit ressemblant à la majorité de la population.**

# Le biais de réponses volontaires

Ici prenons un exemple différent: Les avis Google sur les restaurants

Les personnes qui vont avoir tendance à publier un avis vont être soient:

- ceux qui ont fortement apprécié le moment et souhaitent que celui la soit reconnu comme tel → ils vont laisser un avis positif (5 étoiles)
- ceux qui n'ont pas apprécié le moment et souhaitent que l'établissement soit reconnu comme un lieu peu fréquentable → ils vont laisser un avis négatif (1-2 étoiles)

Les résultats de l'enquête ne contiennent que les personnes qui ont une forte opinion sur la question, laissant de côté le reste de la population, ce qui donne un échantillon surreprésenté.

# Le biais de survie

Prenons un autre exemple: Nous essayons de comprendre les facteurs clés de succès des entrepreneurs

Nous allons avoir tendance à définir ces principaux facteurs de réussite sur les entrepreneurs ayant réussi à développer leur business. Cependant bons nombres d'entrepreneurs qui n'ont pas réussi à transformer leur business ont exactement les memes caractéristiques

**Il ne faut pas se concentrer uniquement sur les individus qui répondent aux critères mais aussi à ceux qui n'y répondent pas**

# Le biais de non-réponses

Il se produit lorsqu'une partie des individus sélectionnés pour participer à une étude choisissent de ne pas répondre ou de ne pas participer. **Ce biais peut se produire lorsque les caractéristiques des non-répondants diffèrent de celles des répondants.**

Exemple: Une entreprise souhaite réaliser une enquête sur la satisfaction de ses clients. Elle envoie des questionnaires par e-mail à 1000 clients sélectionnés au hasard à partir de sa base de données. Cependant, seuls 300 clients retournent les questionnaires complétés. Dans ce cas, il y a un taux de non-réponse de 70 %.

Si les clients qui ont répondu au sondage ont une satisfaction généralement plus élevée que les non-répondants, les résultats de l'enquête peuvent surestimer la satisfaction globale des clients de l'entreprise.

Pour atténuer ce biais, l'entreprise peut essayer de contacter les non-répondants pour encourager leur participation ou utiliser des techniques d'ajustement statistique. **Mais il reste un biais difficile à éliminer.**

# Estimateurs

- **Un estimateur** est utilisé pour évaluer un paramètre inconnu dans un échantillon, représentatif de la population dont il dérive.
- On peut calculer une **estimation de la moyenne, de la variance, etc.**
- On évalue cet estimateur sur uniquement un échantillon, on l'appelle aussi un **estimateur ponctuel**.

Ex: Moyenne de la taille des étudiants aux Pays-Bas

→ L'échantillon permet de faire une inférence sur la taille moyenne des étudiants aux Pays-Bas en utilisant la taille moyenne de l'échantillon

# Population vs. Sample Statistics

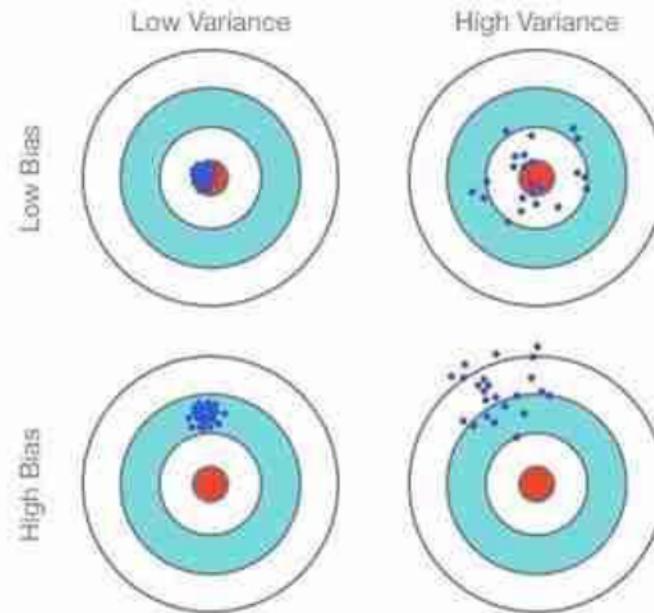
	Population	Sample
# of subjects	$N$	$n$
Mean	$\mu = \frac{\sum_{i=1}^N x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$
Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$

Note:  $S^2$  is the formula for unbiased sample variance, since we're dividing by  $n - 1$ .



