Gestion et analyse d’une base de données

Formation DMG

04/07/2024

Arnaud Cugerone - arnaud.cugerone@u-bordeaux.fr
Thomas Ferté - thomas.ferte@u-bordeaux.fr
Paul Vanderkam - paul.vanderkam@u-bordeaux.fr
Linda Wittkop - linda.wittkop@u-bordeaux.fr

Récupérer la base de données

Moodle séminaire aide thèse :
https://moodle.u-bordeaux.fr/course/view.php?id=6149

Analyser ses données > Base de données petits poids de naissance

Exemple fil rouge

Description

Les données nous viennent de Hosmer et Lemeshow (2000). Appelées les données de faible poids de naissance (lbw), la variable de réponse est une variable binaire, low_birth_weight, qui indique si le poids de naissance d'un bébé est inférieur à 2500g ou supérieur.

Source

Hosmer, D and S. Lemeshow (2000), Applied Logistic Regression, Wiley

Hilbe, Joseph M (2007, 2011), Negative Binomial Regression, Cambridge University Press Hilbe, Joseph M (2009), Logistic Regression Models, Chapman & Hall/CRC

Objectifs

Objectif principal

Évaluer l’association entre la consommation de tabac et le petit poids de naissance.

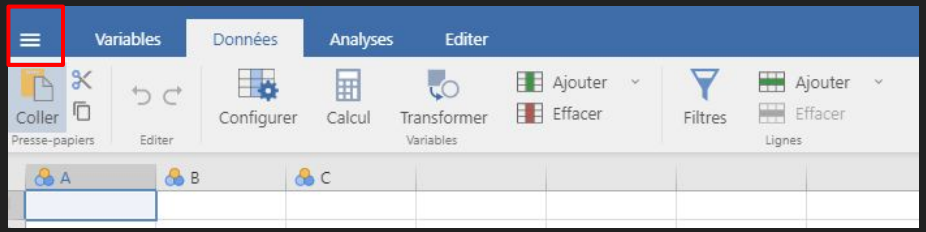
Objectifs secondaires

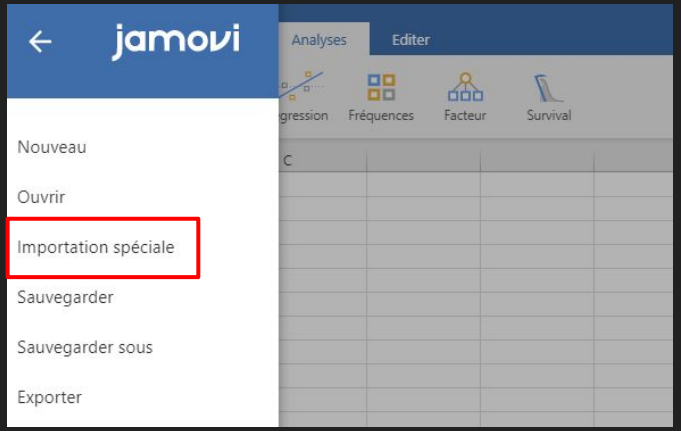
- Évaluer l’association entre l’origine ethnique, l’âge de la mère, le poids de la mère, le nombre de fausses couches antérieures, l’hypertension artérielle et l’irritabilité utérine avec le petit poids de naissance en qualitatif.
- Idem avec le petit poids de naissance en quantitatif

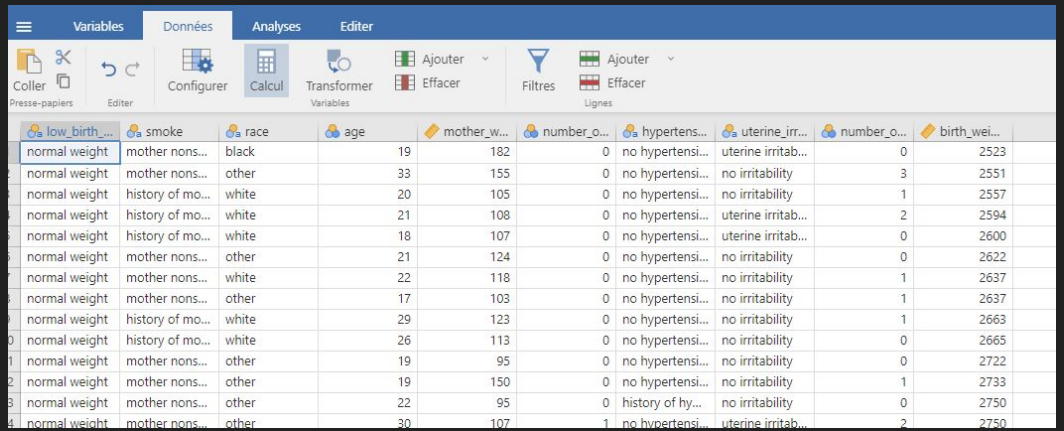
Les données

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	low_birth_weight,"smoke","race","age","mother_weight","number_of_false_or_premature_labors","hypertension","uterine_irritability","number_of_physician_visits_in_1st_trimester","birth_weight_in_grams"																
2	normal weight,"mother nonsmoker","black",19,182,0,"no hypertension","uterine irritability",0,2523																
3	normal weight,"mother nonsmoker","other",33,155,0,"no hypertension","no irritability",3,2551																
4	normal weight,"history of mother smoking","white",20,105,0,"no hypertension","no irritability",1,2557																
5	normal weight,"history of mother smoking","white",21,108,0,"no hypertension","uterine irritability",2,2594																
6	normal weight,"history of mother smoking","white",18,107,0,"no hypertension","uterine irritability",0,2600																
7	normal weight,"mother nonsmoker","other",21,124,0,"no hypertension","no irritability",0,2622																
8	normal weight,"mother nonsmoker","white",22,118,0,"no hypertension","no irritability",1,2637																
9	normal weight,"mother nonsmoker","other",17,103,0,"no hypertension","no irritability",1,2637																
10	normal weight,"history of mother smoking","white",29,123,0,"no hypertension","no irritability",1,2663																
11	normal weight,"history of mother smoking","white",26,113,0,"no hypertension","no irritability",0,2665																
12	normal weight,"mother nonsmoker","other",19,95,0,"no hypertension","no irritability",0,2722																
13	normal weight,"mother nonsmoker","other",19,150,0,"no hypertension","no irritability",1,2733																
14	normal weight,"mother nonsmoker","other",22,95,0,"history of hypertension","no irritability",0,2750																

