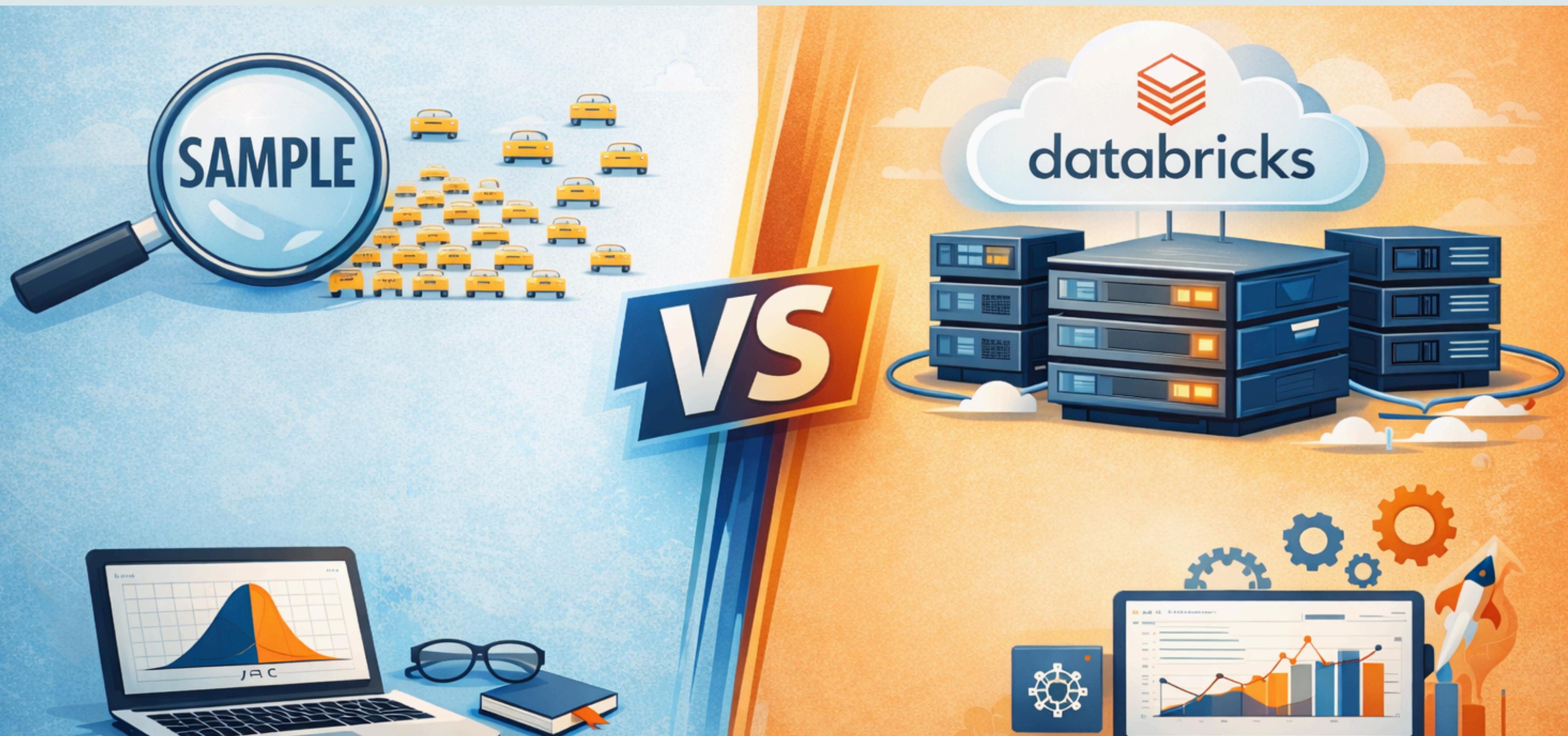


ANALYSE AVEC INFERENTIAL STATISTICS VS BIG DATA ANALYTICS SUR DATABRICKS





CONTEXTE

Dans le cadre du projet DATAKO, l'objectif est d'analyser les données des courses de taxis new-yorkais (2022–2024) en comparant statistiques inférentielles sur un échantillon 1% et Big Data sur la population complète via Spark. L'objectif est d'évaluer la fiabilité des estimations d'un échantillon face aux mesures exactes, en tenant compte de la variabilité, des outliers et de la couverture géographique.