# Biais VS Variance

# Bias-Variance Trade-off



# Comprendre les notions de Biais et de Variance

## Biais

- On veut que le paramètre évalué sur l'échantillon soit **égale (ou du moins le plus proche possible) de la valeur théorique** (la valeur sur le total de la population)
- La différence entre la moyenne calculée sur la base de un ou plusieurs échantillons et la valeur que l'on calcule sur le reste de la population est appelée **biais** ou **erreur systématique**.
- Si l'estimateur donne en moyenne **une valeur correcte** (exactement la valeur pour la population), alors on dit qu'il n'y a **pas de biais**.
- Un modèle de machine Learning avec peu de biais prédit en moyenne correctement les valeurs d'intérêt

# Comprendre les notions de Biais et de Variance

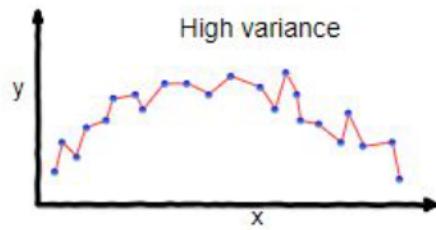
## Variance

- La variance de l'estimateur est la mesure dans laquelle l'estimateur **change** en fonction des différents échantillons.
- Un modèle avec une variance élevée **réflète trop l'échantillon** de formation et changerait beaucoup s'il était basé sur un autre échantillon.

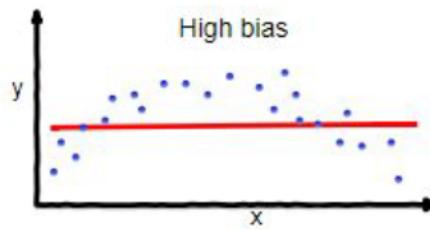
## Arbitrage Bias-Variance

- Lorsqu'on essaie de **modéliser des données** (trouver des estimateurs de la distribution des données sur la base de nos échantillons), notre modèle sera souvent soumis à un certain biais et à une certaine variance.
- Le biais et la variance sont **inversement corrélés**  
Diminuer la variance → augmenter le biais (et vice-versa).  
→ L'objectif sera d'optimiser ce compromis en repensant la modélisation : linéaire vs. non linéaire...

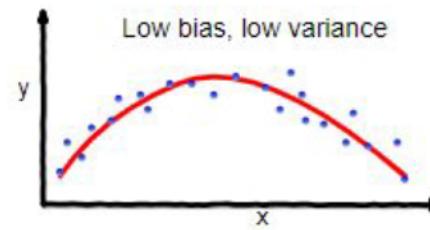
# Bias-Variance Trade-off



overfitting



underfitting



Good balance



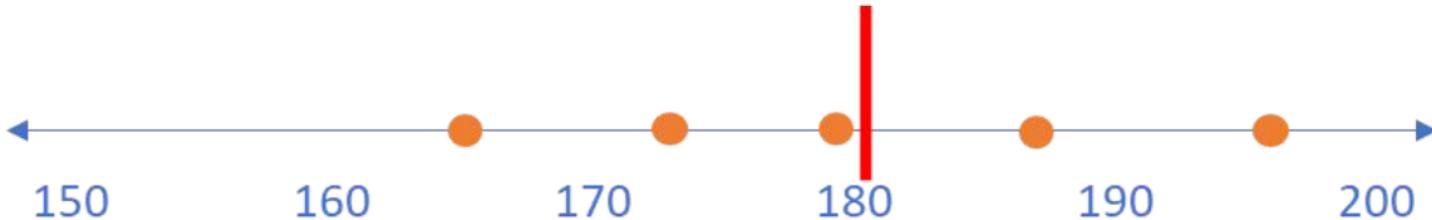
# Intervalles de confiance

# Distribution de l'échantillonnage

Avec les estimateurs ponctuels nous ne sommes **pas sûrs** d'être sur la bonne valeur des paramètres de la distribution, c'est-à-dire la **valeur réelle pour la population**, c'est-à-dire des **paramètres qui vont produire les prédictions**.

## Exemple:

Je prend un échantillon d'étudiants des Pays-Bas au hasard, je récupère leurs tailles (**points orange**) et je positionne la moyenne de leur taille (**barre rouge**).

