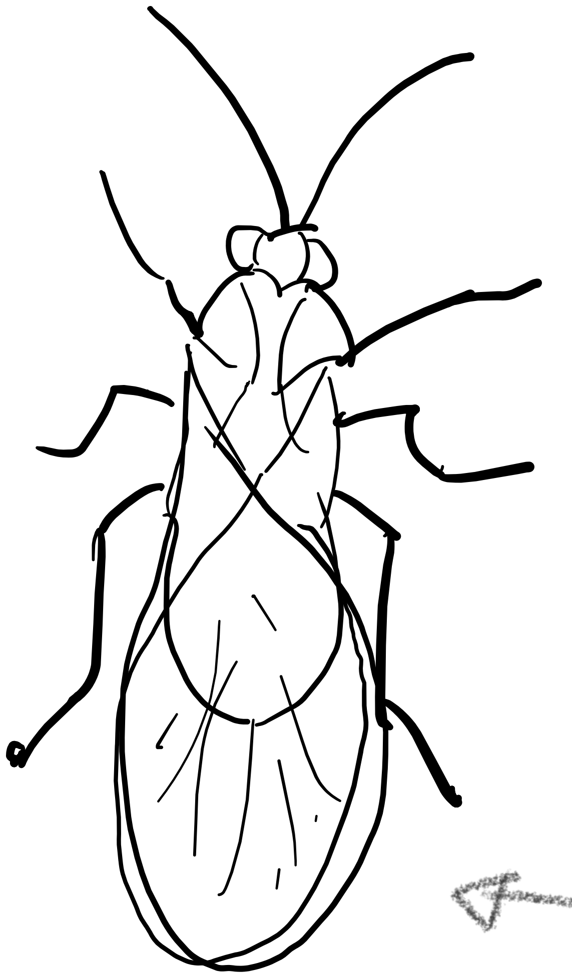
Ça,

c'est le célèbre entomologiste

Alfred KINSEY



Au cours de sa carrière **Kinsey** a amassé une collection d'environ

5,5 millions

de spécimens de **mouches-à-galles.**

Cette collection réside aujourd'hui au musée américain d'histoire naturelle, à New-York.

La collection est incomplète et désorganisée, car avant d'avoir terminé sa recherche sur les mouches-à-galles,

Kinsey s'est tourné vers...

... les comportements sexuels de l'être humain.

En 1948, puis en 1953, il publie deux volumes de la plus vaste enquête jamais menée jusqu'alors sur la **sexualité humaine.**

En tout, près de 11 300 personnes ont participé.

D'après ces données,
Kinsey affirme
qu'il y aurait, parmi
la population,

environ 10%
de personnes homosexuelles.

le 17 mai, c'est la

Journée internationale

contre l'homophobie,

la transphobie,

la biphobie...

contre l'intolérance envers la diversité sexuelle et de genre.

The Observer
Sexuality

Is 10% of the population really gay?

Drawing on the widest survey of sexual behaviour since the Kinsey Report, David Spiegelhalter, in his book Sex By Numbers, answers key questions about our private lives. Here he reveals how Kinsey's contested claim that 10% of us are gay is actually close to the mark

David Spiegelhalter

Sun 5 Apr 2015 08.00 BST

974

Changing times: students stage a kiss-in at a Sainsbury's store in Brighton last year after two gay women were threatened with ejection for kissing. Photograph: Christopher Ison

Capture d'écran le 15 mai 2022.

Comment ?

Petit guide d'autodéfense statistique

une introduction simple
aux probabilités et aux statistiques.

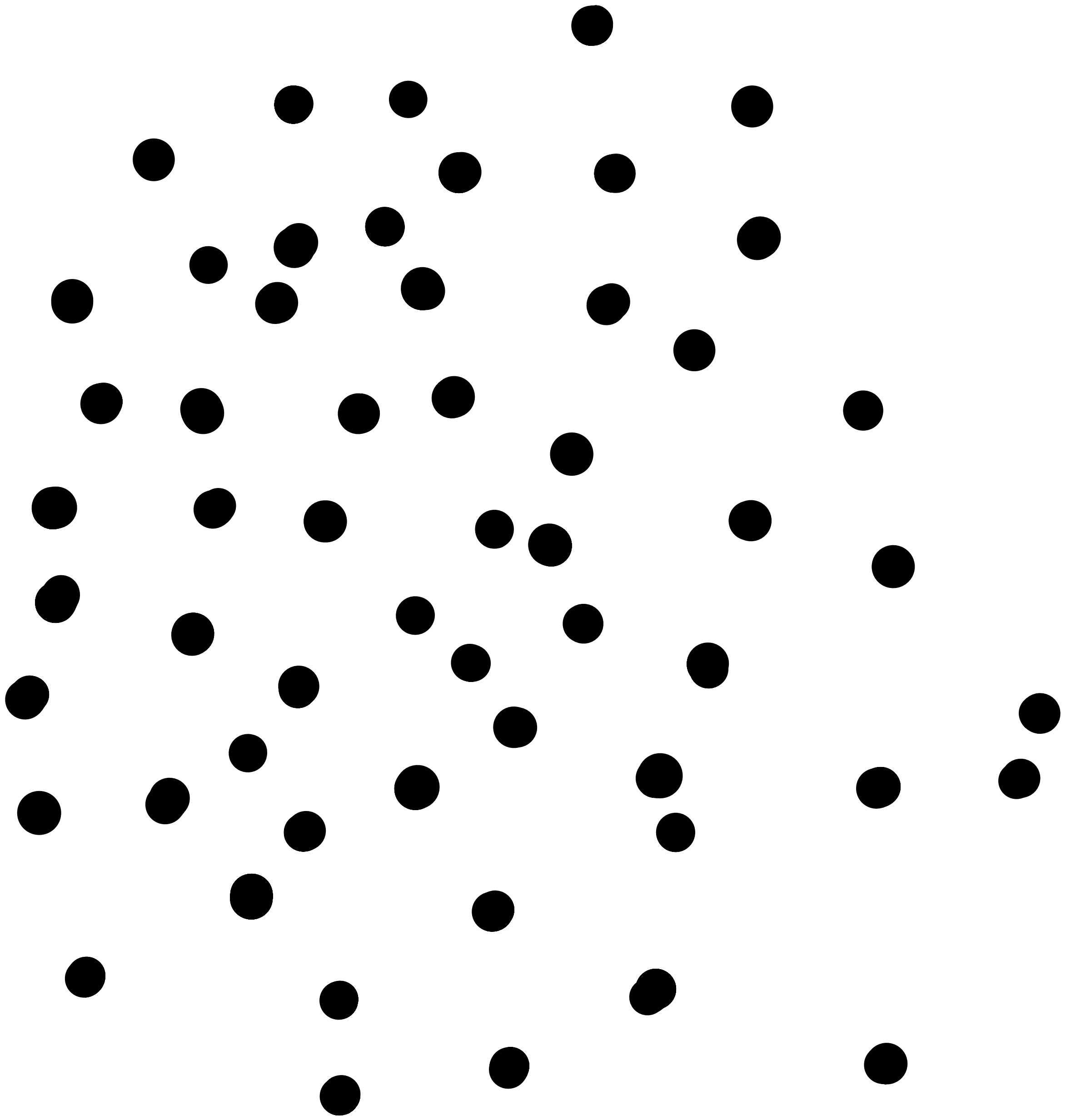
Élise Davignon, M.Sc.
doctorante et chargée de cours à l'U. de Montréal
elise.davignon@umontreal.ca