Importer ses données



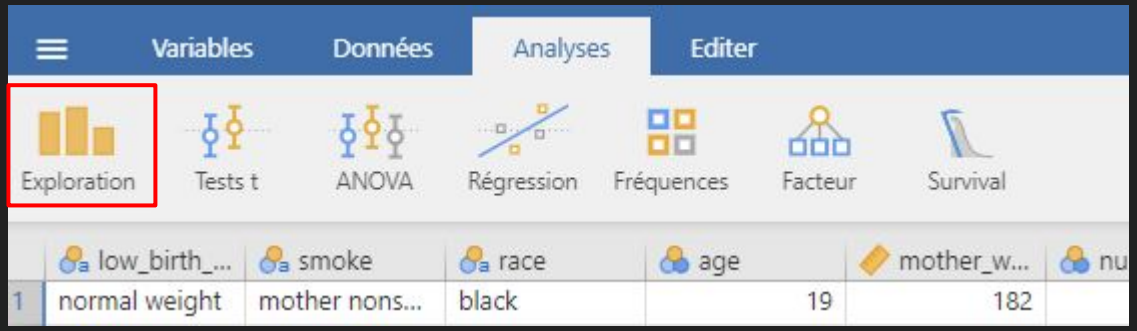




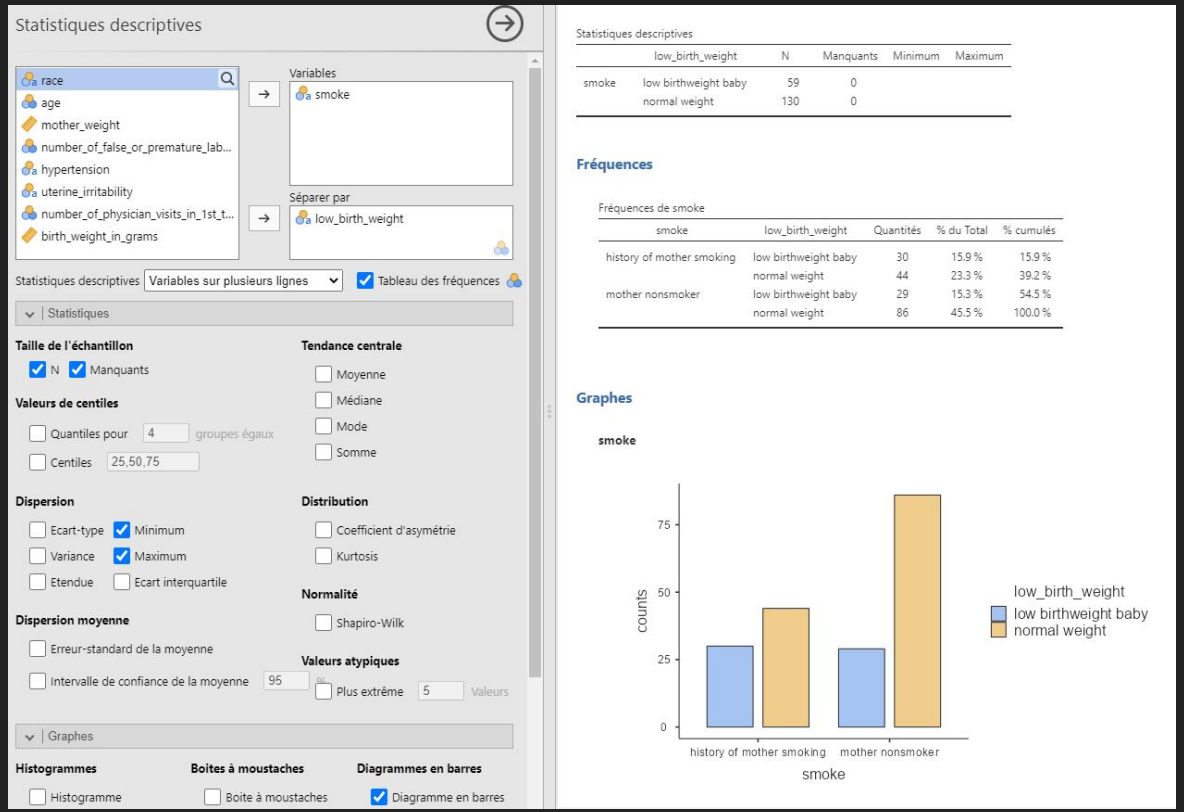
	low_birth_...	smoke	race	age	mother_w...	number_o...	hypertens...	uterine_irr...	number_o...	birth_wei...
1	normal weight	mother nons...	black	19	182	0	no hypertensi...	uterine irritab...	0	2523
2	normal weight	mother nons...	other	33	155	0	no hypertensi...	no irritability	3	2551
3	normal weight	history of mo...	white	20	105	0	no hypertensi...	no irritability	1	2557
4	normal weight	history of mo...	white	21	108	0	no hypertensi...	uterine irritab...	2	2594
5	normal weight	history of mo...	white	18	107	0	no hypertensi...	uterine irritab...	0	2600
6	normal weight	mother nons...	other	21	124	0	no hypertensi...	no irritability	0	2622
7	normal weight	mother nons...	white	22	118	0	no hypertensi...	no irritability	1	2637
8	normal weight	mother nons...	other	17	103	0	no hypertensi...	no irritability	1	2637
9	normal weight	history of mo...	white	29	123	0	no hypertensi...	no irritability	1	2663
10	normal weight	history of mo...	white	26	113	0	no hypertensi...	no irritability	0	2665
11	normal weight	mother nons...	other	19	95	0	no hypertensi...	no irritability	0	2722
12	normal weight	mother nons...	other	19	150	0	no hypertensi...	no irritability	1	2733
13	normal weight	mother nons...	other	22	95	0	history of hy...	no irritability	0	2750
14	normal weight	mother nons...	other	30	107	1	no hypertensi...	uterine irritab...	2	2750