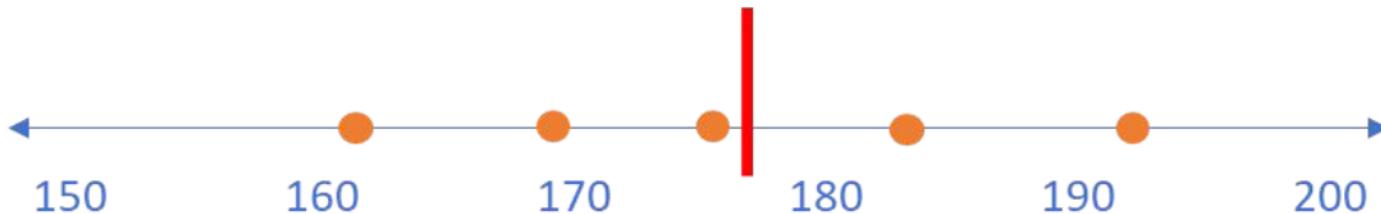


# Distribution de l'échantillonnage

Du coup j'en prend un autre pour vérifier s'il existe bien une différence si je prend un autre échantillon

## Exemple:

Je prend un autre échantillon d'étudiants des Pays-Bas au hasard, je récupère leurs tailles (**points orange**) et je positionne la moyenne de leur taille (**barre rouge**).

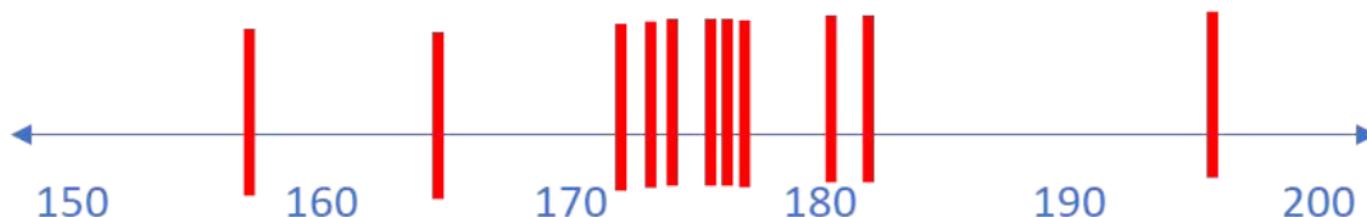


# Distribution de l'échantillonnage

Nous utilisons la théorie des probabilités pour imaginer **ce qui se passerait si nous calculions l'estimateur** (par exemple la moyenne) sur beaucoup d'autres échantillons.

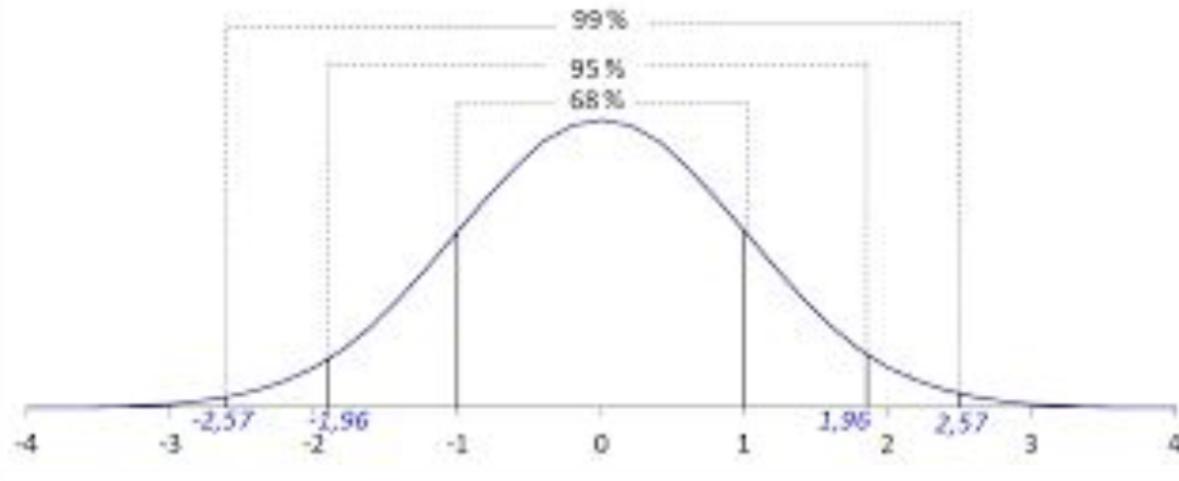
## Exemple:

Je prend un échantillon d'étudiants des Pays-Bas au hasard, je récupère leurs tailles, je calcule la moyenne de leur taille (barre rouges).