Apperçu des

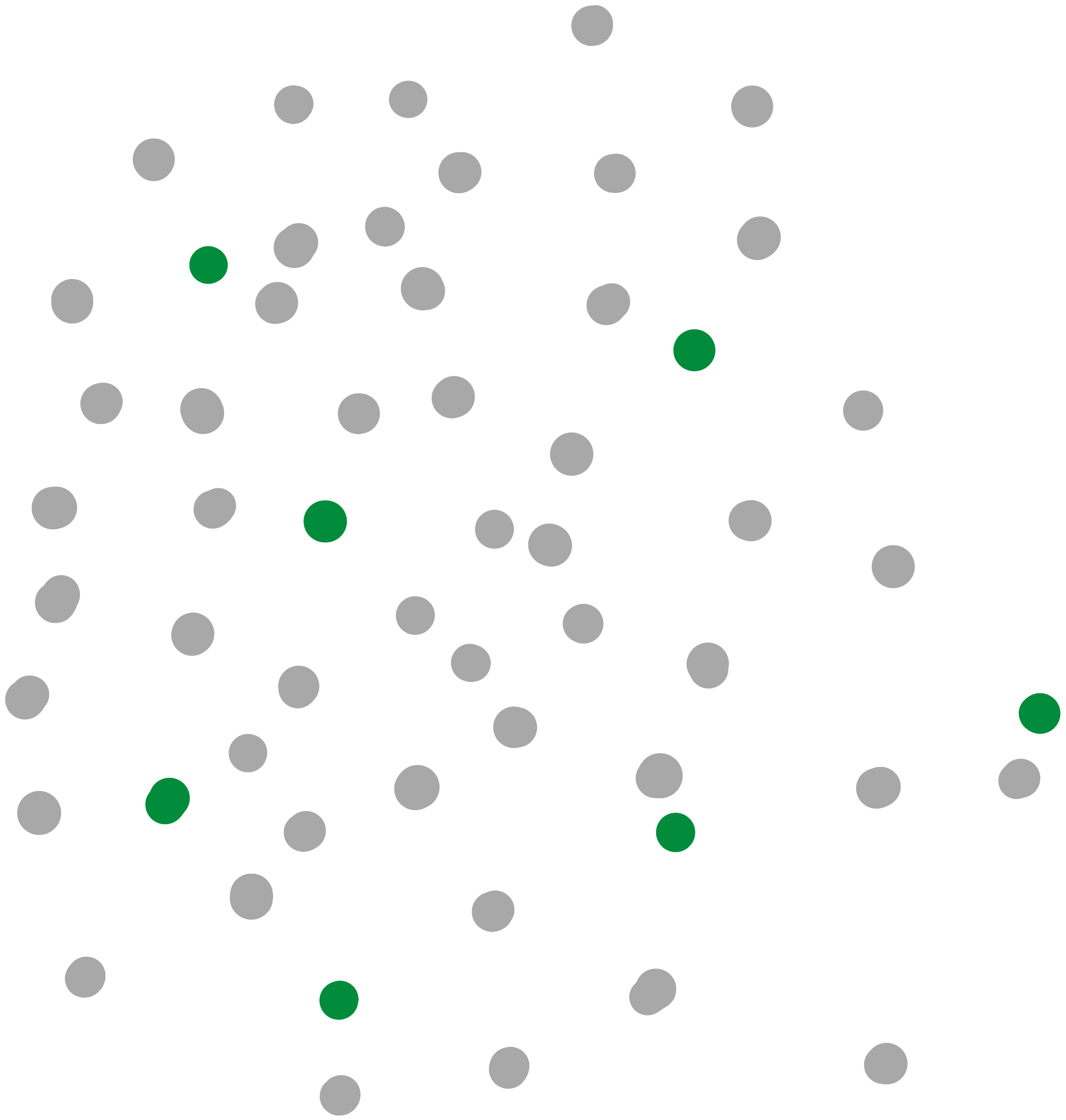
4.) Statistiques

Que nous disent
vraiment les chiffres ?

L'échantillon.



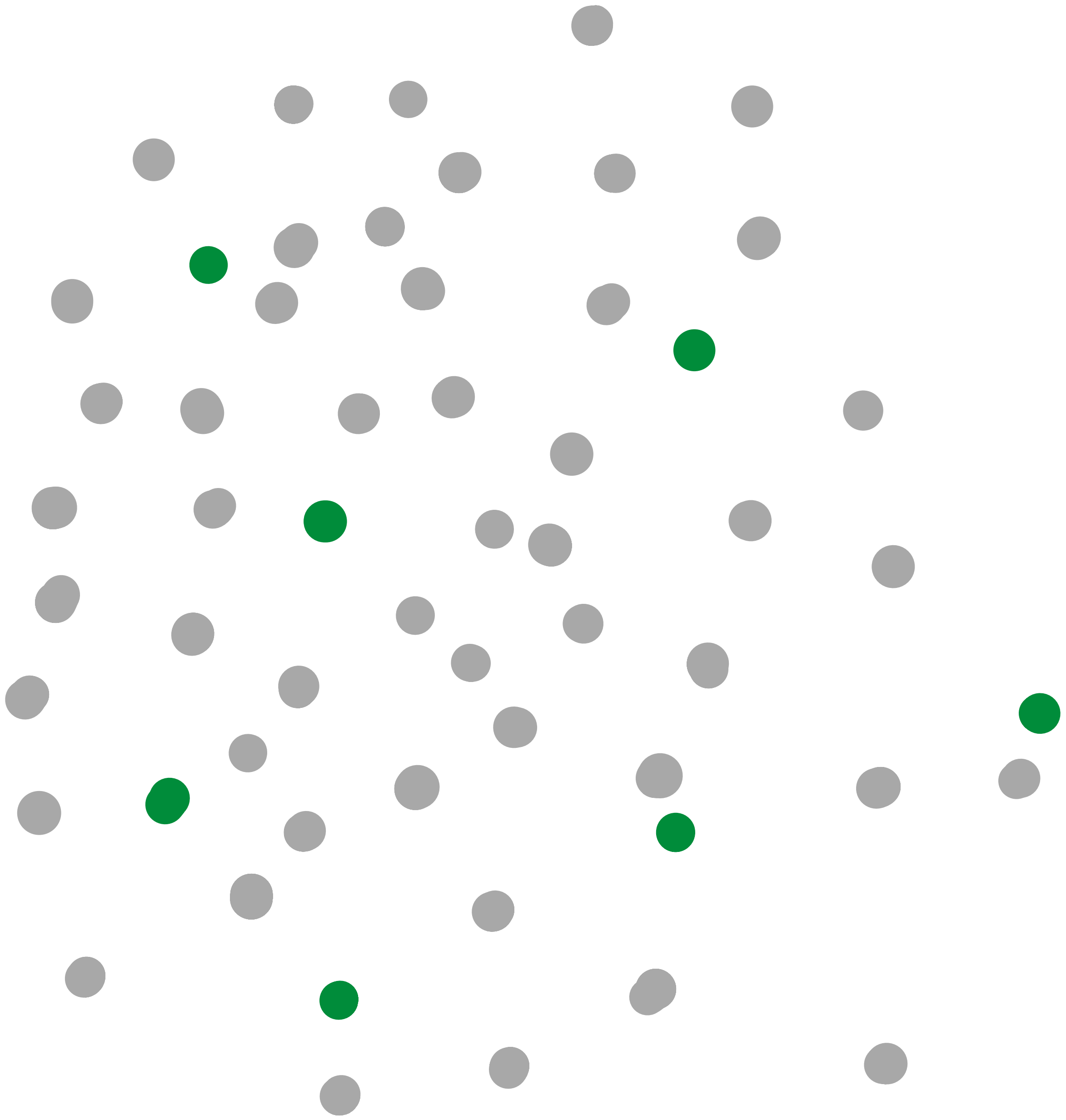
C'est difficile de demander à **tout le monde...**



... alors on prélève un **échantillon.**

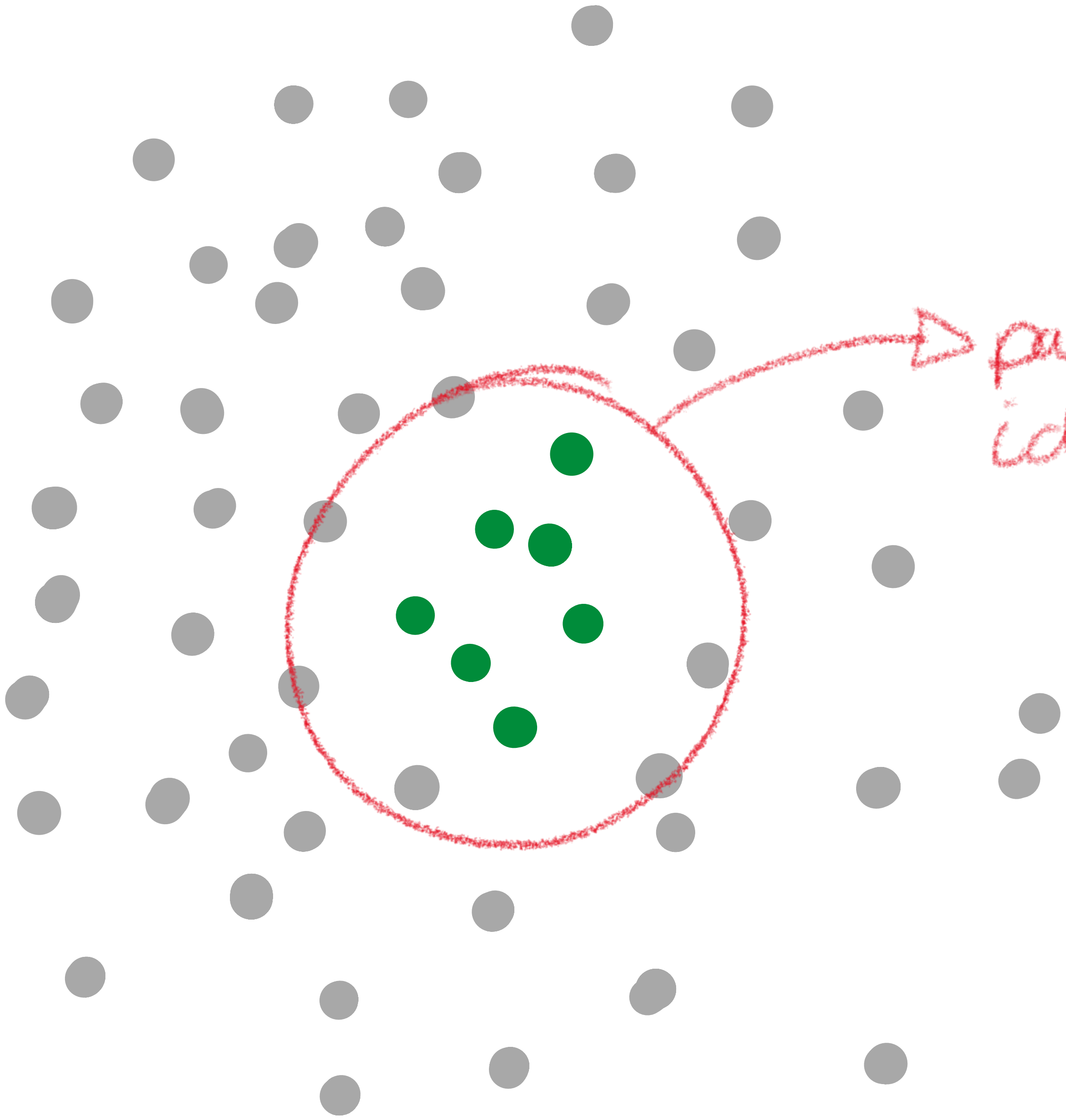
L'échantillon

est l'ensemble des individus d'une population pour lesquels on recueille des données.



