Pour que l'effet trouvé soit statistiquement significatif, il faudrait qu'il soit plus grand que 1,96 fois l'écart-type, c'est-à-dire au-dessus de 8,4% (l'intervalle de confiance ne contiendrait alors pas 0). Exprimé autrement: la probabilité de rejeter à tort la valeur nulle est de 5%.

Mais est-ce bien raisonnable d'envisager des effets significatifs aussi grands que 8,4%? Non, bien sûr. C'est ce que l'on nomme une erreur de magnitude. L'étude est construite de telle façon que tout résultat statistiquement significatif surestime le véritable effet (qui ne peut dépasser 1%). S'y ajoutent des erreurs de signe quand l'estimation trouvée est de signe opposé au véritable effet (ici: «les parents beaux ont plus de garçons que de filles»). Ainsi, on distingue deux types d'effet: positif, si les parents «beaux» ont plus de filles que les autres, et négatif, s'ils en ont moins.

Pour illustrer les probabilités associées à ces erreurs, envisageons quatre scénarios fondés sur des écarts-types égaux à 4,3% (*voir l'encadré ci-contre*). Ils montrent qu'une étude à partir de cette taille d'échantillon (3000 couples environ) n'est pas pertinente pour estimer des petits effets de l'ordre du 1%. Cela est dû notamment à la valeur de l'écart-type qui est particulièrement élevée. C'est pourquoi les études de la répartition des sexes des nouveau-nés utilisent des échantillons beaucoup plus grands, fondés sur de vastes bases de données démographiques, comptant plus d'un million d'individus.

L'ANALYSE BAYÉSIENNE

On résume tout ce que l'on sait sur l'effet à analyser, en utilisant des sources extérieures déjà connues, par une distribution *a priori*. Cette distribution sera modifiée (conditionnée) par le résultat de l'expérience pour donner la distribution *a posteriori*. Si l'on ne savait rien d'avance, on pourrait prendre une distribution *a priori* non informative et le résultat correspondrait à celui de l'approche fréquentiste. Ici, la distribution obtenue *a posteriori* serait alors approximativement une gaussienne, de moyenne 4,7% et d'écart-type 4,3%, ce qui correspondrait à une probabilité de 86% environ que l'effet réel soit positif. En général, plus la distribution *a priori* est concentrée autour de zéro (hypothèse selon laquelle l'effet réel sur la différence des sexes est faible), plus la probabilité *a posteriori* est proche de 50%.

Choisissons, par exemple, une distribution *a priori* gaussienne (en forme de cloche), centrée sur zéro avec une forme telle que la différence réelle de probabilité d'avoir une fille (selon que les parents sont beaux ou non) soit proche de zéro, avec des probabilités 50%, 90% et 94% d'être respectivement dans les intervalles: $[-0{,}3; 0{,}3]$, $[-1; 1]$, et $[-3; 3]$. Pourquoi centrer la distribution sur zéro? Parce que nous n'avons pas d'*a priori* sur le signe de la différence réelle de probabilité de naissance de filles en fonction de la beauté des parents.

À l'étape suivante, nous calculons, à partir de cette distribution *a priori* et des données, la distribution *a posteriori* de l'effet. Pour résumer, la distribution *a posteriori* donne une probabilité que l'effet soit positif (les parents beaux ont plus de filles) de seulement 58%, dont 45% que cette différence positive soit inférieure à 1%. Cette analyse dépend – mais peu – de la distribution *a priori*; par exemple, si l'on décide d'élargir la courbe de distribution autour de zéro, la probabilité que l'effet réel soit positif n'augmente que de 7% (pour atteindre 65%). Le fait de changer de famille de courbes de distribution a peu d'effets sur les résultats: les effets réels restent faibles, ce que confirment les données.

L'idéal scientifique dans l'inférence sur une quantité (ici, le lien entre la beauté des parents et le rapport des sexes à la naissance) est la >

QUATRE SCÉNARIOS POUR COMPRENDRE LES ERREURS

On peut étudier les erreurs liées aux résultats de Kanazawa en déterminant les probabilités associées *via* quatre scénarios. Dans le premier, que nous nommerons «Différence exactement nulle», supposons qu'il n'y a aucune corrélation entre la beauté des parents et le sexe de leur enfant (il n'y a pas de différence entre les deux groupes de parents). Il existe toujours une petite probabilité – 5% – que le résultat trouvé soit statistiquement significatif. Mais ce résultat sera quand même faux, on a cinq% de chances de se tromper sur la valeur et ce quel que soit le signe de l'effet, positif ou négatif.

Dans le scénario dit de 0,3% exactement, les parents beaux ont 0,3 % de probabilité supplémentaire d'avoir une fille – valeur plausible. On a toujours une probabilité de 5% de trouver un résultat statistiquement significatif à l'issue de l'étude. Comment se répartissent ces 5% de part et d'autre de la valeur réelle? Dans ces conditions, on a une probabilité de 3% d'observer un effet statistiquement significatif positif et 2% d'observer un effet statistiquement significatif négatif. Mais dans chaque cas, l'effet estimé – au moins 8,4% pour qu'il soit significatif – sera beaucoup trop éloigné de la valeur réelle de 0,3%. Ainsi, l'erreur de magnitude sera importante. Par ailleurs, la probabilité d'aller dans le mauvais sens – erreur de signe – sera égale à 2/5. Imaginons que l'on se retrouve entre les bornes de l'intervalle de confiance à 95%, donc proches de la valeur réelle. Dans ce cas, l'erreur de magnitude serait faible, mais la probabilité de faire une erreur de signe n'est pas négligeable: 47,5 %, proche des 50% observés en l'absence d'effet. Ainsi, le signe observé n'apporte pratiquement aucune information sur le sens d'une éventuelle différence.

Troisième scénario: si les parents séduisants ont une probabilité d'avoir une fille supérieure de 1% (ce qui semble être l'effet maximal possible), alors on a réciproquement des probabilités de 4 et 1% d'avoir des effets positif et négatif statistiquement significatifs. Globalement, il y a une probabilité de 40% de faire une erreur sur le signe de l'estimation; là encore, l'estimation donne peu d'informations sur le signe ou l'ordre de grandeur de l'effet réel.

Enfin, envisageons le quatrième scénario: même si la différence réelle était de 3% – valeur improbable –, il n'y aurait encore que 10% de probabilité d'obtenir un résultat statistiquement significatif. Dans ce cas, nous aurions une probabilité de 24% de faire une erreur de signe. Ce type d'étude est donc peu informatif.