TYPE D'ÉTUDE STATISTIQUE		
CARACTÉRISTIQUE	**OBSERVATIONNELLE (mode passif)**	**EXPÉRIMENTALE (mode actif)**
provenance des données	- historiques - suivi de processus dans des conditions normales d'opération; -le processus n'est pas manipulé (volontairement perturbé);	on fait varier le processus sous différentes conditions (variables) dans le but d'obtenir des données
quantité de données	-généralement abondantes; -on peut obtenir des observations additionnelles	fixe et limitée
qualité de données	peut présenter des difficultés : changements non documentés, données manquantes etc.	excellente
coût	généralement faible	généralement élevé
but	modélisation et exploration	détecter des changements
hypothèse sous jacente	homogénéité des données (*)	hétérogénéité causée par les perturbations induites
méthodes d'analyse	- carte données individuelles et étendues mobiles XmR; - carte Xbar&R ou Xbar&S afin de vérifier l'homogénéité des données - méthodes de Data Mining	- analyse de la variance - analyse de régression - autres méthodes - cartes XmR, Xbar&R,…

L'homogénéité des données est fondamentale lors de leur l'analyse.
Cette question est clarifiée dans l'article suivant :

Wheeler, Donald J. (2009) ***The four Questions of Data Analysis***
https://www.spcpress.com/pdf/DJW204.pdf

Nous présentons les éléments essentiels de cet article dans les pages suivantes.

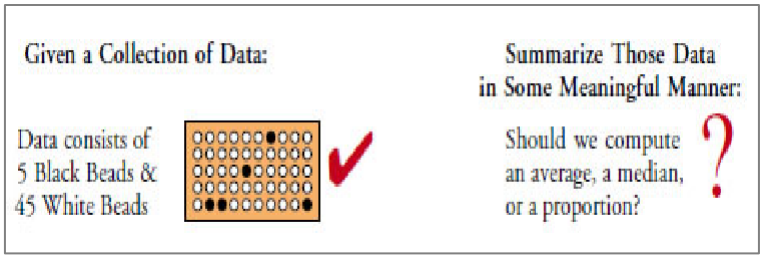
La question fondamentale de l'homogénéité des données : les données sont-elles homogènes? Comment interpréter correctement la variabilité?

L'analyse statistique des données s'appuie sur 4 thèmes :

DESCRIPTION PROBABILITÉ INFÉRENCE HOMOGÉNÉITÉ

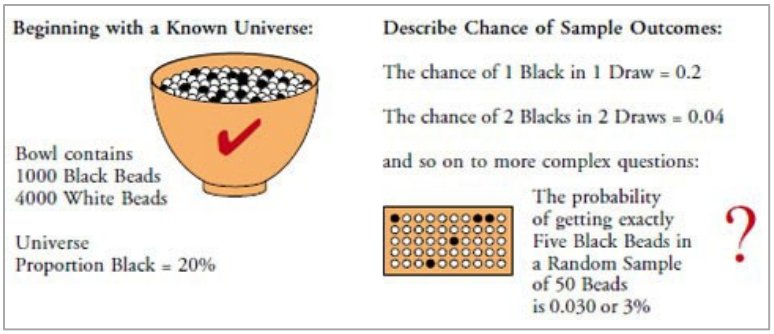
Les 3 premiers thèmes sont relativement bien connus. C'est le thème de l'homogénéité qui l'est beaucoup moins. La mise en œuvre de toute analyse repose sur le savoir-faire pour organiser et utiliser ces 4 questions afin d'obtenir des résultats compréhensibles et des réponses exactes. Nous fournirons un bref résumé de chacun des 3 premiers thèmes avant de focaliser sur celui de l'homogénéité car il représente la question préalable indispensable avant tous les autres. En effet, les autres thèmes font du sens que dans la mesure où le caractère d'homogénéité des données est satisfait.

DESCRIPTION : étant donné une collection de nombres, comment résumer d'une manière compréhensible, l'information contenue dans ces nombres?



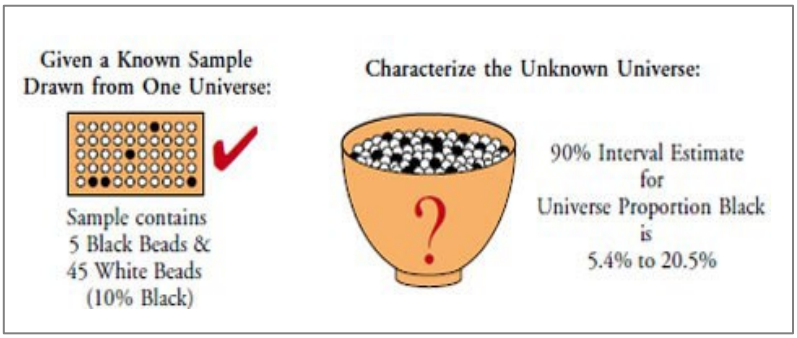
Graphique 1: La question de la description
Wheeler, Donald J. (2009) *The four Questions of Data Analysis*

PROBABILITÉ : étant donné univers de référence connu (population), que pouvons-nous dire à propos d'échantillons tirés de cet univers?