... alors on prélève un **échantillon.**

on assume implicitement que l'**échantillon** et la **population** ont les mêmes *propriétés statistiques*.



pas une bonne idée.

... alors on prélève un **échantillon.**

on assume implicitement que l'échantillon et la **populationt** ont les mêmes propriétés statistiques.

Les données.

	Taille en cm.	Masse corporelle (k.g.)
Robert	177 cm	70 kg
Paul	183 cm	86 kg
Marie	169 cm	61 kg

Une donnée est une **mesure** de la valeur d'une variable pour un individu connu de l'échantillon.

Les données.

Robert
Paul
Marie

→ échantillon

	Taille en cm.	Masse corporelle (k.g.)
Robert	177 cm	70 kg
Paul	183 cm	86 kg
Marie	169 cm	61 kg

Une donnée est une **mesure** de la valeur d'une variable pour un individu connu de l'échantillon.

Les données.

Taille en cm. Masse corporelle (k.g.)

→ variables.

Robert	177 cm	70 kg
Paul	183 cm	86 kg
Marie	169 cm	61 kg

Une donnée est une **mesure** de la valeur d'une variable pour un individu connu de l'échantillon.

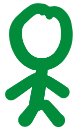
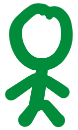
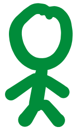
Les données.

	Taille en cm.	Masse corporelle (k.g.)
Robert	177 cm	70 kg
	183 cm	86 kg
Marie	169 cm	61 kg

données ←

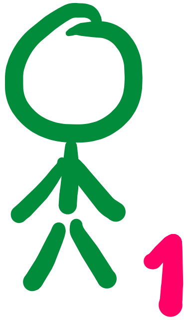
Une donnée est une **mesure** de la valeur d'une variable pour un individu connu de l'échantillon.

Les données.

	T	M
 1	t_1	m_1
 2	t_2	m_2
 3	t_3	m_3

Une donnée est une **mesure** de la valeur d'une variable pour un individu connu de l'échantillon.

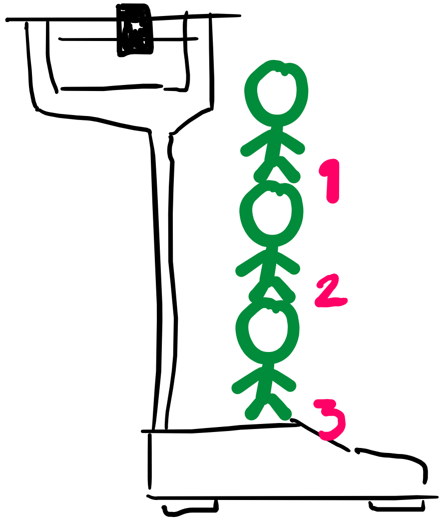
Les données.

$T($  $) = t_1 = 177\ \text{cm}$

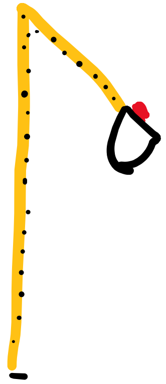
T	individu	t_1	$177\ \text{cm}$
variable	individu de l'échantillon	donnée (symbole)	donnée (valeur)

On utilise souvent des **lettres minuscules** pour les données pour les différentier des variables.

Les statistiques.



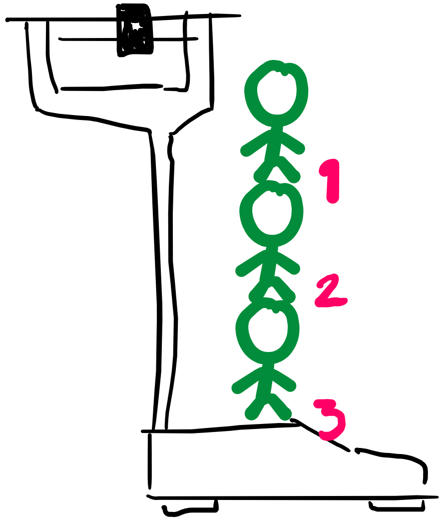
La masse totale

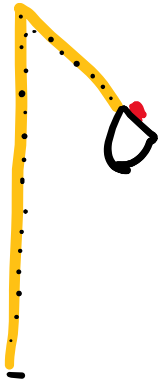


La hauteur totale

Une **statistique** est une quantité qui dépend des **données** d'un **échantillon**.

Les statistiques.



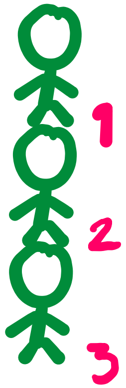


$t_{tot.}$

$m_{tot.}$

Une **statistique** est une quantité qui dépend des **données** d'un **échantillon**.

Les statistiques.

$m_{tot.}$ () $= m_1 + m_2 + m_3$

$= 217 \text{ kg}.$

La statistique

L'échantillon

sa valeur.

On dit que la statistique est une fonction des données de l'échantillon.

Des exemples.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

La **moyenne échantillonnale**, ou simplement la « moyenne » est une **statistique**.

Remarquons que les statistiques peuvent aussi dépendre de la **taille de l'échantillon** (n).

Des exemples.

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

La variance échantillonnale, et

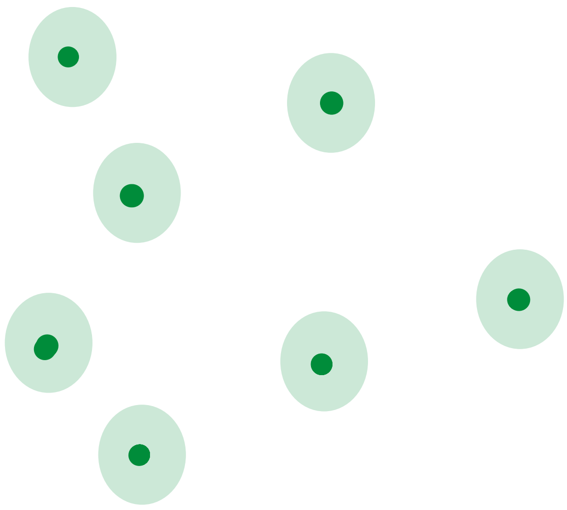
$$s = \sqrt{s^2}$$

L'écart-type échantillonnal aussi.

L'inférence.

L'inférence

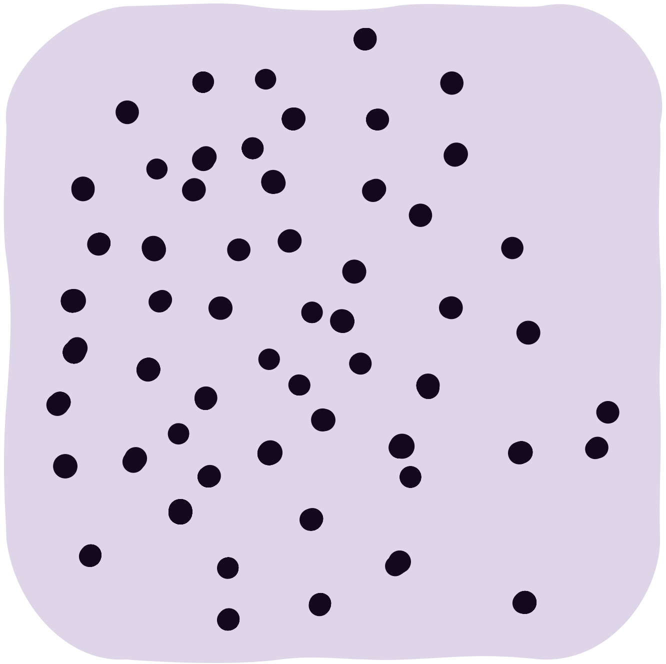
Ce qu'on **a :**



des **données** d'un **échantillon** pour diverses **variables**.

⇒

Ce qu'on **veut :**



des informations sur la **distribution** de ces **variables** dans la **population**.

L'inférence

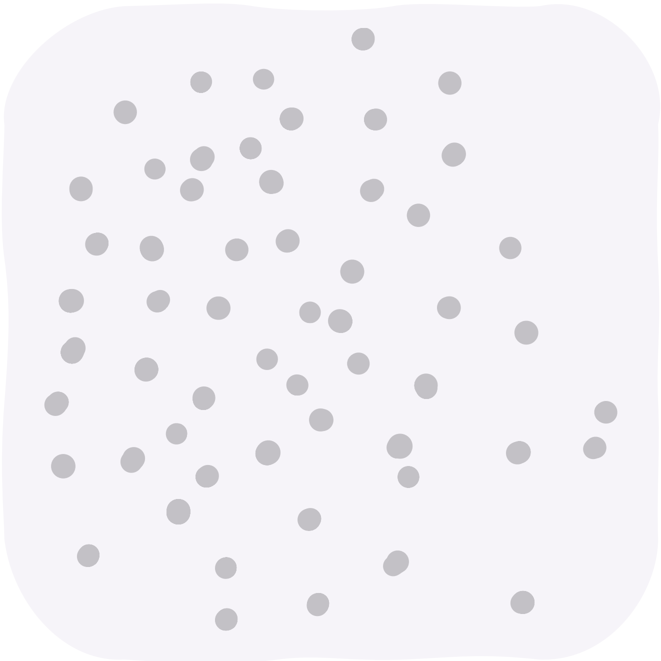
Ce qu'on **a** :

des **données** d'un **échantillon** pour diverses **variables**.