Description des données

Analyses > Exploration > Statistiques descriptives

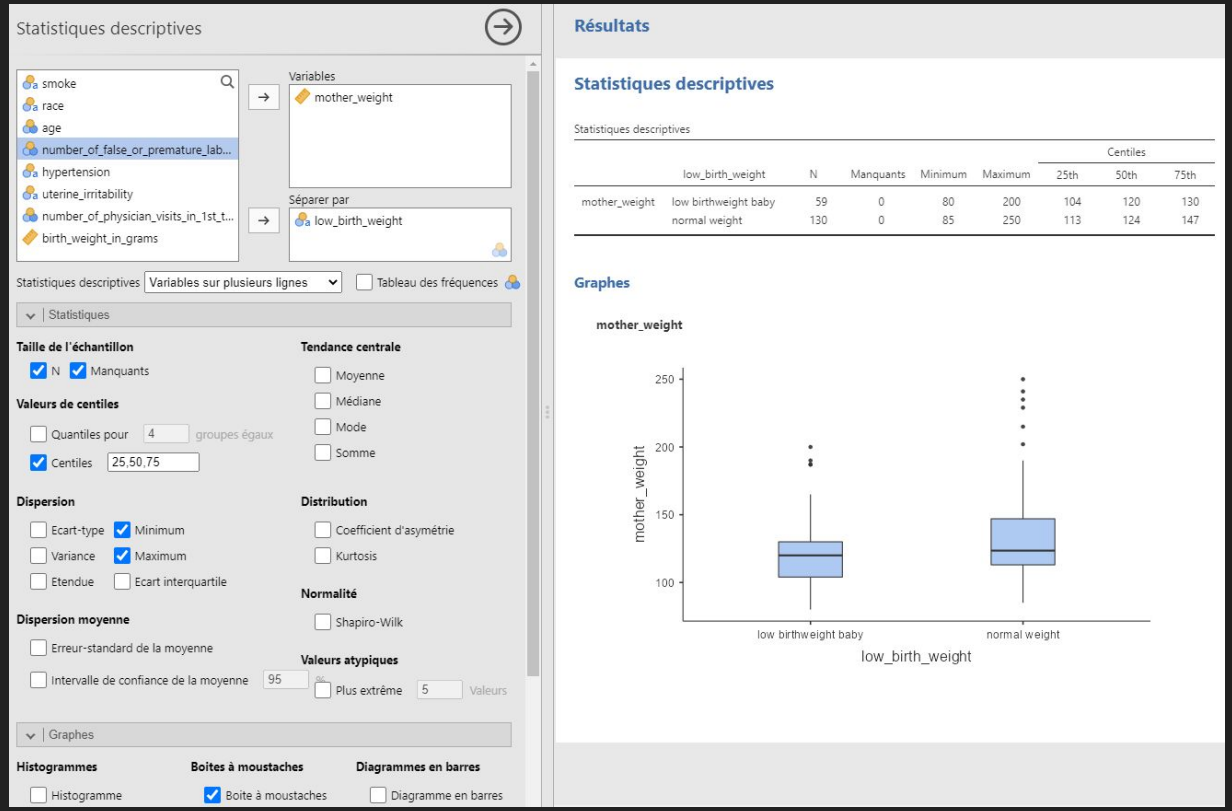


Exemple : Tabac en fonction du petit poids de naissance



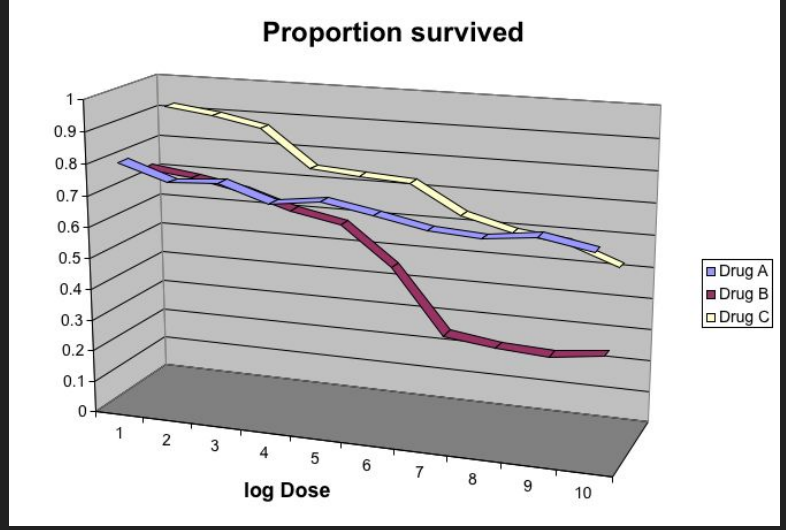
Exercice : décrivez le poids de la mère en fonction du petit poids de naissance. Réalisez le graphique adapté

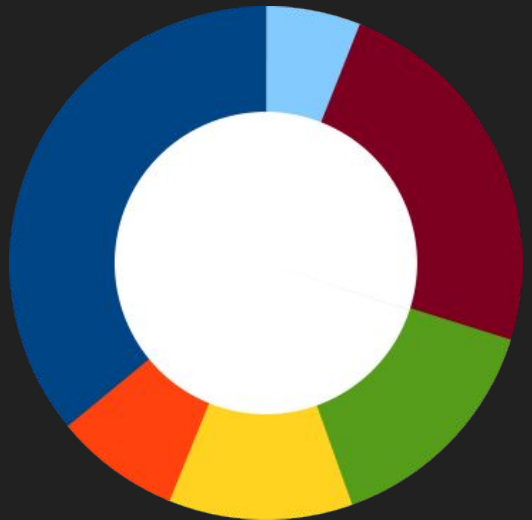
Exercice : décrivez le poids de la mère en fonction du petit poids de naissance. Réalisez le graphique adapté

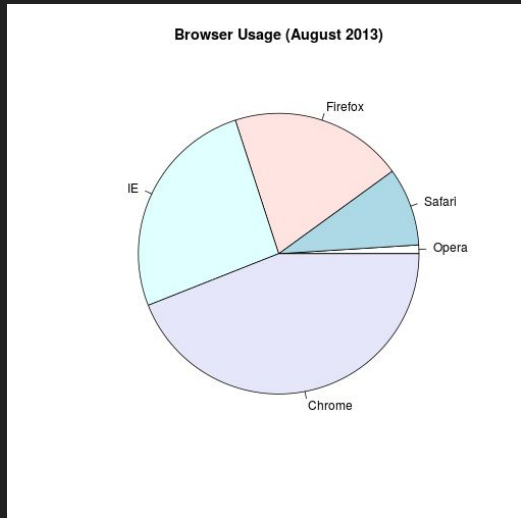


Ne pas faire

- Camemberts/Donuts
- Plots en 3D
- Un graphique et un tableau pour représenter la même chose
- Un graphique lorsque le tableau est plus parlant