Graphique 2: La question de la probabilité - Wheeler

INFÉRENCE : étant donné un univers inconnu, étant donné un échantillon tiré de cet univers, étant donné que l'on connait tout de cet échantillon, que pouvons-nous dire à propos de cet univers?



Graphique 3: La question de l'inférence - Wheeler

l'inférence est le domaine privilégié des tests d'hypothèses, des intervalles de confiance et de la régression.

La question importante : les données sont-elles homogènes ou non ?

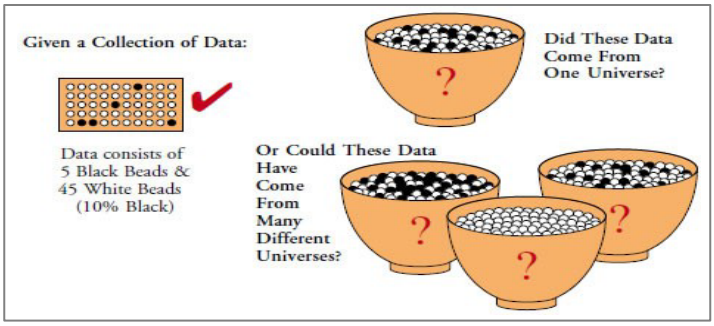
Des données homogènes c'est quoi ?

HOMOGÉNÉITÉ : étant donné une collection d'observations, est-il raisonnable de supposer qu'elles proviennent d'un univers unique ou montrent-elles l'évidence de provenir de plusieurs univers?

Afin de comprendre le rôle prépondérant de l'homogénéité, considérons les conséquences lorsque la collection d'observations ne provient pas d'un univers unique.

- **Si les données furent générées par plusieurs sources (populations), comment les *statistiques descriptives* peuvent –elles être employées pour caractériser une propriété unique mais qui, dans les faits, représentent plusieurs propriétés ?**
- **Si on ne peut pas répondre à la question de l'homogénéité alors on ne sait pas si on a affaire à un modèle de probabilité ou plusieurs ; les calculs de probabilité conduisent à des résultats variés et contradictoires.**
- **L'inférence statistique suppose que l'échantillon provient d'un seul univers. Si les données proviennent de sources différentes que représente, par exemple, un intervalle de confiance?**

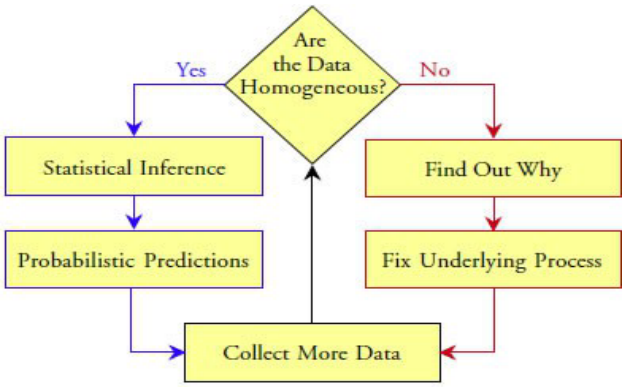
L'hypothèse d'homogénéité est implicite dans l'application de toutes les méthodes statistiques. Si elle n'est pas vérifiée, alors elle constitue un obstacle majeur lorsque l'on analyse des données. Considérons le graphique 4. Supposons par exemple que l'échantillon contient 10% de billes noires. Le résultat provient-il d'un univers unique ou de plusieurs autres possibles? Quel univers l'échantillon représente-t-il ?



Graphique 4: La question de l'homogénéité - Wheeler

Comment analyser les données?

La présence de non-homogénéité dans l'échantillon met sérieusement en péril les résultats d'inférence statistique. Le manque d'homogénéité représente un signal que des évènements en sont responsables et, aussi longtemps que l'on n'identifie pas et on n'enlève pas les causes de ce dérèglement, on est en présence d'un obstacle majeur pour analyser les données. Le graphique 5 représente les étapes du processus. Il faut vérifier l'homogénéité avant de procéder plus avant avec les analyses.



Graphique 5 : L'homogénéité est la question principale - Wheeler

La prudence nous dicte de ne pas faire l'hypothèse implicite de l'homogénéité des données.

La principale et l'unique méthode pour examiner si des données sont homogènes ou non est la *carte de contrôle* connue sous ce et intimement associée aux méthodes d'ingénierie de la qualité. Elle fut développée par Shewhart en 1924.
Ces cartes représentent le comportement du processus de collecte des données et on pourrait appelé cet outil extrêmement puissant *carte de comportement de processus*.

Le document suivant présente les éléments essentiels de cette méthode fondamentale.
Clément, B. (2011) *Introduction aux cartes de Shewhart*

https://cours.polymtl.ca/mth6301/mth8302-Cours&Plus/Clement-ControleStatistiqueProcessus.pdf