⇒

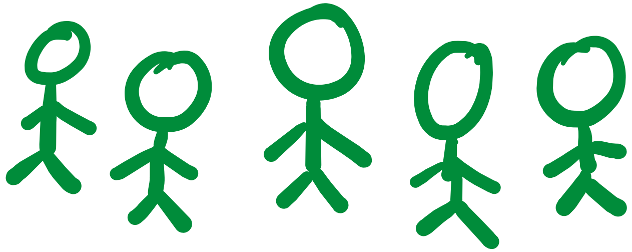
comment on **l'obtient.**

Ce qu'on **veut** :



des informations sur la **distribution** de ces **variables** dans la **population**.

Les données-modèle.



Le «vrai»
échantillon

t_1 m_1 ...
t_2 m_2 ...
t_3 m_3 ...
... ...

Les «vraies» données.

T? M?

Hypothèses et suppositions sur la **distribution** dans la **population.**

T_1 M_1 ...
T_2 M_2 ...
T_3 M_3 ...
... ...

données-modèle.

Les données-modèle.

Les **données-modèle** sont des variables aléatoires qui **simulent** un jeu de données pour un **échantillon aléatoire** de taille égale à l'**échantillon réel**, conformément à nos **hypothèses** et nos suppositions.

T? M?

Hypothèses et suppositions sur la **distribution** dans la **population.**

$$
\begin{array}{lll}
T_1 & M_1 & \dots \\
T_2 & M_2 & \dots \\
T_3 & M_3 & \dots \\
\dots & \dots &
\end{array}
$$

données-modèle.

Les données-modèle.

$t_1 \quad m_1 \quad \cdots$
$t_2 \quad m_2 \quad \cdots$
$t_3 \quad m_3 \quad \cdots$
$\cdots \quad \cdots$

$\bar{t} \quad \bar{m}$

$s_t \quad s_m$

Les **statistiques échantillonnales**

$T_1 \quad M_1 \quad \cdots$
$T_2 \quad M_2 \quad \cdots$
$T_3 \quad M_3 \quad \cdots$
$\cdots \quad \cdots$

$\bar{T} \quad \bar{M}$

$S_T \quad S_M$

Les **statistiques-modèle.**

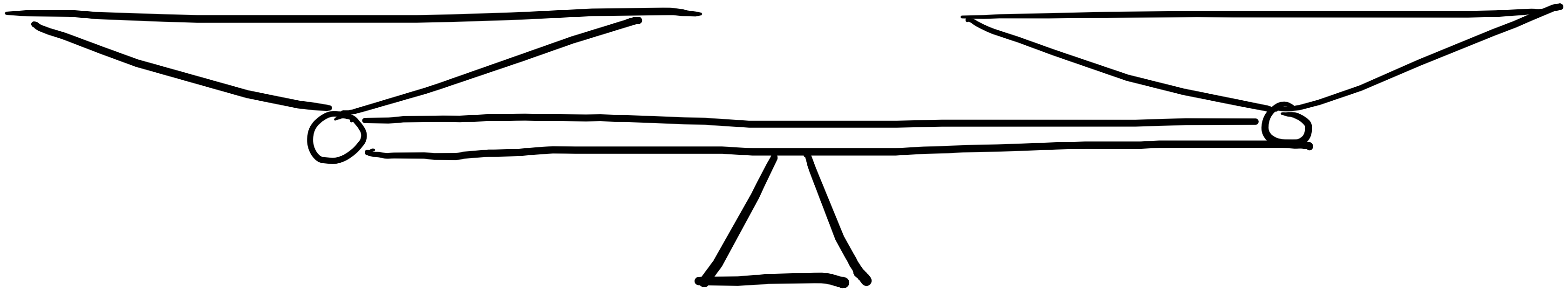
Les données-modèle.

$\bar{t}$ $\bar{m}$

s_t s_m

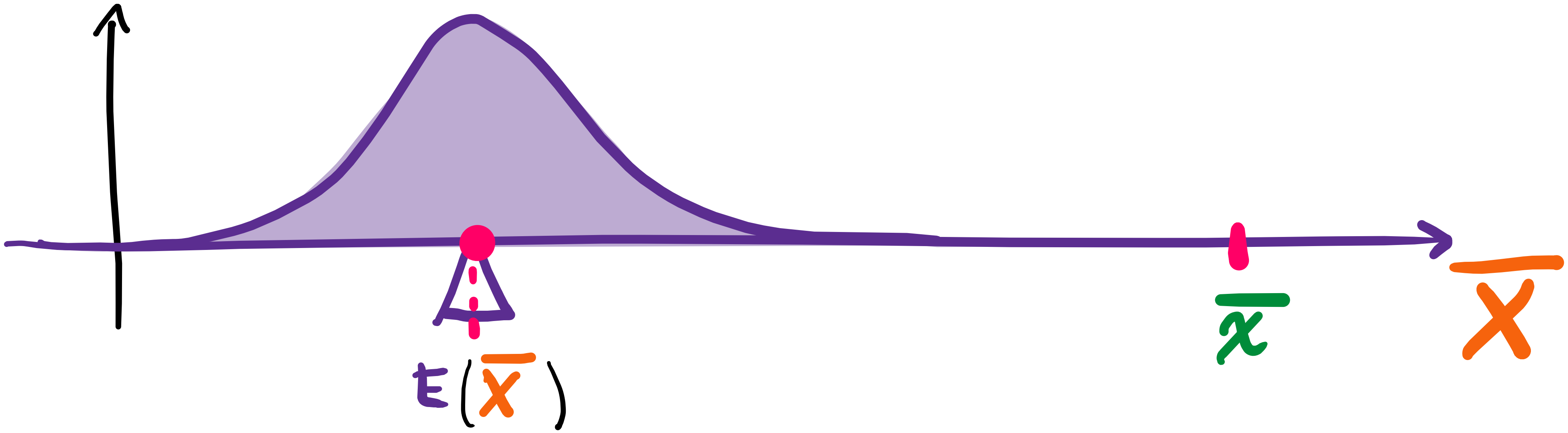
$\bar{T}$ $\bar{M}$

S_T S_M



en comparant les statistiques échantillonnales aux distributions prévues des statistiques-modèle ...

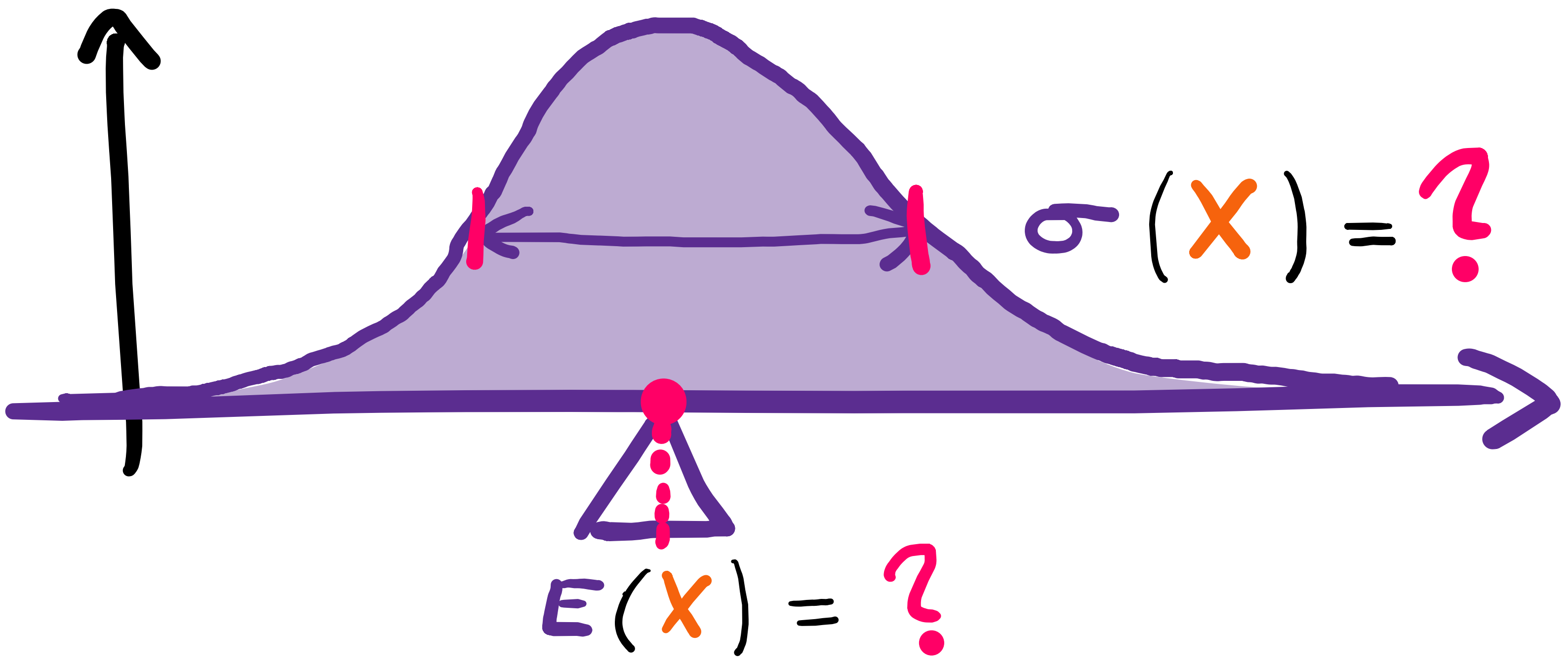
Les données-modèle.