STATISTIQUES INFÉRENTIELLES (ÉCHANTILLON 1%)

Indicateur	Moyenne (échantillon)	IC 95 %
Prix moyen d'une course (fare_amount)	17.86 \$	17.83 – 17.90
Distance moyenne (trip_distance)	5.67 miles	4.78 – 6.57
Durée moyenne des courses (minutes)	26.081	9.10 – 43.06
Proportion de courses avec tip > 0	0.744	0.743 – 0.745

RÉPARTITION PAR HEURE / JOUR

Heures de pointe : 16h–18h

Jour le plus fréquent : Thursday

Répartition hebdomadaire : semaine 50 la plus chargée

COMPARAISON PAR ZONES GÉOGRAPHIQUES (PU/DO)

PULocationID	mean_fare	IC 95 %
1	87.75	80.73 – 94.77
265	69.38	65.81 – 72.94
132	55.65	55.38 – 55.91

ANALYSE DES OUTLIERS

Moyenne avec outliers : 17.86

Moyenne sans outliers : 13.52

Les courses très chères ou longues influencent fortement la moyenne.

RATIO TIP / FARE PAR TYPE DE PAIEMENT

Type paiement	Ratio moyen tip/fare	IC 95 %
Card	0.266	0.248 - 0.283
Cash	~0	~0

STATISTIQUES BIG DATA (POPULATION COMPLÈTE)

Indicateur	Valeur population
Prix moyen d'une course	16.39 \$
Distance moyenne	5.02 miles
Durée moyenne des courses	29.46 min
Proportion de courses avec tip > 0	0.744
Outliers (fare_amount)	12 490 247 courses
Moyenne sans outliers	13.52 \$

DISTRIBUTION POPULATIONNELLE

Par heure : pic 17-18h (≈ 8.46 M courses)

Par jour : jeudi le plus chargé (≈ 18.56 M courses)

Par semaine ISO : semaine 29 la plus chargée (≈ 2.25 M courses)

STATISTIQUES PAR ZONE (TOP 3 PULOCATIONID)

PULocationID	mean_fare	std_fare	n
1	84.89	50.78	21 173
204	75.28	60.88	60
265	67.39	70.52	139 634

RATIO TIP/FARE PAR TYPE DE PAIEMENT

payment_type	mean_ratio	n
1 (Card)	0.274	90 387 135
2(Cashe)	~0	19 687 622

COMPARAISON STATISTIQUES INFÉRENTIELLES VS BIG DATA

Indicateur	Échantillon 1 % (Inférentiel)	Population 100 % (Big Data)	Commentaire
Prix moyen	17.86 ± 0.03	16.39	Inférence légèrement surestimée
Distance moyenne	5.67 ± 0.89	5.02 miles	Échantillon légèrement plus long
Durée moyenne	26.08 ± 17.96	29.46	IC large → variabilité importante sur échantillon
Proportion tip>0	0.744 ± 0.00078	0.744	Très proche → échantillon représentatif
Fares par zone	Certaines zones sous-représentées	Exact	Big Data capture toutes les zones
Outliers	impact sur la moyenne (17.86 → 13.52 sans outliers)	12 490 247 courses	Big Data fiable pour valeurs extrêmes
Ratio tip/fare (Card)	0.266 ± 0.018	0.274	Inférence proche, échantillon suffisamment grand

CONCLUSION

- 
- L'échantillon 1 % permet d'avoir des estimations proches des valeurs exactes pour la majorité des indicateurs.
 - Les outliers et zones sous-représentées peuvent influencer la moyenne.
 - L'approche Big Data est indispensable pour les analyses précises et la couverture complète des zones et des valeurs extrêmes