Bonnes idées

- Variable quantitative continue : histogramme
- Variable quantitative discrète : histogramme ou barplot
- Variable qualitative : souvent tableau > graphique
- Variable quantitative ~ Variable quantitative : nuage de points
- Variable quantitative ~ Variable qualitative : boite à moustache
- Variable qualitative ~ Variable qualitative : souvent tableau > graphique

Test simples

Choix du test

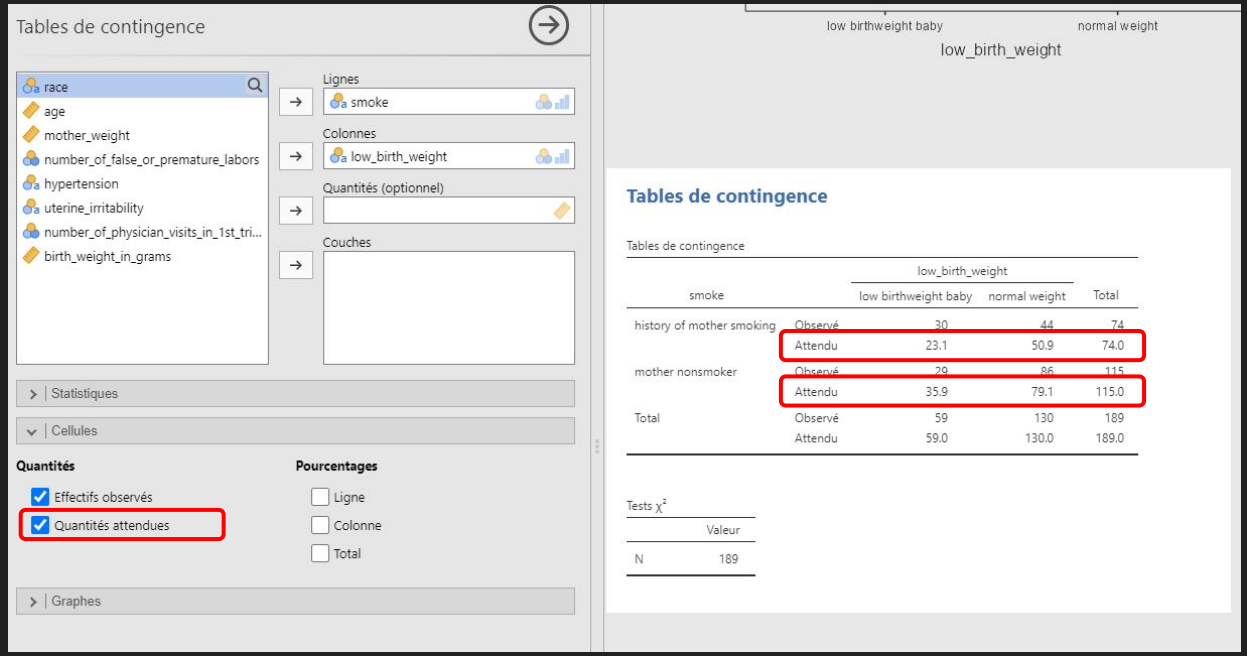
	Variable qualitative	**Variable quantitative**
Variable qualitative	Chi-2 / Fisher	Student / U de Mann Whitney
Variable quantitative	Student / U de Mann Whitney	Corrélation Spearman / Pearson
Survie	Log-rank	

Évaluer l’association entre le statut tabagique et le petit poids de naissance

Analyses > Fréquences > Table de contingence > échantillons indépendants

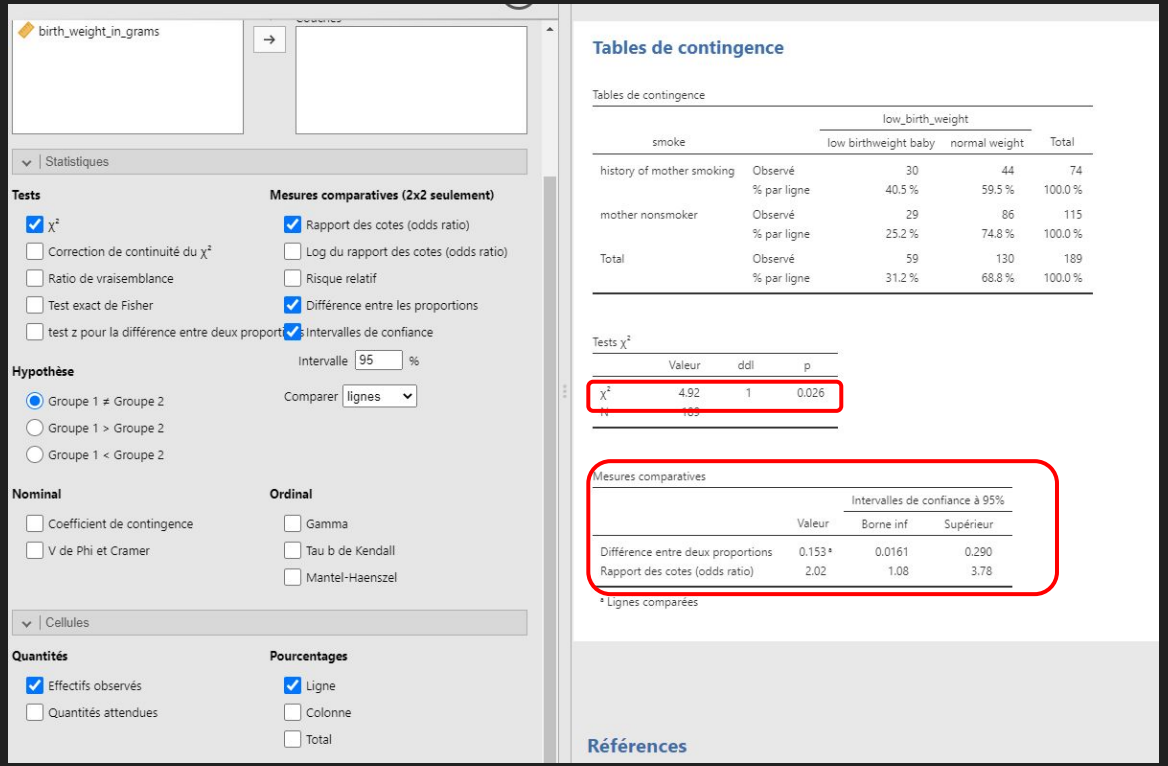
Évaluer l'association entre le statut tabagique et le petit poids de naissance