… On peut tester notre modèle.

L'inférence paramétrique.

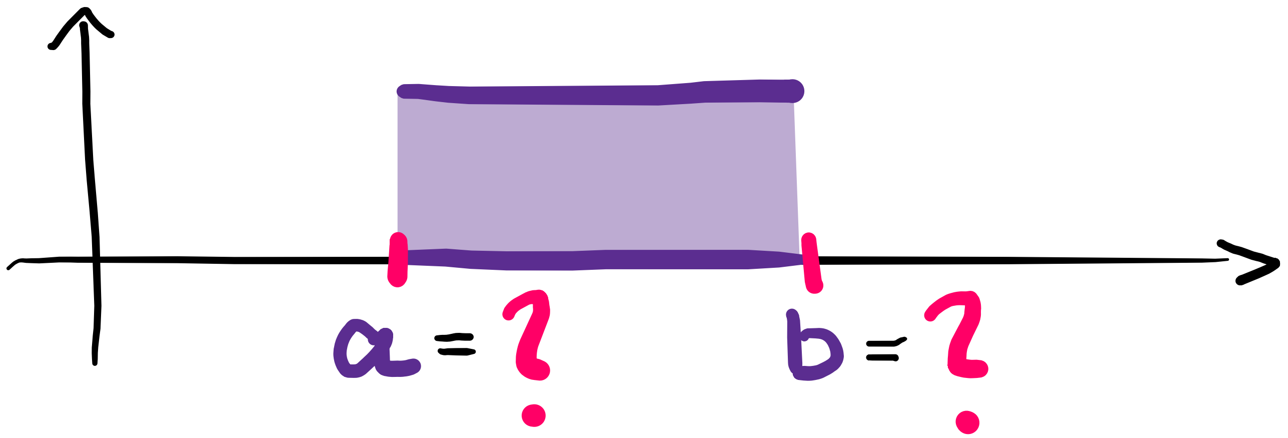
L'inférence paramétrique.



On **assume** que X suit une certaine distribution.

On cherche alors à connaître la valeur de paramètres.

L'inférence paramétrique.

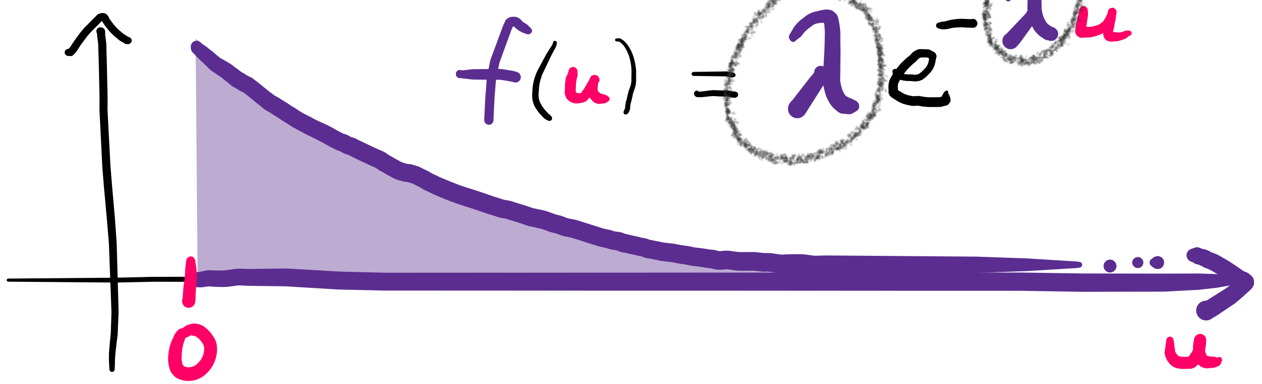


On **assume** que X suit une certaine distribution.

On cherche alors à connaître la valeur de paramètres.

L'inférence paramétrique.

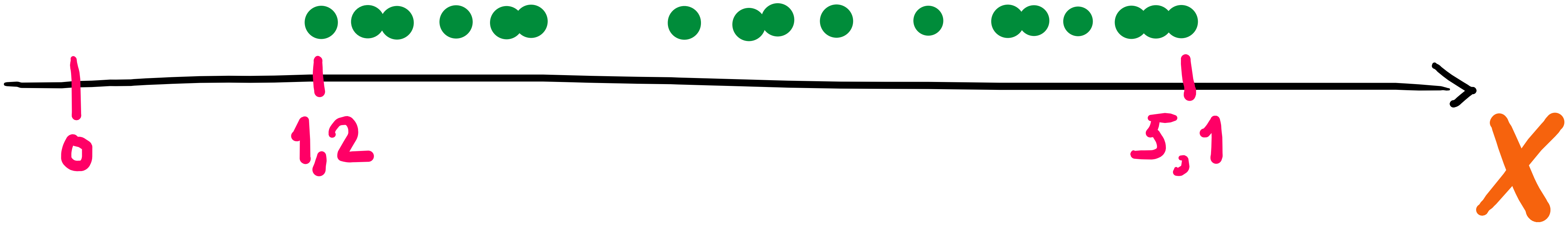
$$f(u) = \lambda e^{-\lambda u} \qquad \lambda = ?$$



On **assume** que X suit une certaine distribution.

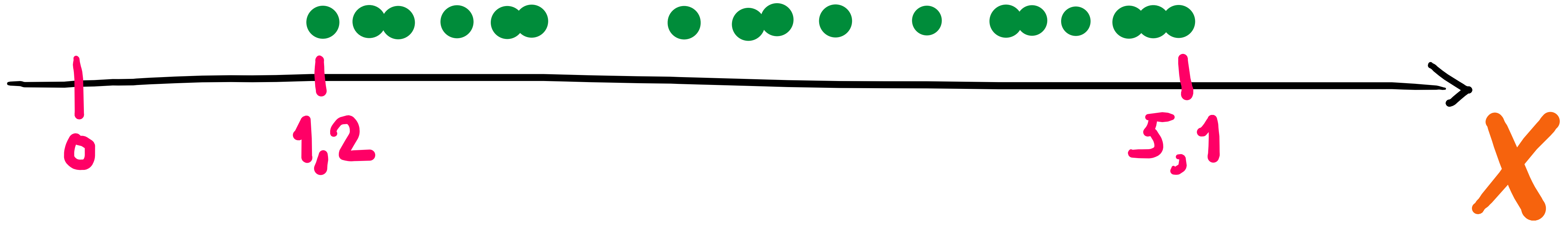
On cherche alors à connaître la valeur de paramètres.

Les estimateurs.



Avec seulement ces données, $(n=20)$
quel serait un bon estimé
de l'étendue de la distribution de X ?

Les estimateurs.

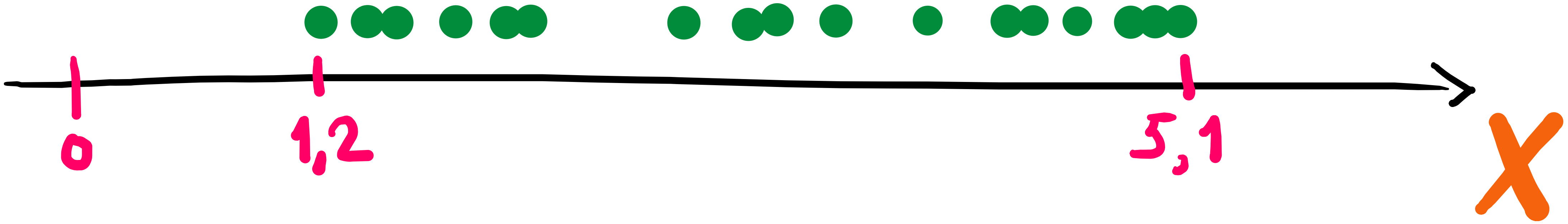


Rappel: L'étendue est la largeur de la plage de valeurs possibles.

$(n = 20)$

Avec seulement ces données, quel serait un bon estimé de l'étendue de la distribution de X ?

