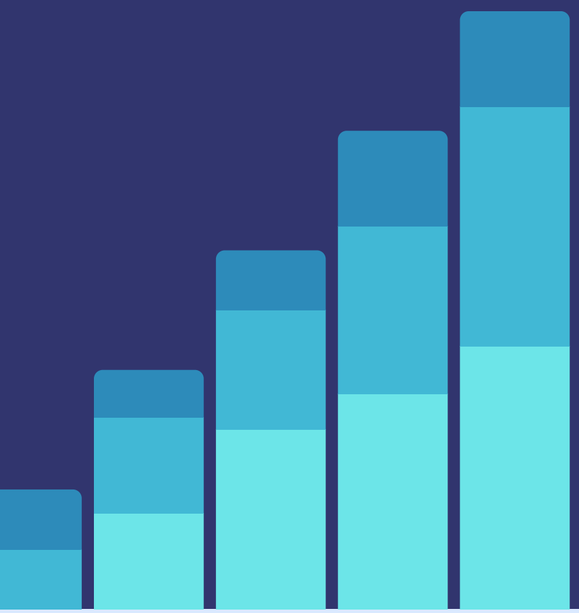
PUIS-JE UTILISER DES DONNÉES SYNTHÉTIQUES POUR MES ÉTUDES STATISTIQUES ?

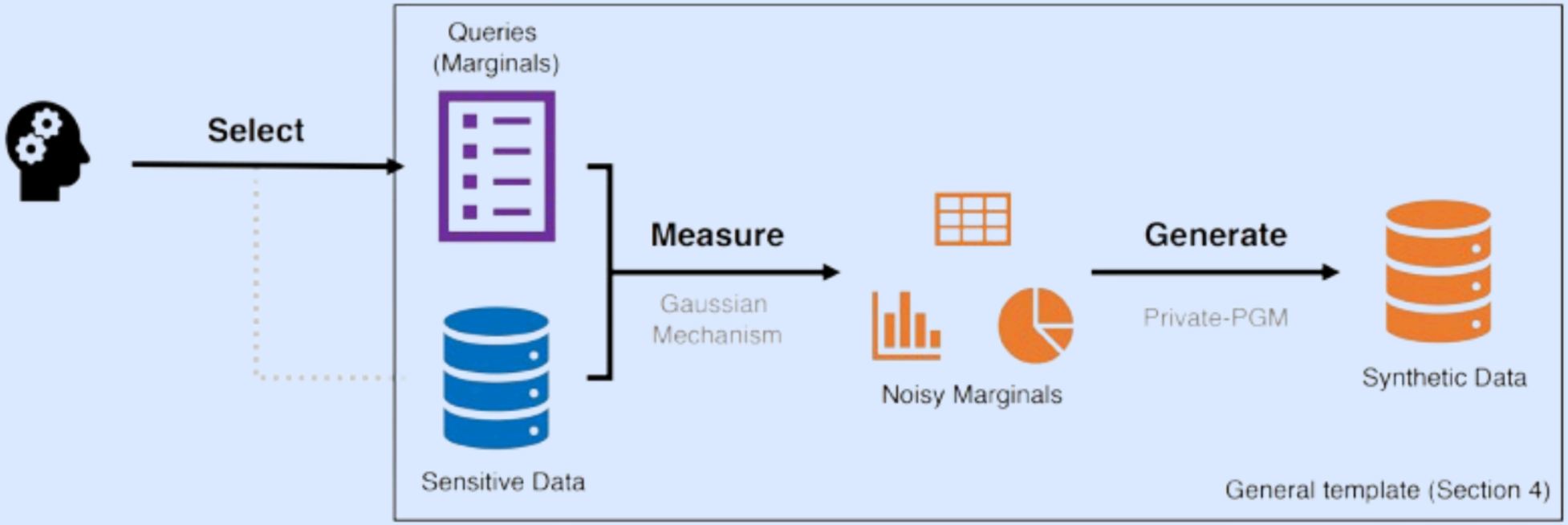
DELMARE BASTIEN & COTTET CORALIE



ALGORITHME À CONFIDENTIALITÉ DIFFÉRENTIELLE :

LE MST (MAX SPANNING TREE)

3 GRANDES ETAPES :



QU'EST CE QU'UNE MARGINALE?