Si toutes les quantités attendues >= 5
=> **Chi-2**
Si au moins une quantité attendue entre 3 et 5
=> **Chi-2 avec correction de continuité**
Si au moins une quantité attendue < 3
=> **Fisher**



Évaluer l’association entre le statut tabagique et le petit poids de naissance

Si toutes les quantités attendues >= 5
=> **Chi-2**
Si au moins une quantité attendue entre 3 et 5
=> **Chi-2 avec correction de continuité**
Si au moins une quantité attendue < 3
=> **Fisher**



Évaluer l’association entre l’irritabilité utérine et le poids de naissance

Évaluer l'association entre le statut tabagique et le petit poids de naissance

Si toutes les quantités attendues >= 5

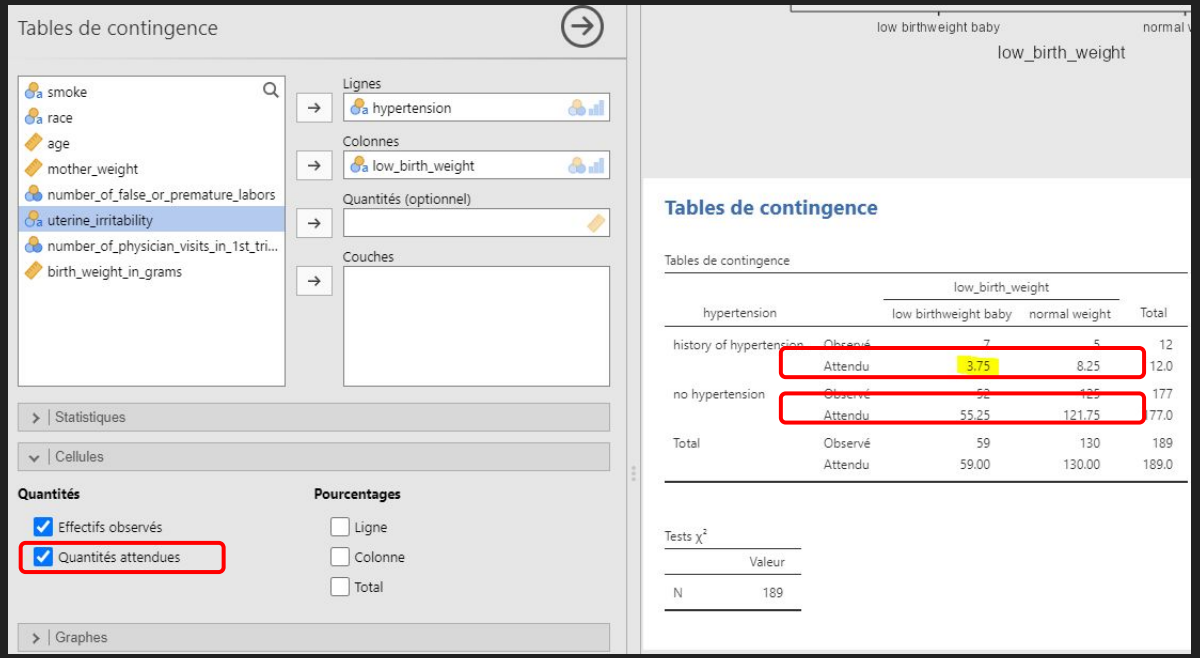
=> **Chi-2**

Si au moins une quantité attendue entre 3 et 5

=> **Chi-2 avec correction de continuité**

Si au moins une quantité attendue < 3

=> **Fisher**



Évaluer l'association entre le statut tabagique et le petit poids de naissance

Si toutes les quantités attendues >= 5