# Distribution de l'échantillonnage

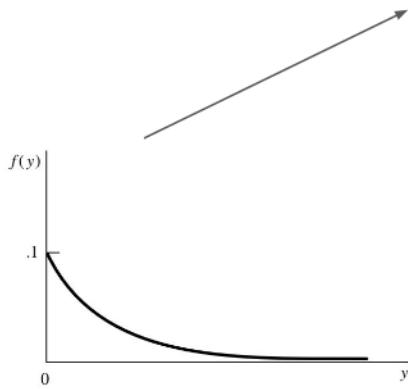
La théorie nous dit que la distribution de la moyenne sur un grand nombre d'échantillons (appelée distribution d'échantillonnage) est normale.



# Distribution d'échantillonnage : le Théorème Central Limite

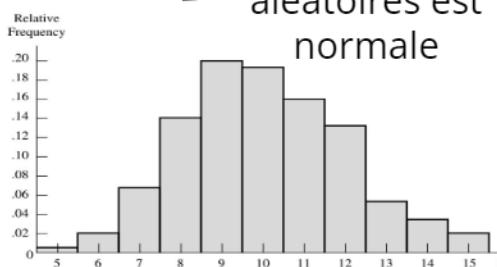
Soit  $X_1, X_2, \dots, X_i$  des variables aléatoires indépendantes et identiquement distribuées. La distribution de la somme de ces variables aléatoires tend vers une **distribution normale**.

FIGURE 7.5  
An exponential density function



Nous générerons  
30 variables  
indépendantes  
aléatoires

FIGURE 7.7  
Relative frequency histogram: sample  
means for 1000  
samples ( $n = 25$ )  
from an exponential  
distribution



La distribution de  
ces 30 variables  
aléatoires est  
normale

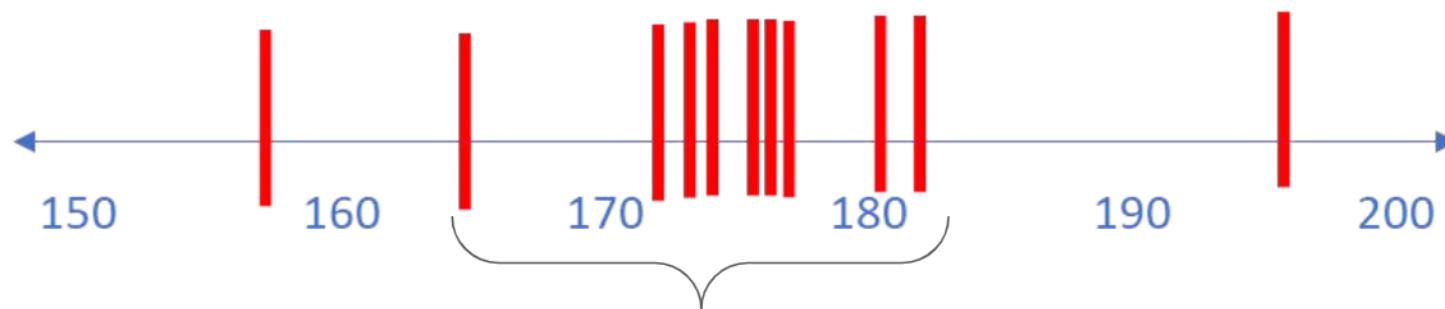
# Intervalles de Confiance

Un intervalle de confiance (de 95% par exemple) c'est simplement un intervalle qui couvre 95% des valeurs que l'on a trouvé

On sait aussi dire que tout ce qui est en dehors de l'intervalle de confiance arrive moins de 5% du temps.