Sexe	Travaille	Heures travaillées
Homme	Oui	35
Femme	Oui	42
Femme	Oui	40
Homme	Non	20
Homme	Oui	33

Marginales à deux dimensions :

- Homme qui travaille = 2
- Homme qui ne travaille pas = 1
- Femme qui travaille = 2
- Femme qui ne travaille pas = 0

LE MÉCANISME GAUSSIEN

Soit $f : \mathcal{D} \to \mathbb{R}^p$ une fonction vectorielle de l'ensemble de données D. Le mécanisme Gaussien ajoute du bruit Gaussien indépendant et identiquement distribué (i.i.d.)

$$\mathcal{M}(D) = f(D) + \mathcal{N}(0, \sigma^2 \mathbf{I})$$

LE MÉCANISME EXPONENTIEL

Soit $q : \mathcal{D} \times \mathcal{R} \to \mathbb{R}$ une fonction de score de qualité et epsilon un paramètre. Le mécanisme exponentiel génère une sortie candidate r∈R selon la distribution suivante :

$$\Pr[\mathcal{M}(D) = r] \propto \exp\left(\epsilon \cdot q(D, r)\right)$$

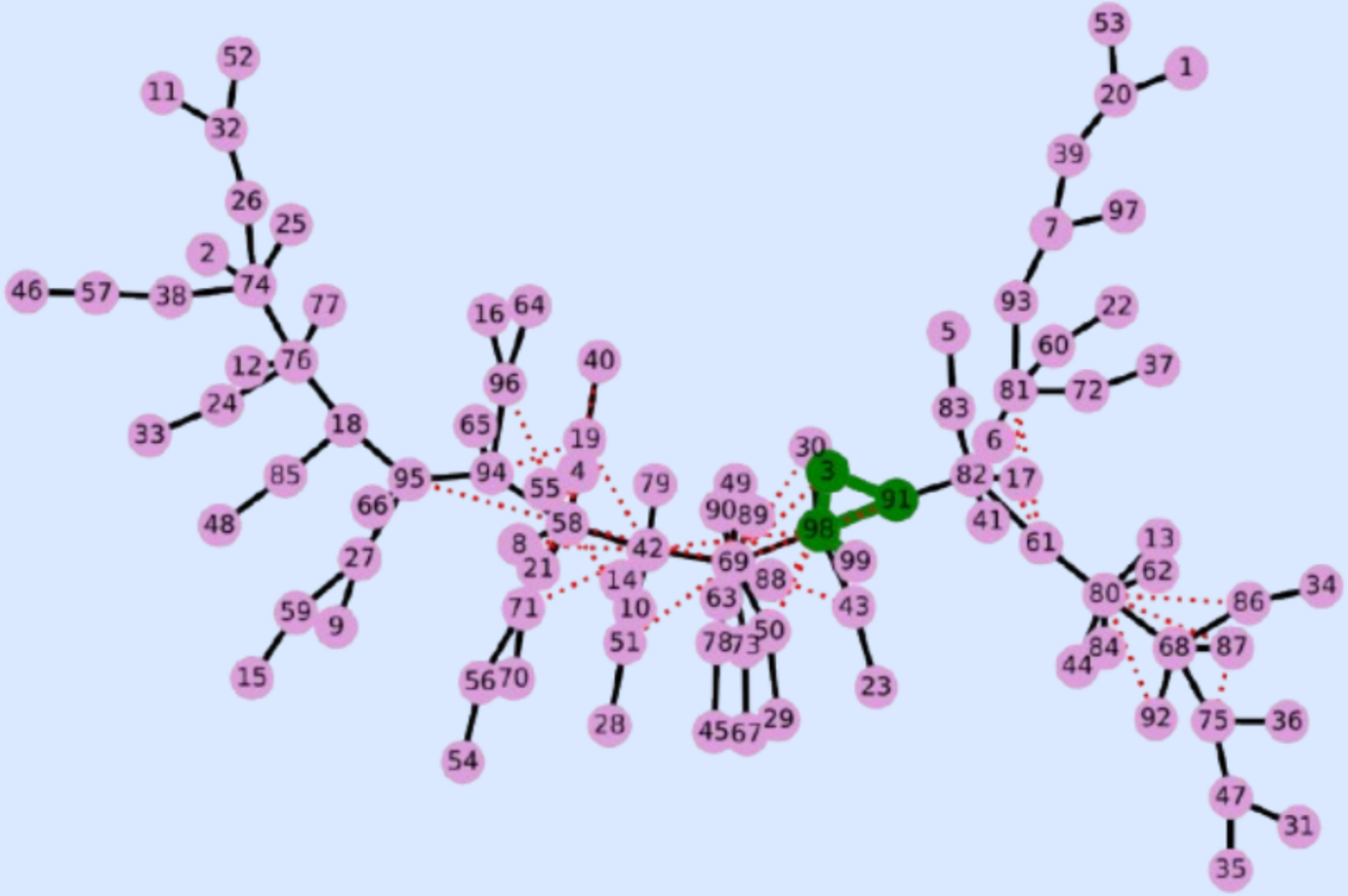
1ER ÉTAPE : SELECTION DES MARGINALES

Soient i et j deux variables de notre base de données.

Soit p le modèle indépendant obtenu à l'aide des marginales bruitées de dimension 1.

L'algorithme initialise un graphe, et ajoute à ce graphe au fur et à mesure les arêtes qui ont la valeur donné par le mecanisme exponentielle avec le score de qualité le plus élevé : $q_{ij}(D) = \|\mu_{ij}(D) - \mu_{ij}(p)\|_1$