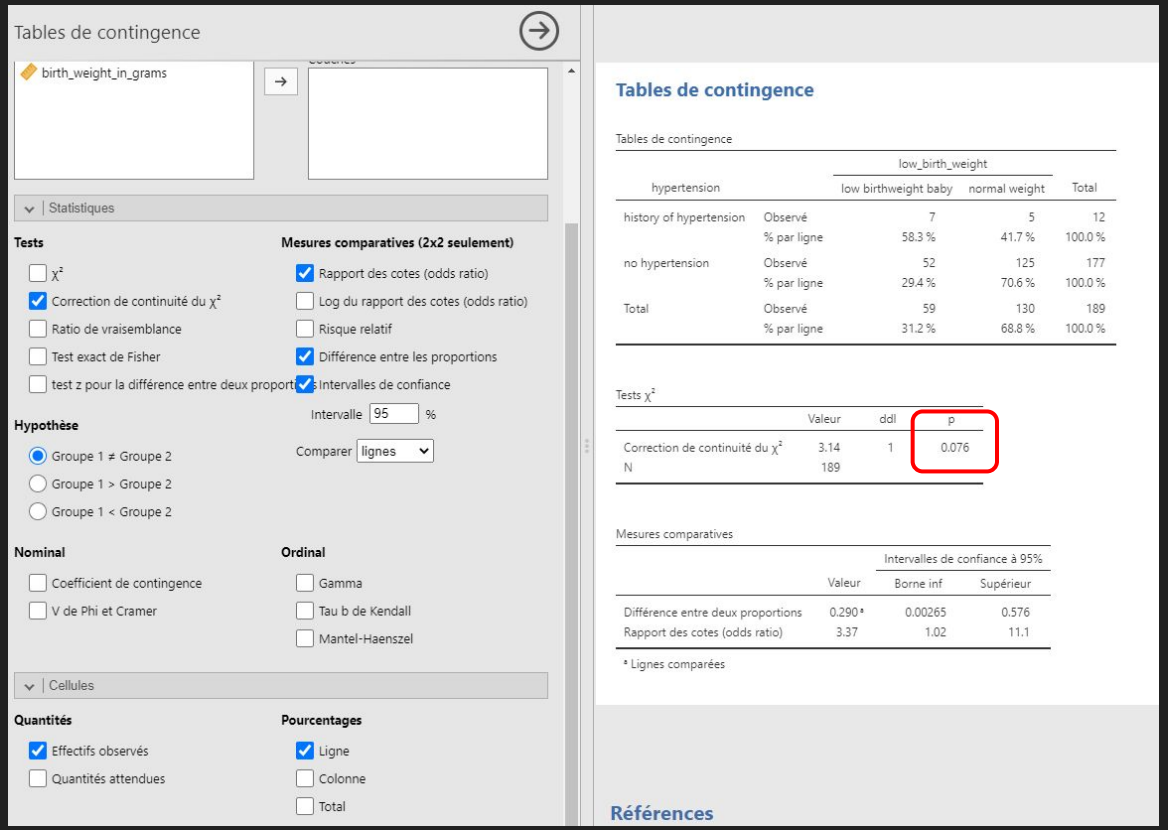
=> **Chi-2**

Si au moins une quantité attendue entre 3 et 5

=> **Chi-2 avec correction de continuité**

Si au moins une quantité attendue < 3

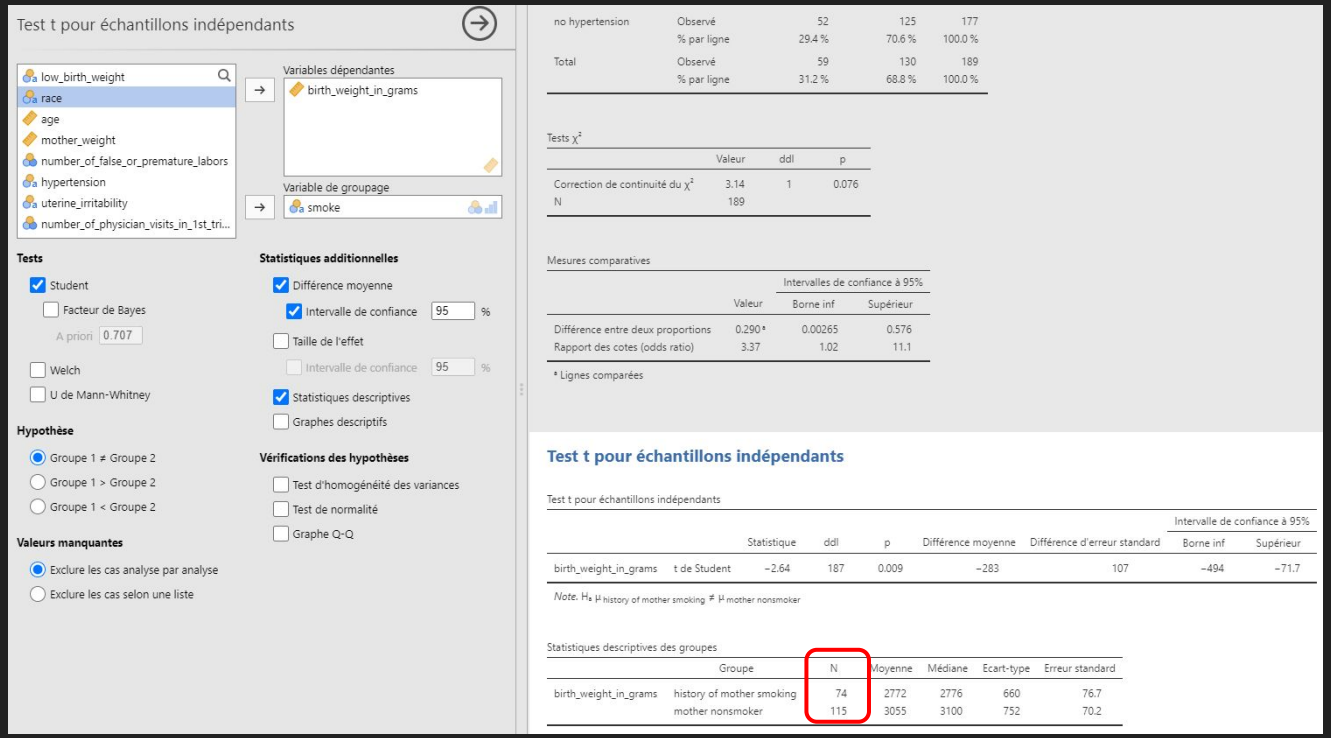
=> **Fisher**



Évaluer l’association entre le statut tabagique et le poids de naissance

Analyses > Test t > Test t pour échantillons indépendants

Évaluer l'association entre le statut tabagique et le poids de naissance

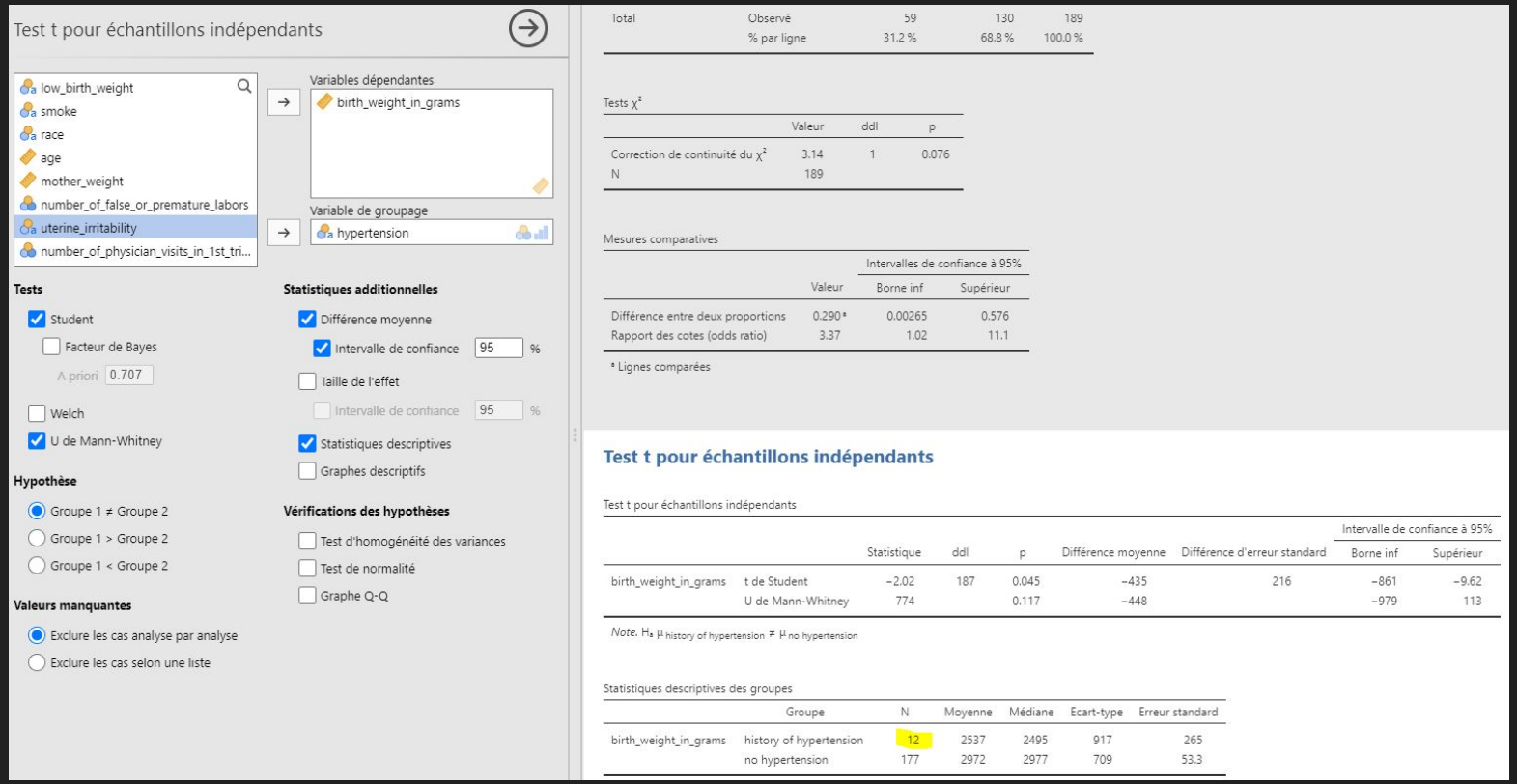


N >= 30 dans les 2 groupes
=> Student

N < 30 dans un des groupes
=> U de Mann Whitney
(= Wilcoxon)

Évaluer l’association entre l’hypertension et le poids de naissance

Évaluer l’association entre l’hypertension et le poids de naissance



Total	Observé	59	130	189
	% par ligne	31.2 %	68.8 %	100.0 %

Tests χ^2

	Valeur	ddl	p
Correction de continuité du χ^2	3.14	1	0.076
N	189		

Mesures comparatives

		Intervalles de confiance à 95%	
	Valeur	Borne inf	Supérieur