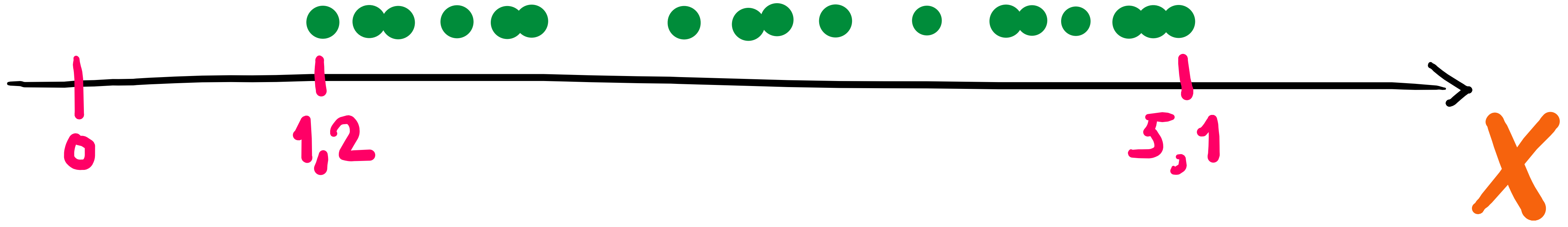
Les estimateurs.



$$\text{max} - \text{min} = 5{,}1 - 1{,}2$$
$$= 3{,}9.$$

serait un bon estimé.

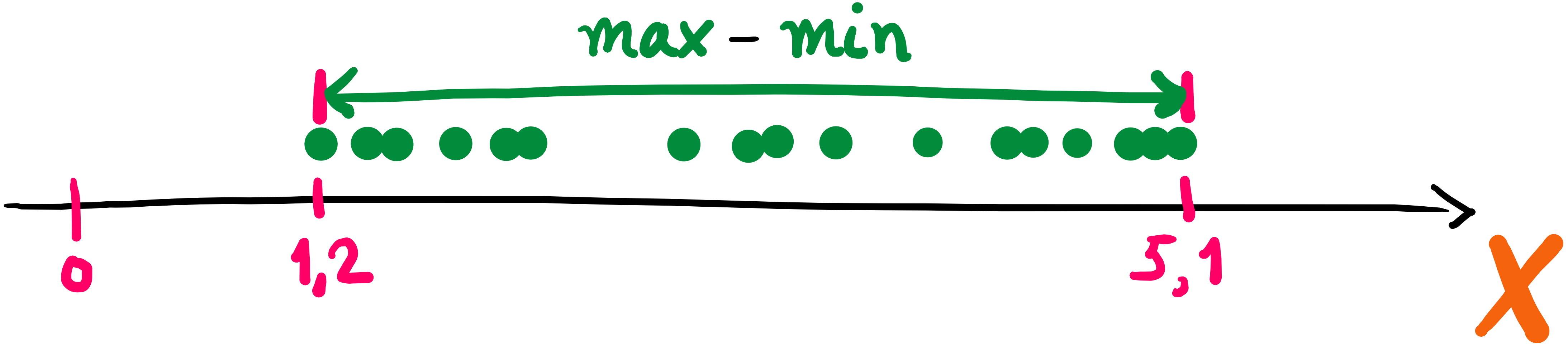
Les estimateurs.



$$(\max - \min) \times \frac{n+1}{n} = 3{,}9 \times \frac{21}{20} = 4{,}05$$

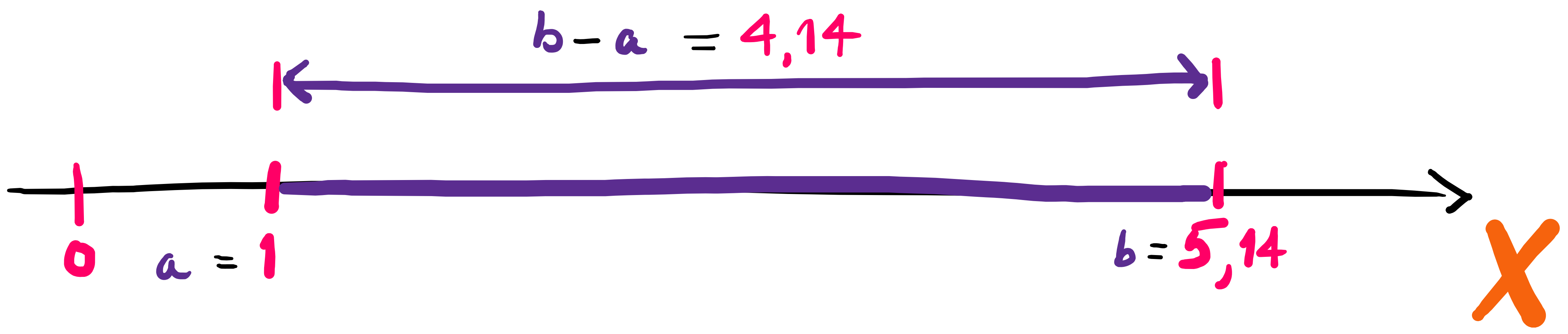
Serait en fait marginalement meilleur !
(surtout lorsque n est relativement petit).

Les estimateurs.



Un estimateur d'un paramètre est une statistique de l'échantillon qui a la propriété d'être une meilleure approximation du paramètre plus l'échantillon est grand. ($n \to \infty$)

Les estimateurs.



Un estimateur d'un paramètre est une statistique de l'échantillon qui a la propriété d'être une meilleure approximation du paramètre plus l'échantillon est grand. ($n \longrightarrow \infty$)

Les estimateurs.

$$\bar{x} \xrightarrow[n \to \infty]{} E(X)$$

La moyenne échantillonnale est un estimateur de l'espérance

$$s^2 \xrightarrow[n \to \infty]{} Var(X)$$

La variance échantillonnale est un estimateur de la variance

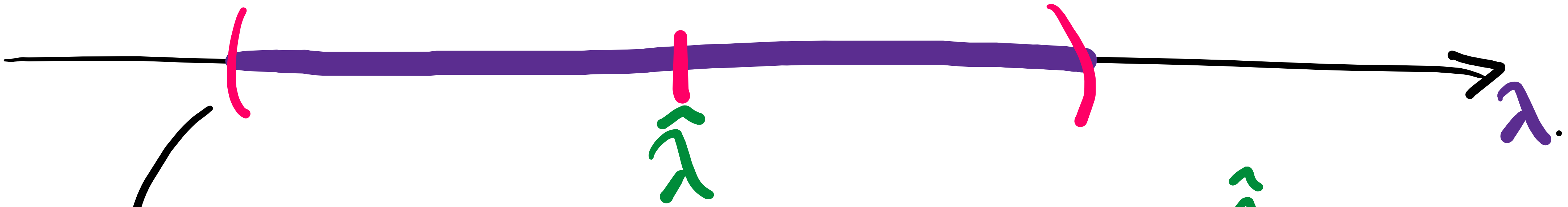
$$s \xrightarrow[n \to \infty]{} \sigma(X)$$

L'écart-type échantillonnal est un estimateur de l'écart-type

$$\hat{\lambda} \xrightarrow[n \to \infty]{} \lambda(X)$$

On utilise souvent le ^ pour les estimateurs.

Les intervalles de confiance.



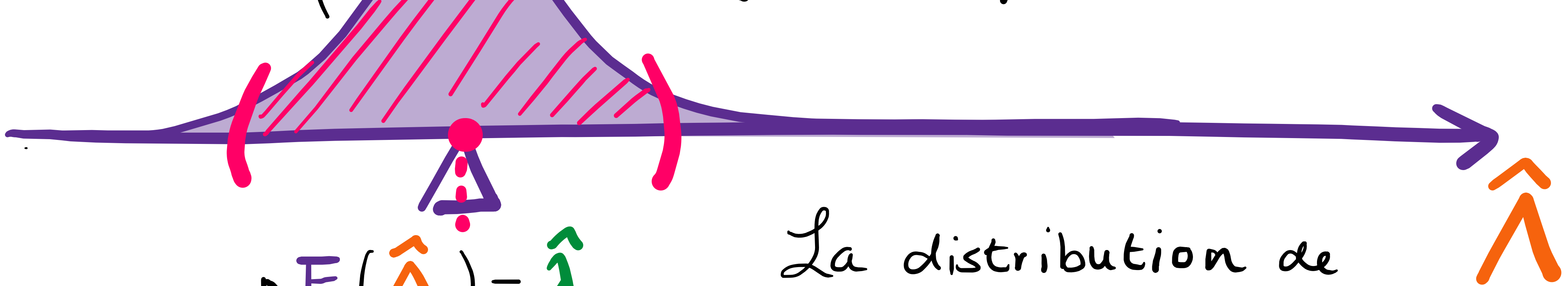
L'**estimateur** $\hat{\lambda}$ est notre prédiction pour la valeur du **paramètre** λ.

On voudrait un **intervalle** où il serait « probable » de retrouver λ.

Les intervalles de confiance.