1ER ÉTAPE : SELECTION DES MARGINALES



(Similaire à l’algorithme de Kruskal’s)

2ème étape : Calcule des marginales

Mesure avec un mécanisme gaussien :

Normalisation des poids $w_C \leftarrow w_C / \sqrt{\sum_C w_C^2}$

Calcule des marginales bruitées $\tilde{\mu} = w_C M_C(D) + \mathcal{N}(0, \sigma^2 I)$

Collecte du 4-uplet $(w_C I,\ \tilde{\mu},\ \sigma,\ C)$

3ème étape : Generation des données synthétiques

Private-PGM

Private-PGM est un outil de post-traitement polyvalent permettant d'inférer une distribution de données à partir de mesures bruitées.

Il permet de résoudre un problème d'optimisation pour trouver une distribution de données qui produirait des mesures proches de celles observées

PRIVATE-PGM

Un peu de formalisme

Supposons que les mesures soient de la forme : $y_C = Q_C M_C(D) + \xi$

Private-PGM infère une distribution de données P qui explique le mieux les mesures en résolvant le problème d'optimisation suivant :

$$\underset{P}{\operatorname{argmin}} \sum_{C \in \mathcal{C}} \|Q_C M_C(P) - y_C\|_2^2$$

PRIVATE-PGM

Avantages

1 - Capable de s'adapter à des domaines de très haute dimension.

2 -produit des réponses aux requêtes qui sont cohérentes entre elles

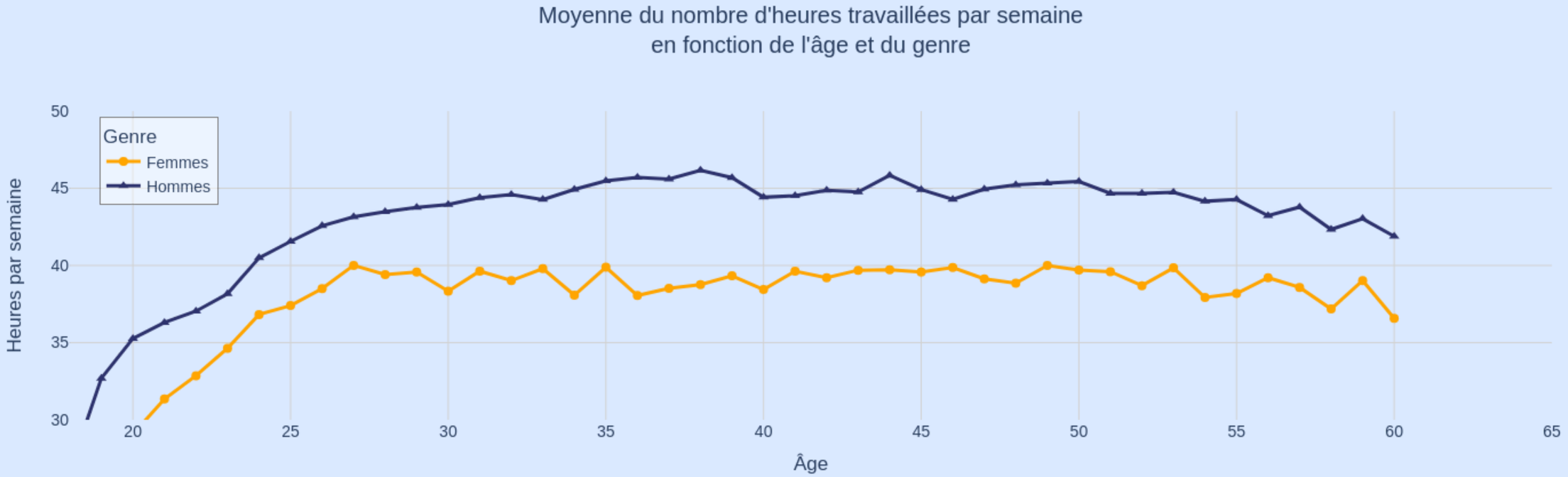
3 - Permet d’estimer des marginales non mesurées