Différence entre deux proportions	0.290[a]	0.00265	0.576
Rapport des cotes (odds ratio)	3.37	1.02	11.1

[a] Lignes comparées

Test t pour échantillons indépendants

Test t pour échantillons indépendants

						Intervalle de confiance à 95%	
		Statistique	ddl	p	Différence moyenne	Différence d'erreur standard	Borne inf	Supérieur
birth_weight_in_grams	t de Student	−2.02	187	0.045	−435	216	−861	−9.62
	U de Mann-Whitney	774		0.117	−448		−979	113

Note. H_a $\mu_{\text{history of hypertension}} \neq \mu_{\text{no hypertension}}$

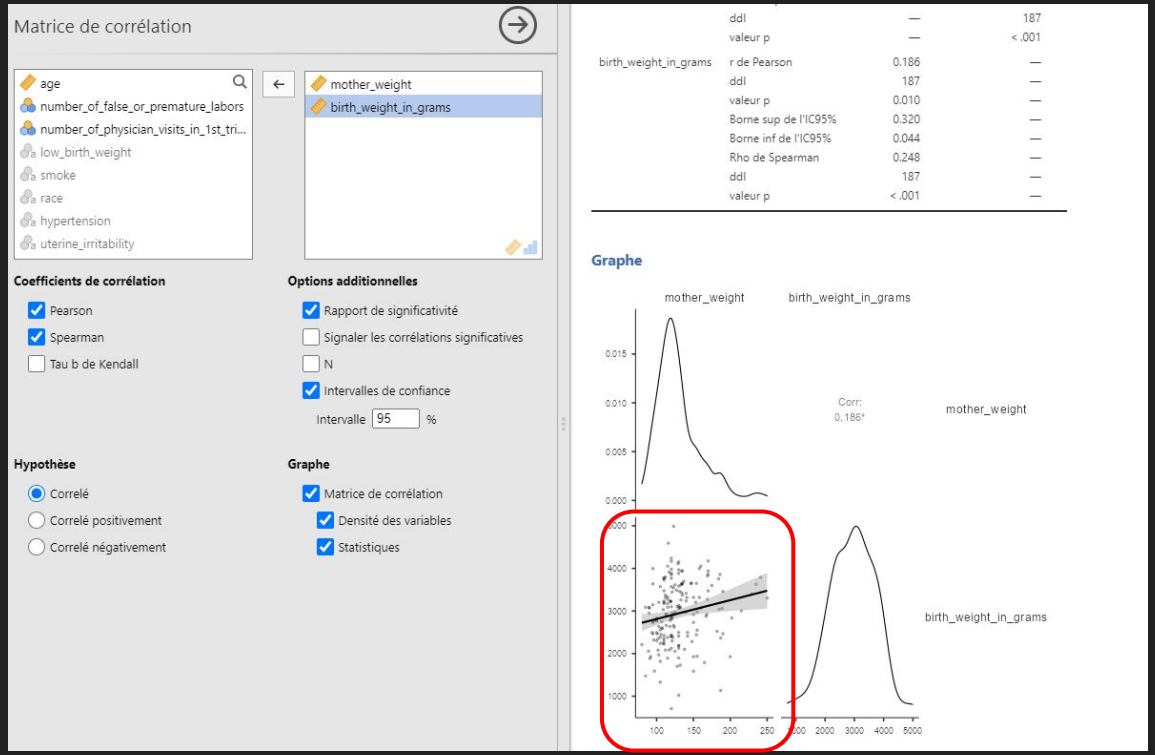
Statistiques descriptives des groupes

	Groupe	N	Moyenne	Médiane	Ecart-type	Erreur standard
birth_weight_in_grams	history of hypertension	12	2537	2495	917	265
	no hypertension	177	2972	2977	709	53.3

Association entre poids de la mère et poids de l’enfant

Analyses > Régression > Matrices de corrélation

Association entre poids de la mère et poids de l'enfant



Relation linéaire

=> Pearson

Relation non linéaire (monotonique)