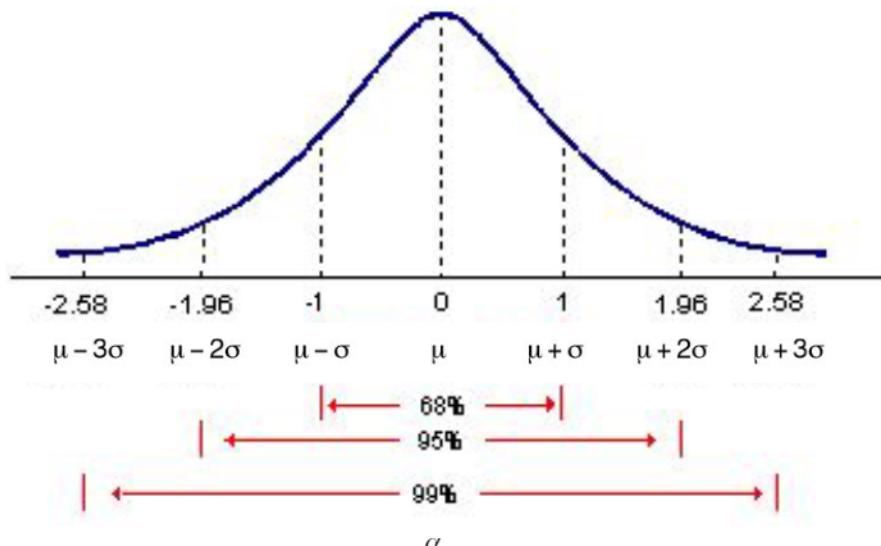
## Exemple:

Je reprend les moyennes des tailles des étudiants aux Pays-Bas, et l'accolade en noir représente 95% des valeurs que l'on a trouvé, qui sont le plus proche



# Intervalles de Confiance

- Nous pouvons donc obtenir un intervalle de confiance pour la moyenne réelle, à un niveau de **probabilité alpha** souhaité.
- Par exemple, on peut considérer que la valeur réelle de l'estimateur **se situe dans un intervalle avec  $\alpha = 95\%$**  de chances (c'est-à-dire dans 95% des échantillons).



# Bonus: Intervalles de Confiance

- Nous pouvons utiliser une **Table de Distribution Normale Standard** pour trouver l'aire cumulée sous la courbe normale standard.
- Nous introduisons cette valeur dans la formule ci-dessous pour calculer l'IC autour de la moyenne de l'échantillon :

The diagram shows the formula for a confidence interval (CI) around a sample mean ( $\bar{x}$ ):

$$CI \text{ for } \mu = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

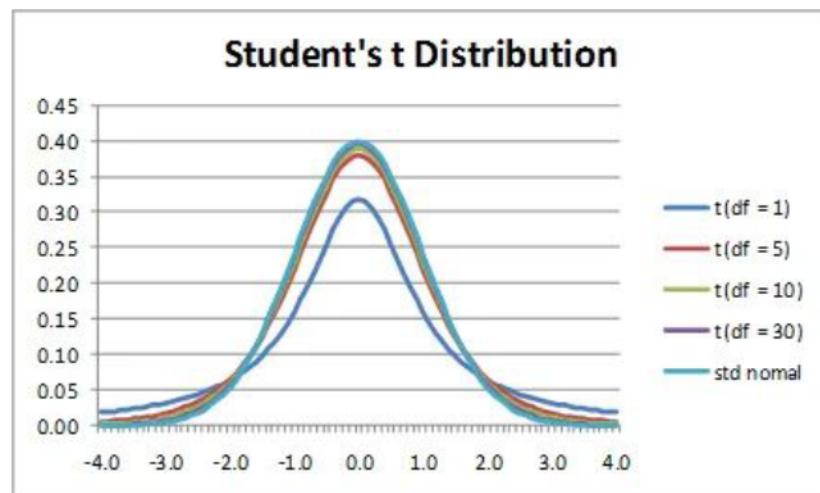
Annotations explain the components:

- Mean**: Points to  $\bar{x}$ .
- „Confidence level“**: Points to  $z^*$ .
- Variance in population**: Points to  $\sigma / \sqrt{n}$ .
- Sample size**: Points to  $n$ .
- Standard Error**: Points to the entire term  $\sigma / \sqrt{n}$ .

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

## Bonus: IC sur la moyenne en utilisant une distribution t-student

Si on ne connaît pas l'écart-type théorique  $\sigma$  (le plus fréquent), on utilise l'écart-type empirique et on approxime la distribution normale avec la distribution t de Student.