4 - Génère des données synthétiques à partir des marginales biaisées

Problème :

Quel est le lien entre sexe et nombre d'heures travaillées ?

SUR LES DONNÉES NON BRUITÉES



Moyenne du nombre d’heures travaillées (hommes) : 40.11

Moyenne du nombre d’heures travaillées (femmes) : 37.30

P-valeur du test de Kolgomorov-smirnov : 4.83e-16

P-valeur du test de Student : 1.40e-06

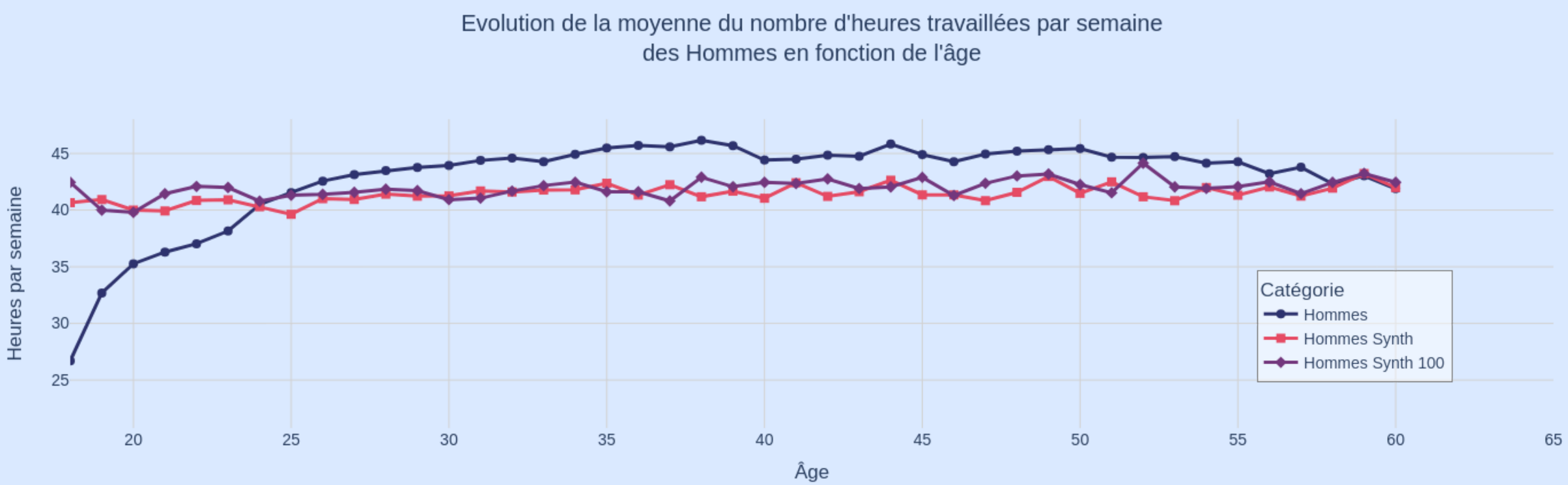
ORIGINALES VS SYNTHÉTIQUES (HOMMES)

