Supporting Text

The likelihood function for male models that incorporate bias correction factor

Denote r_{nm}, the bias correction factor for families of shape $F(n,m)$, $3 \le n \le 5$ and $m \ge 2$. The log-likelihood function incorporating r_{nm} is given by

[Eq. 1] $LL(\theta) = \sum_{n=3}^{5} \sum_{m=2}^{n} obs(n,m) \log(P_\theta(n,m \mid m \ge 2))$

where $P_\theta(n,m \mid m \ge 2)$ is the conditional probability that equals to $\frac{r_{nm} P_\theta(n,m)}{\sum_{k \ge 2} r_{nk} P_\theta(n,k)}$. In the paper, when *m* = *2*, we set $r_{nm} = 1$ and when m >2, r_{nm}s have the same value, estimated using the gender asymmetry method in the second section of *Results*.

The likelihood function for models that distinguish males and females by penetrance.

Denote $P(n, m, n_{AM}, n_{UM}, n_{AF}, n_{UF})$ the probability that a family of size *n* with *m* children with autism has n_{AM} males with autism, n_{UM} typical males, n_{AF} females with autism and n_{UF} typical females. That probability under the model that distinguishes males and females by penetrance *p* is given by

[Eq. 2]
$$\begin{aligned} &P(n, m, n_{AM}, n_{UM}, n_{AF}, n_{UF}) \\ &= \int_0^1 \binom{n}{n_{AM}, n_{AF}, n_{UM}, n_{UF}} (q_m x)^{n_{AM}} (q_f p x)^{n_{AF}} (q_m (1-x))^{n_{UM}} (q_f (1-px))^{n_{UF}} f(x) dx \end{aligned}$$

where x is the probability of male offspring with autism, $q_m = 105/205$ and $q_f = 100/205$.

Let $F(n,m,n_{AM},n_{UM},n_{AF},n_{UF})$ denote the type of such a family and we observe $obs(n,m,n_{AM},n_{UM},n_{AF},n_{UF})$ families of that type. The log likelihood function for families with $3 \le n \le 5$ and $m \ge 2$ is given by

[Eq. 3] $$LL(\theta) = \sum_{n=3}^{5}\sum_{m=2}^{n} obs(n,m,n_1,n_2,n_3,n_4)\log(P_\theta(n,m,n_1,n_2,n_3,n_4 \mid m \ge 2)),$$

where $P_\theta(n,m,n_1,n_2,n_3,n_4 \mid m \ge 2)$ equals to $\frac{r_{nm}P_\theta(n,m,n_1,n_2,n_3,n_4)}{\sum_{k\ge2} r_{nk}P_\theta(n,k,n_1,n_2,n_3,n_4)}$. Here r_{nm} are ascertainment bias factors as described previously.

Simulation in goodness-of-fit test for model 2(m) and 2(m/f/p)

For model 2(m), based on the parameters estimated from the AGRE set, we calculated the probability $P_\theta(n,m \mid m \ge 2)$ for $3 \le n \le 5$ and $2 \le m \le n$, and then computed the expected number of families of shape *F(n,m)* conditional on the total number of male sibships of size *n* with at least two autistic children. Bias correction factor 1.14 was used for AGRE set, 1.0 was used for UMICH set, and 1.06 was used for IAN set. The Pearson χ^2 statistic was calculated to summarize the difference between the observed number of *F(n,m)* families (in the AGRE, University of Michigan, or IAN sets) and the expected number of *F(n,m)* families from our model, for $n = 3, 4, 5$, and $2 \le m \le n$. Simulations based on multinomial distribution $P_\theta(n,m \mid m \ge 2)$ were carried out 1,000 times to compute the empirical distribution of the χ^2 statistic which was then used to calculate the *P* value.

Similarly for model 2(m/f/p), based on the parameters estimated from the AGRE set, we calculated the probability $P_\theta(n,m,n_{1,}n_{2,}n_{3,}n_4 \mid m \ge 2)$ for $3 \le n \le 5$ and $2 \le m \le n$, and then computed the expected number of families $F(n,m,n_1,n_2,n_3,n_4)$ conditional on the total number of families of size n with at least two autistic children. Pearson χ^2 statistic was calculated to summarize the difference between the observed number of $F(n,m,n_{AM},n_{UM},n_{AF},n_{UF})$ families (in the AGRE, University of Michigan, or IAN sets) and the expected number of $F(n,m,n_{AM},n_{UM},n_{AF},n_{UF})$ families from our model, for n = *3, 4, 5* and $2 \leq m \leq n$. Simulations based on multinomial distribution $P_\theta(n,m,n_{1,}n_{2,}n_{3,}n_4 \mid m \geq 2)$ were carried out 1,000 times to compute the empirical distribution of the χ^2 statistic which was then used to calculate the *P* value.

Monte Carlo simulation for calculating P values in likelihood ratio test

When testing a three-component model (alternative hypothesis) against a two-component model (null hypothesis), we estimated the *P* value using a Monte Carlo method. We simulated 100 sets of data under the null model based on the parameters estimated from the real data, as done in the simulation for goodness-of-fit test. We then computed the log-likelihood ratio between the null model and the alternative model on the simulated data, to establish a reference distribution of the log-likelihood ratio. The *P* value was then obtained by comparing the observed log-likelihood ratio to this reference distribution. Similar procedures were applied to calculate *P* values for other comparisons in Table 2.