=> Spearman

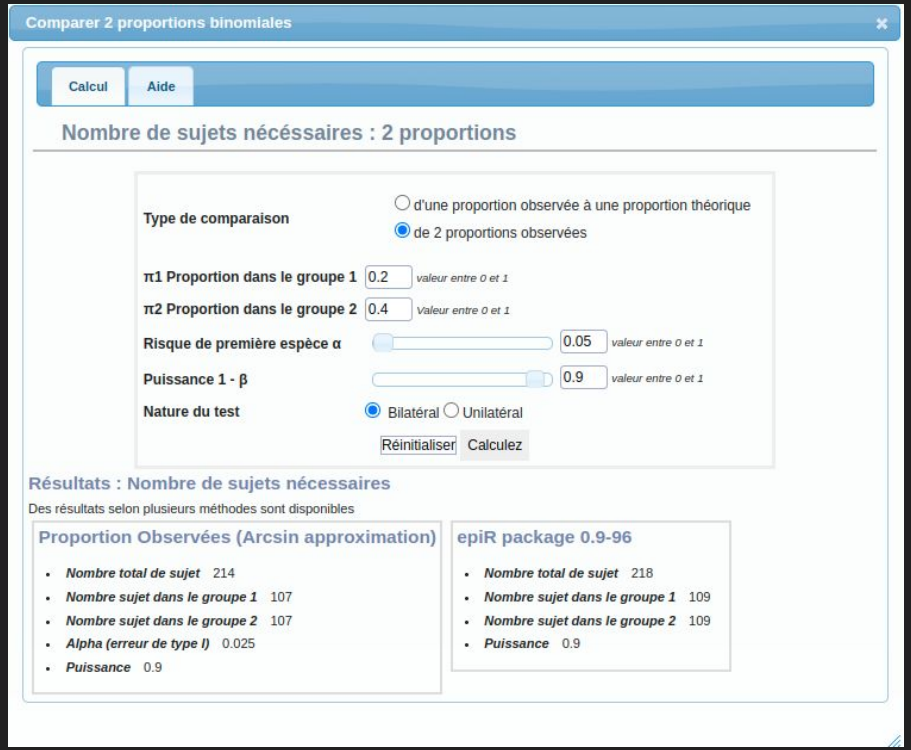
Choix du test

	Variable qualitative	**Variable quantitative**
Variable qualitative	effectif attendu >= 5 **=> Chi-2** effectif attendu >= 3 et < 5 **=> Chi-2 + correction** effectif attendu < 3 **=> Fisher**	effectif >= 30 (dans chaque groupe) **=> Student** effectif < 30 **=> U de Mann Whitney (= Wilcoxon)**
Variable quantitative	effectif >= 30 (dans chaque groupe) **=> Student** effectif < 30 **=> U de Mann Whitney (= Wilcoxon)**	relation linéaire **=> Pearson** relation non linéaire (monotone) **=> Spearman**
Survie	Log-rank	

Taille d’étude avec Biostatgv

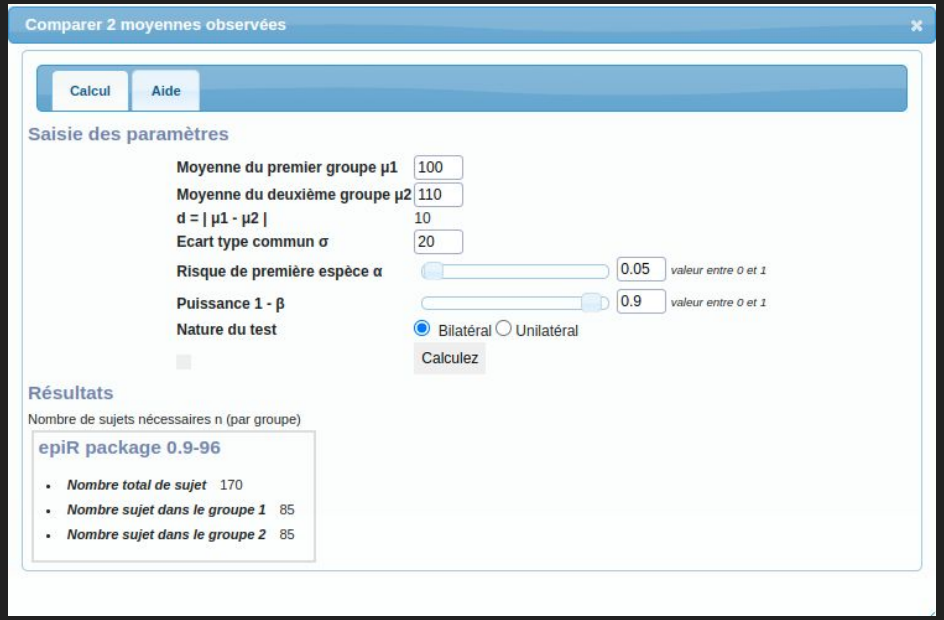
https://biostatgv.sentiweb.fr/?module=etudes/sujets

Outcome qualitatif



Nombre de sujets nécessaire pour une proportion attendue d'événement dans les deux groupes de **20%** et **40%**, avec un puissance de **90%** (probabilité de mettre en évidence si elle existe) et un risque alpha de **5%** (probabilité de mettre en évidence une différence alors qu'elle n'existe pas)

Outcome quantitatif



Nombre de sujets nécessaire pour mettre en évidence une différence moyenne de **10** entre les deux groupes, avec un écart-type de **20** (variabilité de la mesure) avec un puissance de **90%** (probabilité de mettre en évidence si elle existe) et un risque alpha de **5%** (probabilité de mettre en évidence une différence alors qu'elle n'existe pas)

Take home message

Choix du test

	Variable qualitative	**Variable quantitative**
Variable qualitative	effectif attendu >= 5 **=> Chi-2** effectif attendu >= 3 et < 5 **=> Chi-2 + correction** effectif attendu < 3 **=> Fisher**	effectif >= 30 (dans chaque groupe) **=> Student** effectif < 30 **=> U de Mann Whitney (= Wilcoxon)**
Variable quantitative	effectif >= 30 (dans chaque groupe) **=> Student** effectif < 30 **=> U de Mann Whitney (= Wilcoxon)**	relation linéaire **=> Pearson** relation non linéaire (monotone) **=> Spearman**
Survie	Log-rank	

Questions ?