On cherche un intervalle I t.q.

$$P(\hat{\Lambda} \in I) = 1 - \alpha$$



$$E(\hat{\Lambda}) = \hat{\lambda}$$

On assume que

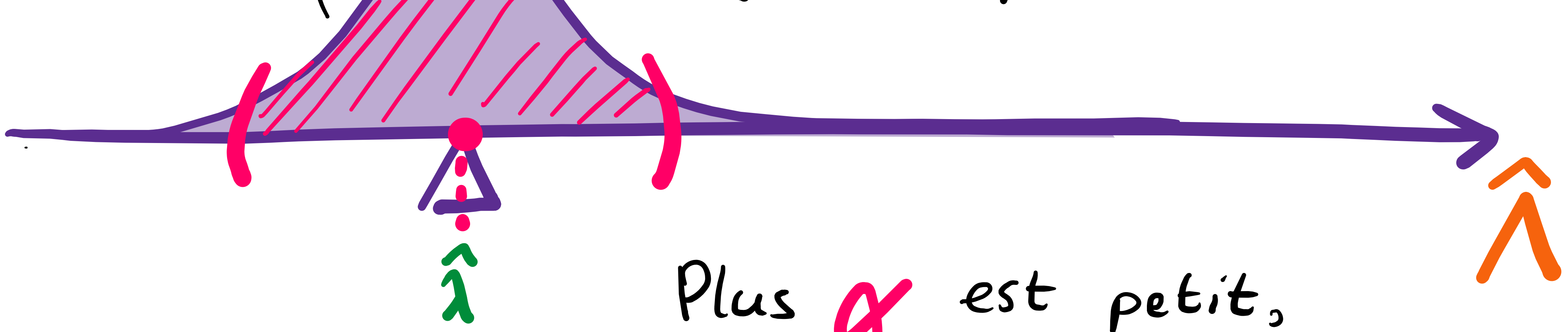
$$\lambda = \hat{\lambda}.$$

La distribution de l'estimateur-modèle $\hat{\Lambda}$ est centrée autour de $\hat{\lambda}$.

Les intervalles de confiance.

On cherche un intervalle I t.q.

$$P(\hat{\Lambda} \in I) = 1 - \alpha$$

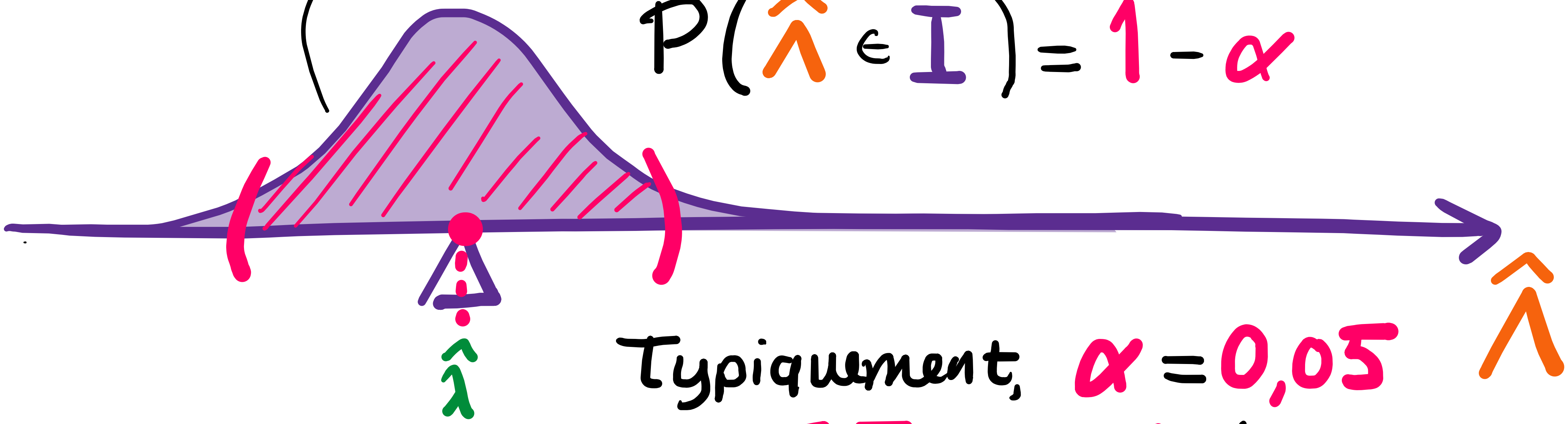


Plus α est petit,
plus on a confiance,
mais plus l'intervalle est large.

Les intervalles de confiance.

On cherche un intervalle I t.q.

$$P(\hat{\lambda} \in I) = 1 - \alpha$$



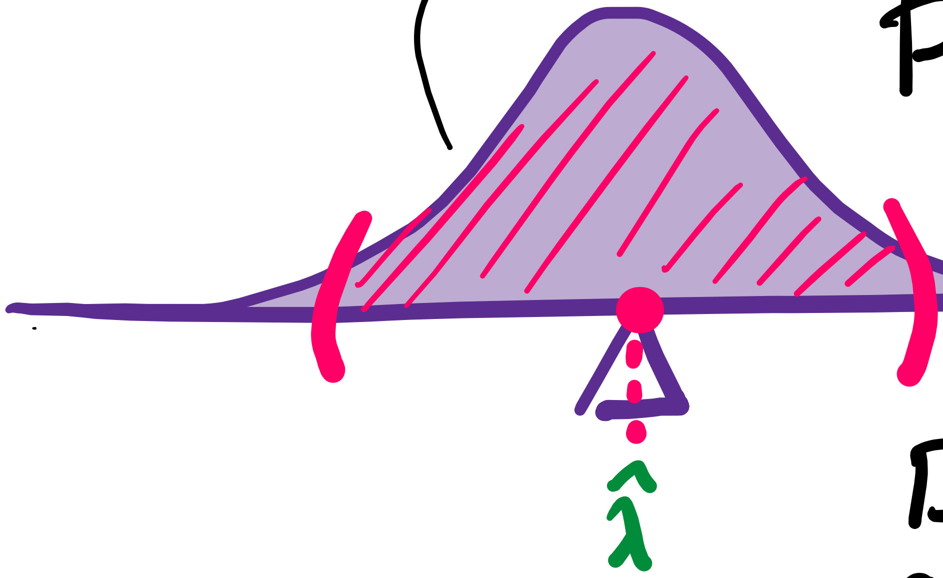
Typiquement, $\alpha = 0,05$

ou 5%, ou $1/20$.

Les intervalles de confiance.

On cherche un intervalle I t.q.

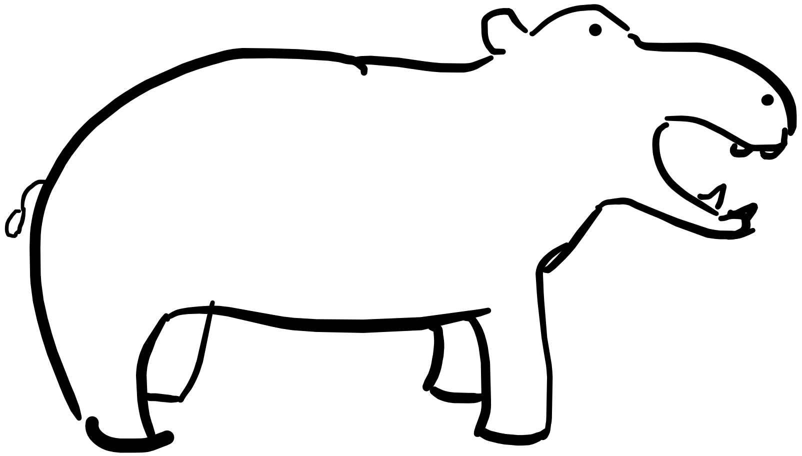
$$P(\hat{\lambda} \in I) = 1 - \alpha$$



Dans ce cas, λ se trouve dans l'intervalle **19** fois sur **20** !

L'inférence non-paramétrique
et les tests d'hypothèse.

L'hypothèse nulle.



« l'hippopo-thèse »

H_0

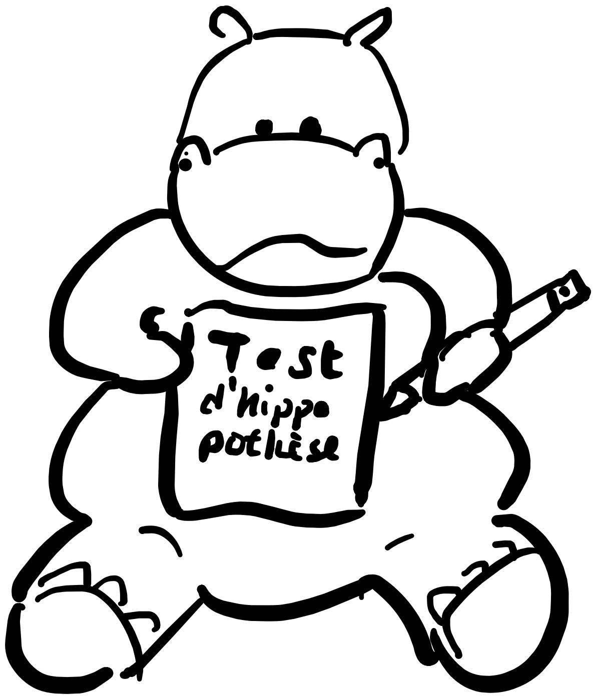
Il s'agit d'une hypothèse à tester concernant la **distribution** de **variables** dans la population.

ex : $X \perp Y$ — X et Y sont indépendantes

$X \sim \mathcal{N}$ — X est de **loi normale**.

etc.

