Quelle est la fiabilité des chandelles standard ?

C'est une question difficile. Les relations qui donnent la luminosité intrinsèque sont-elles réellement universelles ? Une céphéide brille-t-elle de la même façon dans toutes les galaxies ? En effet, les galaxies n'ont pas toutes la même métallicité (le contenu en éléments plus lourds que l'hydrogène, l'hélium et le lithium). La composition des étoiles y est donc différente, ce qui peut influer sur leur comportement. Même question pour les supernovæ : explosaient-elles de la même façon dans l'Univers plus jeune ?

Valider ces hypothèses est un défi majeur. Plus on regarde loin dans l'Univers et plus il est difficile de détecter des supernovæ. Il faut regarder dans la bonne direction, au bon moment. Les futurs grands instruments comme le télescope spatial *James-Webb*, l'*E-ELT* (le futur observatoire géant que l'Europe construit au Chili) ou l'observatoire *Vera-Rubin* (nouveau nom du *LSST*) amélioreront la situation, mais ne permettront probablement pas de confirmer l'universalité des chandelles standard.

Existe-t-il d'autres méthodes qui soient indépendantes de ces chandelles standard ?

Une toute nouvelle méthode consiste à utiliser les ondes gravitationnelles émises lors de la fusion de deux étoiles à neutrons, ondes que la collaboration *Ligo-Virgo* a détectées pour la première fois en 2017. On parle de « sirènes standard » ! L'amplitude des ondes gravitationnelles captées permet de déterminer la distance de l'événement. Et contrairement à la coalescence de deux trous noirs, l'événement s'accompagne d'une émission intense d'ondes électromagnétiques dont on déduit, grâce au *redshift*, la vitesse de récession de la paire d'étoiles. Avec seulement six fusions de ce type détectées pour l'instant, la collaboration *Ligo-Virgo* a obtenu pour H_0 une estimation de 68 kilomètres par seconde et par mégaparsec. Mais les barres d'erreur sont encore très grandes. Les spécialistes de cette méthode estiment qu'il faudrait une cinquantaine d'événements pour avoir une valeur compétitive.

Cette technique est encore assez jeune. Il se pose beaucoup de questions sur sa robustesse. Par exemple, il faut déterminer avec précision la masse des étoiles qui ont fusionné pour connaître l'amplitude initiale, ce qui est loin d'être évident et dépend des modèles. Autre difficulté, les ondes gravitationnelles peuvent être amplifiées sur leur chemin, par un effet de lentille gravitationnelle ; la coalescence peut alors paraître plus proche qu'elle ne l'est en réalité.

Et vous, sur quelle méthode travaillez-vous ?

Dans notre projet *HoLiCOW*, l'idée est d'utiliser ce phénomène bien connu en relativité générale qu'est la lentille gravitationnelle. Prenez un objet lointain comme un quasar (une galaxie au noyau très actif et émettant beaucoup de lumière) et imaginez qu'une galaxie très massive se trouve sur la ligne de visée. Celle-ci agit comme une lentille dans un dispositif optique. En déformant l'espace-temps dans son voisinage, la galaxie-lentille dévie la lumière du quasar qui passe près d'elle et la focalise vers nous. La lumière nous parvient en suivant différents chemins de longueurs différentes, ce qui se manifeste par des images multiples du quasar. Quand la lumière du quasar fluctue, celle de ses images aussi, mais en décalé dans le temps (de plusieurs mois ou plus), car les chemins parcourus par la lumière ne sont pas de même longueur. En 1964, Sjur Refsdal a montré qu'il est possible d'exprimer la constante de Hubble en fonction de ce délai temporel et du *redshift*.

L'analyse est complexe, car divers phénomènes perturbent la mesure du délai temporel. Il faut aussi modéliser de la façon la plus réaliste possible la distribution de masse de la galaxie-lentille pour reproduire correctement le champ gravitationnel de la lentille qui, par dilatation du temps, influe sur le délai. Enfin, il faut aussi tenir compte de la matière présente sur le parcours de la lumière.

Qui plus est, de telles configurations de lentilles pour un quasar sont rares. Nous en avons étudié sept avec précision. Nous avons obtenu des valeurs de la constante de Hubble toutes très voisines, ce qui nous donne confiance en notre méthode d'analyse. Notre estimation est d'environ 73,3 kilomètres par seconde et par mégaparsec. Ce résultat est très proche de celui de *SHoES*. Avec les instruments d'observation à venir, nous espérons exploiter 50 quasars d'ici à quelques années. C'est l'objectif du nouveau consortium international *TDCOSMO* (*Time delay cosmography*).