Mesures	Originales	Synthétiques (epsilon = 100)	Synthétiques (epsilon = 1)
Moyennes	42.41	41.44	41.94
Ecarts-types	4.94	0.75	0.83
Médiane	44.28	41.35	42.02
1er quartile	42.51	41.00	41.45
3ème quartile	44.91	41.83	42.45

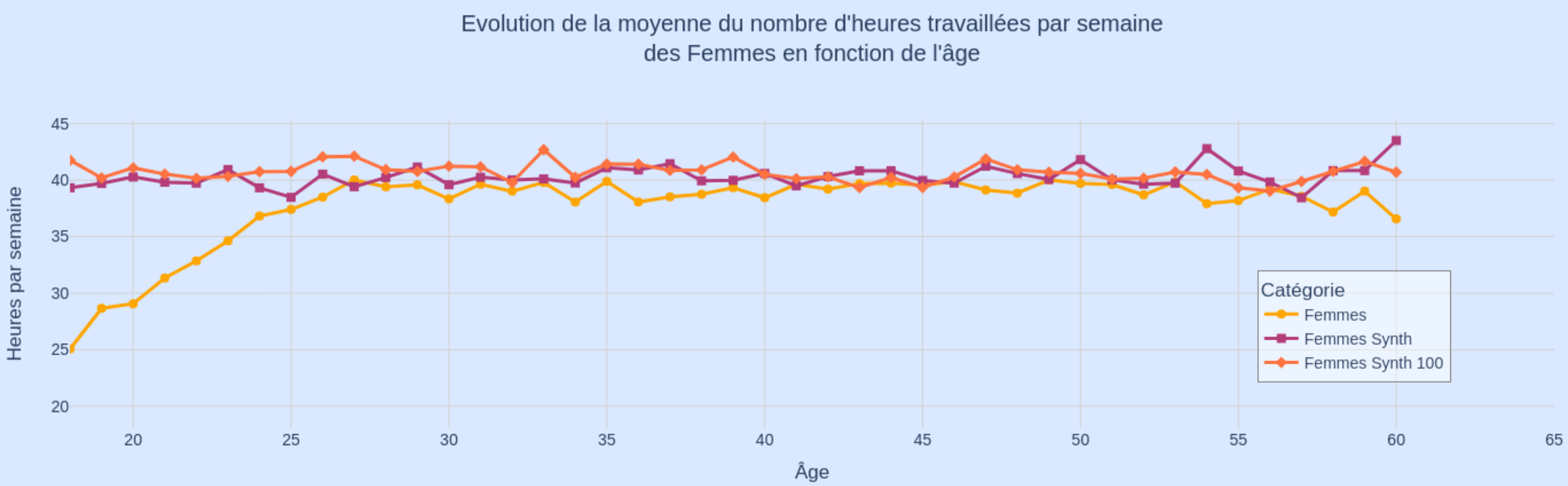
Mesures	Originales	Synthétiques (epsilon = 100)	Synthétiques (epsilon = 1)
Moyennes	37.30	40.29	40.71
Ecarts-types	4.26	0.97	0.80
Médiane	38.81	40.08	40.71
1er quartile	37.79	39.73	40.23
3ème quartile	39.61	40.83	41.11

ORIGINALES VS SYNTHÉTIQUES (FEMMES)

DONNÉES CLAIRES
VS
DONNÉES BRUITÉES



DONNÉES CLAIRES
VS
DONNÉES BRUITÉES



DES RÉSULTATS SIMILAIRES

P-valeur du test de Kolmogorov-Smirnov :

1 - Pour les données claires : 4.84e-16

2 - Pour les données modérément bruités : 2.42e-09

3 - Pour les données fortement bruités : 2.31e-05

P-valeur du test de Student :

1 - Pour les données claires : 1.41e-6

2 - Pour les données modérément bruités : 2.44e-10

3 - Pour les données fortement bruités : 2.64e-06

COMMENT LA RELATION ENTRE LE REVENU ET DES VARIABLES TELLES QUE L'ÂGE, LE NIVEAU D'ÉDUCATION, LE NOMBRE D'HEURES TRAVAILLÉES PAR SEMAINE ET LE SEXE INFLUENCE-T-ELLE LA PROBABILITÉ D'OBTENIR UN REVENU SUPÉRIEUR À 50K ?