Confidence Interval =

$$\bar{x} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

Where  $t_{n-1,\alpha/2}$  is the critical value of the t distribution with  $n-1$  degrees of freedom and an area of  $\alpha/2$  in each tail:  $P(t > t_{n-1,\alpha/2}) = \alpha/2$

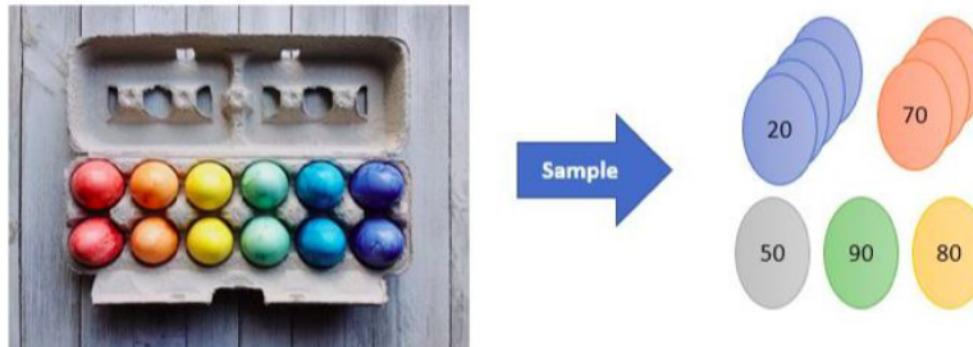


# Bootstrapping

# Exemple: Oeufs de Pâques

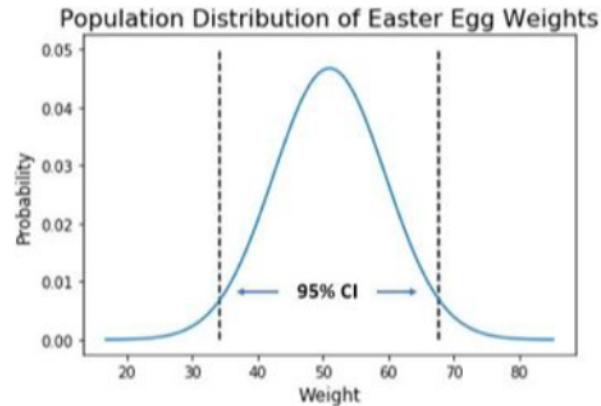
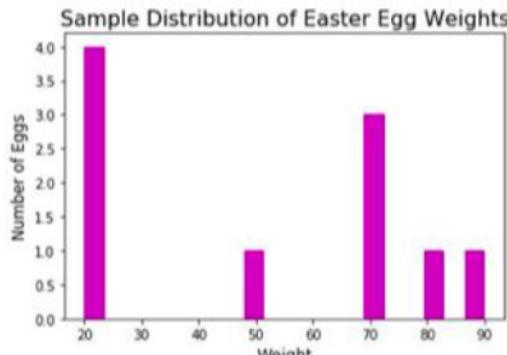
Vous avez reçu une importante cargaison d'œufs de Pâques et vous voulez déterminer le poids moyen de chaque œuf pour le contrôle de la qualité. Mais vous ne pesez pas tous les œufs. Vous ne testez qu'un échantillon de 10 œufs.

Vous obtenez:  
moyenne ( $\mu$ ) = 51,  
écart-type ( $\sigma$ ) = 27,  
erreur-type (z) = 8,53.



# Exemple: Oeufs de Pâques

La théorie des probabilités nous dit que vous pouvez multiplier cette erreur standard par 1,96 pour obtenir l'intervalle de confiance à 95%, entre 34,27 et 67,73.



**Nous sommes passés d'une distribution d'échantillon à une distribution de population.** Mais ce calcul comporte des **hypothèses cachées** : la distribution des poids est **normale**, et l'intervalle de confiance est **symétrique**. Ces hypothèses peuvent donc ne pas être vraies. Que pouvons-nous faire ?