Le **test**



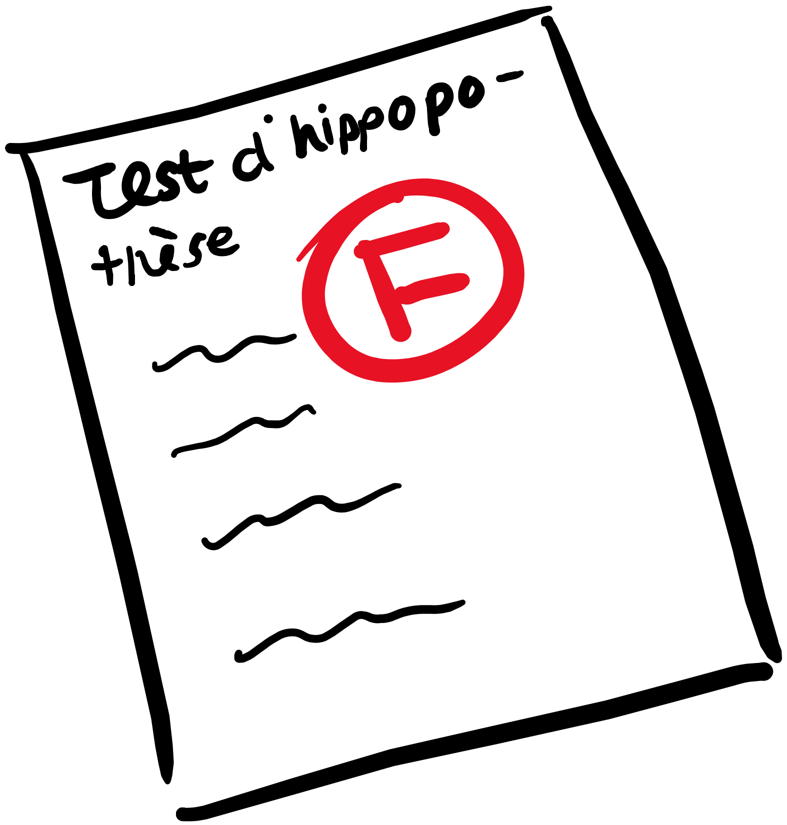
On sélectionne une **statistique de test**

On la **compare** à la **distribution prédite** pour la **statistique-modèle** en supposant que

$$\mathcal{H}_0$$

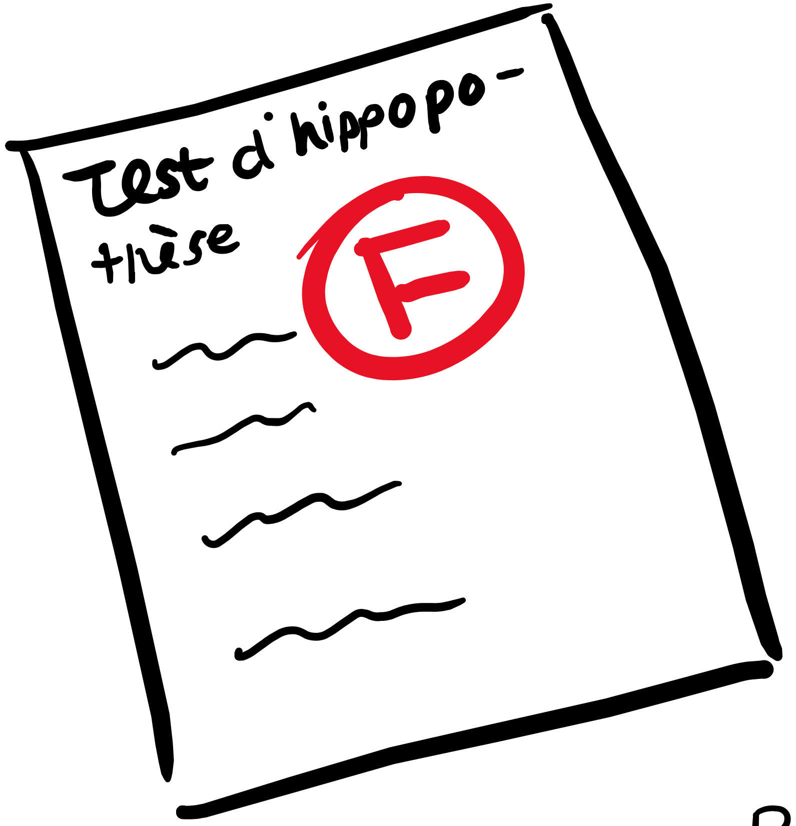
est **vraie.**

L'échec.



On **rejette** l'hypothèse si la valeur de la statistique de test est **trop improbable** pour la statistique-modèle.

L'échec.



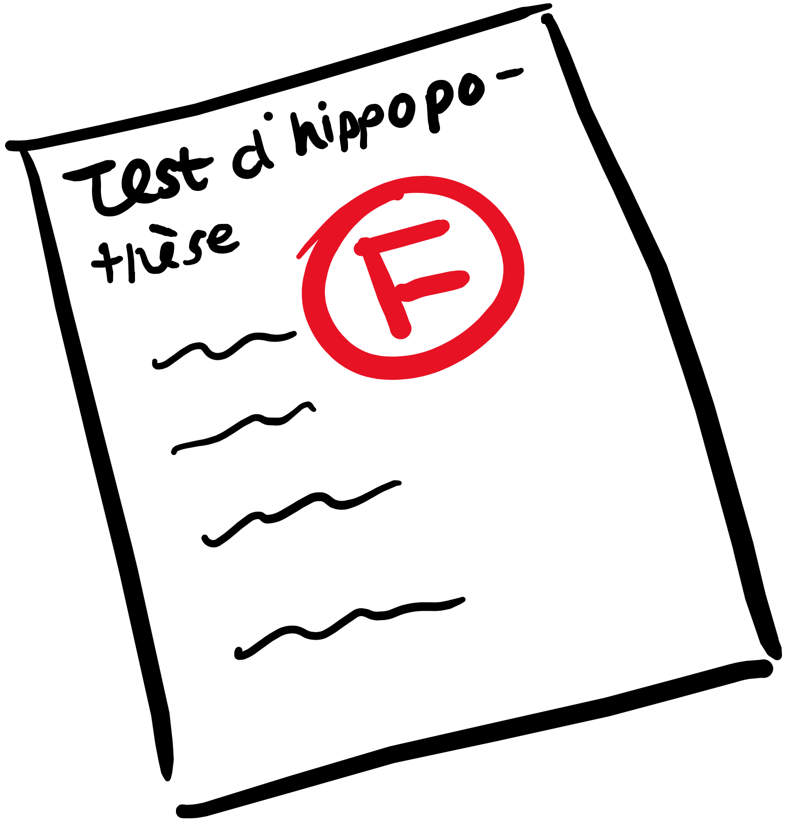
Le seuil de signification

α est la probabilité d'échouer le test si H_0 est vraie:

$$\alpha = P(F \mid H_0)$$

plus α est petit, plus un échec au test est **significatif**.

L'échec.



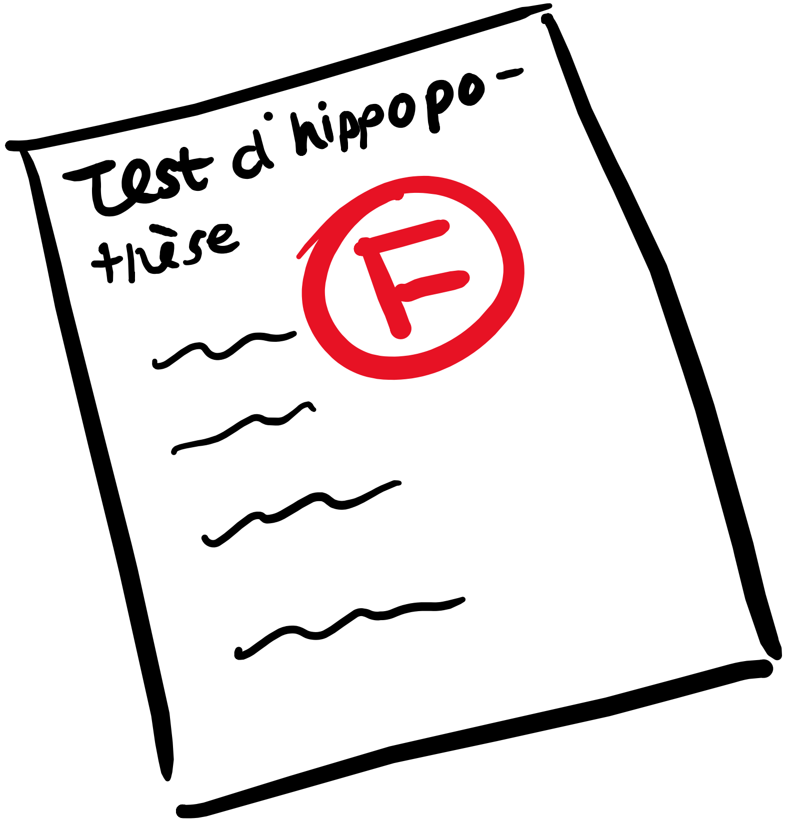
La valeur p

est la probabilité que la **statistique-modèle** soit au moins aussi improbable que la **statistique de test.**

$$p = P\left(\hat{\Lambda} \text{ «pire que» } \hat{\lambda} \mid H_0\right)$$

plus p est petit, plus on risque d'échouer.

L'échec.



Donc, le test échoue et l'hypothèse est rejetée si

$$p < \alpha.$$

Le non-échec.

Si le test n'échoue pas, on ne peut <u>**pas**</u> conclure que H_0 est vraie. On ne peut rien conclure.


Merci
pour votre attention!

Voici l'éclipse de la nuit du 15 au 16 mai 2022.


Cette photo
n'a pas rapport
elle est juste
belle.