income ~ age + education_num + hours_per_week + sex

Non-private results :

Test du chi-deux validé,
significatif au niveau de 5%

Logit Regression Results

Dep. Variable:	income	No. Observations:	32561
Model:	Logit	Df Residuals:	32556
Method:	MLE	Df Model:	4
Date:	Thu, 03 Oct 2024	Pseudo R-squ.:	0.2234
Time:	13:11:29	Log-Likelihood:	-13959.
converged:	True	LL-Null:	-17974.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-9.1334	0.116	-78.934	0.000	-9.360	-8.907
sex[T. Male]	1.1612	0.038	30.804	0.000	1.087	1.235
age	0.0456	0.001	38.464	0.000	0.043	0.048
education_num	0.3551	0.007	53.666	0.000	0.342	0.368
...						
age	0.0214	0.001	22.241	0.000	0.019	0.023
education_num	-0.0054	0.005	-1.035	0.301	-0.016	0.005
hours_per_week	0.0044	0.001	4.031	0.000	0.002	0.006

COMMENT LA RELATION ENTRE LE REVENU ET DES VARIABLES TELLES QUE L'ÂGE , LE NIVEAU D'ÉDUCATION, LE NOMBRE D'HEURES TRAVAILLÉES PAR SEMAINE ET LE SEXE INFLUENCE-T-ELLE LA PROBABILITÉ D'OBTENIR UN REVENU SUPÉRIEUR À 50K ?

Epsilon = 100 :

Logit Regression Results

Dep. Variable:	income	No. Observations:	32561
Model:	Logit	Df Residuals:	32556
Method:	MLE	Df Model:	4
Date:	Thu, 03 Oct 2024	Pseudo R-squ.:	0.05147
Time:	13:15:59	Log-Likelihood:	-16995.
converged:	True	LL-Null:	-17918.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.0147	0.081	-37.209	0.000	-3.174	-2.856
sex[T.male]	1.0823	0.033	32.326	0.000	1.017	1.148
age	0.0200	0.001	21.834	0.000	0.018	0.022
education_num	0.0154	0.005	2.925	0.003	0.005	0.026
...						
age	0.0214	0.001	22.241	0.000	0.019	0.023
education_num	-0.0054	0.005	-1.035	0.301	-0.016	0.005
hours_per_week	0.0044	0.001	4.031	0.000	0.002	0.006

Test du chi-deux validé

COMMENT LA RELATION ENTRE LE REVENU ET DES VARIABLES TELLES QUE L'ÂGE , LE NIVEAU D'ÉDUCATION, LE NOMBRE D'HEURES TRAVAILLÉES PAR SEMAINE ET LE SEXE INFLUENCE-T-ELLE LA PROBABILITÉ D'OBTENIR UN REVENU SUPÉRIEUR À 50K ?

Epsilon = 1 :

Logit Regression Results

Dep. Variable:	income	No. Observations:	32561
Model:	Logit	Df Residuals:	32556
Method:	MLE	Df Model:	4
Date:	Thu, 03 Oct 2024	Pseudo R-squ.:	0.05147
Time:	13:13:54	Log-Likelihood:	-16995.
converged:	True	LL-Null:	-17918.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.0147	0.081	-37.209	0.000	-3.174	-2.856
sex[T.male]	1.0823	0.033	32.326	0.000	1.017	1.148
age	0.0200	0.001	21.834	0.000	0.018	0.022
education_num	0.0154	0.005	2.925	0.003	0.005	0.026
hours_per_week	0.0025	0.001	2.374	0.018	0.000	0.005

Test du chi-deux validé

BIBLIOGRAPHIE

WINNING THE NIST CONTEST: A SCALABLE AND GENERAL APPROACH TO DIFFERENTIALLY PRIVATE SYNTHETIC DATA : (AUG 2021) RYAN MCKENNA, GEROME MIKLAU, AND DANIEL SHELDON

Marginal-based Methods for Differentially Private Synthetic Data
Google TechTalks

W. Qardaji, W. Yang, and N. Li. Priview: practical differentially private release of marginal contingency tables. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data, pages 1435–1446. ACM, 2014.