# Rééchantillonnage : la méthode bootstrap

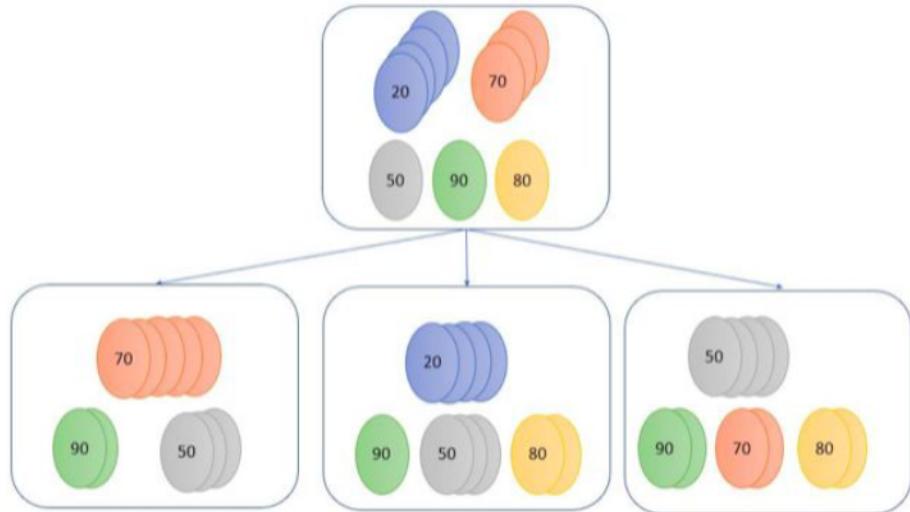
Nous utilisons l'ensemble de données existant pour simuler plusieurs ensembles de données différents (= rééchantillonnage).

(Au lieu de créer pleins d'échantillon différents à partir de l'ensemble de la population, on en crée pleins à partir du même échantillon)

Echantillons bootstrapped:

échantillons de même taille, prélevés avec remplacement (c'est-à-dire que chacun des 10 œufs a une probabilité égale d'être prélevé à chaque fois).

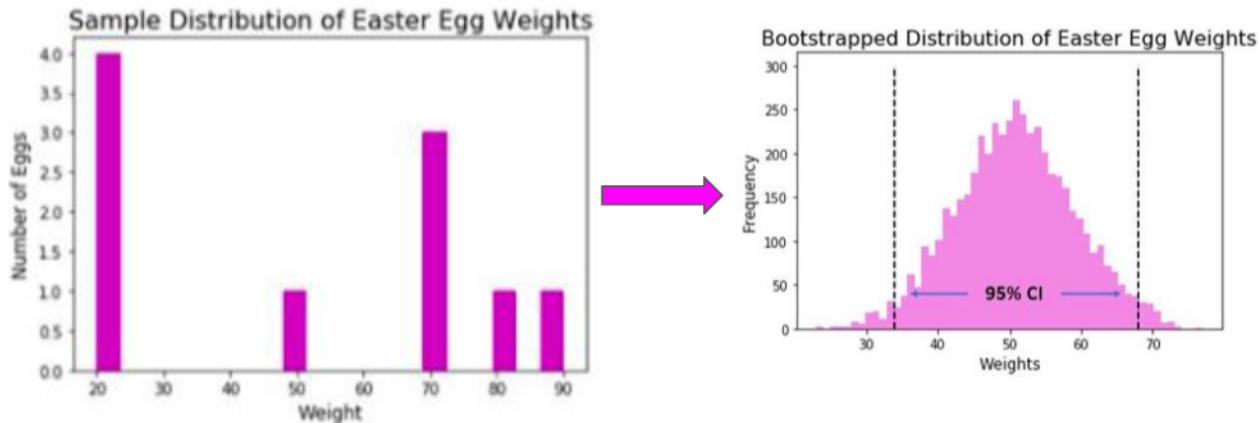
Après avoir tiré plusieurs échantillons bootstrap, nous pouvons **calculer le poids moyen pour chacun de ces échantillons.**



# Distribution Bootstrappée

En utilisant **5000 itérations**, j'obtiens un poids moyen de 50,8g avec un intervalle de confiance de 95% entre 35 et 67,03.

Remarquez que l'IC n'est pas exactement symétrique autour de la moyenne. Le résultat serait une bonne approximation de l'IC quelles que soient les hypothèses sous-jacentes.



## Pourquoi cela fonctionne-t-il ?

La méthode bootstrap permet d'estimer précisément la variation relative de la moyenne bootstrap autour de la moyenne de l'échantillon, grâce à la loi des grands nombres (grand nombre d'échantillons bootstrap). Cela permet d'approcher la variation relative de la moyenne de l'échantillon autour de la vraie moyenne.

# Avec Python: Numpy & SciPy

- "Numeric Python" et "Scientific Python" : fonctions à vocation scientifique
- Nombreux sous-packages dont `scipy.stats` et `numpy.random` :
  - Large gamme de simulations de variables aléatoires
  - Large gamme de statistiques (moyenne, stdev, médiane, etc.)
  - Large gamme d'intervalles de confiance et de tests d'hypothèses