Peut-on espérer que la crise cosmologique sera bientôt résolue ?

Aujourd'hui, la diversité de méthodes indépendantes de mesure de H_0, avec des erreurs systématiques de natures différentes, enrichit le débat. Nous ne savons pas encore expliquer tous les écarts dans les mesures, mais, en comparant méticuleusement ces approches, nous devrions bientôt commencer à y voir plus clair.

Si le désaccord persiste, il faudra probablement repenser les modèles. Différentes pistes existent – les conditions dans le plasma primordial, la nature de l'énergie sombre, et bien d'autres. Nous avons même trop de solutions : pour l'instant, nous avons une serrure, mais toutes les clés fonctionnent. Nous ne savons pas encore laquelle est la bonne !

Propos recueillis par Sean Bailly

BIBLIOGRAPHIE

K. C. Wong *et al.*, **H0LiCOW XIII. A 2.4 % measurement of H_0 from lensed quasars : 5.3 σ tension between early and late-Universe probes,** à paraître dans *MNRAS*, 2020. (prépublication sur arXiv).

A. G. Riess *et al.*, **Large Magellanic Cloud cepheid standards provide a 1 % foundation for the determination of the Hubble constant and stronger evidence for physics beyond ΛCDM,** *Astrophys. J.*, vol. 876(1), article 85, 2019.

W. L. Freedman *et al.*, **The Carnegie-Chicago Hubble program. VIII. An independent determination of the Hubble constant based on the tip of the red giant branch,** *Astrophys. J.*, vol. 882(1), article 34, 2019.

L'ESSENTIEL

> L'utilisation de la valeur-*p*, qui sert depuis près d'un siècle à jauger la qualité statistique de résultats expérimentaux, a donné une fausse illusion de certitude et a mené à des crises de reproductibilité dans de nombreux champs scientifiques.

> Refonte totale ou simples aménagements ? Malgré le consensus grandissant pour une réforme de l'analyse statistique, les chercheurs sont en désaccord sur l'ampleur à lui donner.

> Diverses pistes sont envisagées, mais une chose est sûre : il est temps pour les scientifiques et le grand public d'accepter l'inconfort du doute.

L'AUTRICE

LYDIA DENWORTH
journaliste scientifique à New York, aux États-Unis

La valeur-*p* : un problème significatif

Les méthodes statistiques pour jauger la pertinence d'un résultat expérimental sont attaquées de toutes parts. Résisteront-elles ?

En 1925, le généticien et statisticien britannique Ronald Fisher publiait un livre intitulé *Statistical Methods for Research Workers* (*Les Méthodes statistiques adaptées à la recherche scientifique*). Il n'avait *a priori* rien d'un best-seller, pourtant il remporta un énorme succès et installa Fisher comme le père des statistiques modernes. Dans l'ouvrage, Fisher expliquait aux chercheurs comment utiliser les statistiques pour éprouver la robustesse de leurs conclusions, afin qu'ils puissent décider ou non de donner une suite à leurs travaux. Fisher y décrivait notamment un test qui examinait la compatibilité des résultats avec un modèle et produisait en sortie un nombre : la valeur-*p* ou *p*-valeur (*p-value* en anglais). Il fixa un seuil à $p = 0{,}05$: « Il est commode de prendre ce point comme limite pour juger si une déviation [par rapport au modèle] doit être considérée comme significative ou non. » Fisher conseillait aux chercheurs de poursuivre leurs travaux s'ils obtenaient des valeurs-*p* en deçà. Ainsi naissait l'idée qu'une valeur-*p* inférieure à 0,05 apposait le sceau « statistiquement significatif » sur des recherches.

Aujourd'hui, presque un siècle plus tard, une valeur-*p* inférieure à 0,05 est considérée comme la référence pour jauger la qualité statistique d'une expérience. Elle ouvre aux chercheurs les portes du monde académique – au financement et à la publication – et dès lors détermine en partie la majorité des résultats scientifiques publiés. Mais même Fisher avait conscience des limites de ce critère statistique. La plupart de ces travers sont connus depuis des décennies. « Le recours excessif aux tests de pertinence statistique, écrivait le psychologue américain Paul Meehl en 1978, est une façon appauvrie de faire de la science. » La valeur-*p* est fréquemment mal interprétée et l'on a tendance à oublier qu'un résultat statistiquement significatif n'est pas forcément significatif en pratique.

En outre, la méthodologie pratiquée dans les laboratoires donne la possibilité aux expérimentateurs, consciemment ou non, d'augmenter >