Fake Correlations in Data

Scientific popular news is bombarded with questionable studies providing rather bizarre correlations between two different things, which are in each case "statistically significant". One can (and should) ask the question: how likely are we to find statistically significant correlations between data which is, in truth, uncorrelated. We will provide a heuristic answer to this question in this problem.

For simplicity, let us provide a very crude model of a data set. Consider a set of MN i.i.d Bernoulli(1/2) random variables $x_i^\alpha = \pm 1$. $i = 1, \ldots, N$ denotes an individual in the data set, and $\alpha = 1, \ldots, M$ represents one bit of information about that individual. If N is very large, we can approximately assume that the $2^M - 1$ random variables

$$X^{\alpha\cdots\beta} \equiv \sum_{i=1}^{N} x_i^\alpha \cdots x_i^\beta$$

are i.i.d. There are $2^M - 1$ random variables because we can choose to draw any of the possible 2^M combinations of either including or excluding each bit of information α, although we must include at least one bit. Assume for this problem that $M \gg 1$, but $N \gg M$. Also, for simplicity, you may assume that N is an even number, if this helps your calculations.

(a) What is the probability distribution on $X^{\alpha\cdots\beta}$?[1] What is the mean and variance of $X^{\alpha\cdots\beta}$?

(b) Historically, statisticians would say that if

$$\frac{X^{\alpha\cdots\beta} - \langle X^{\alpha\cdots\beta}\rangle}{\sqrt{\mathrm{Var}(X^{\alpha\cdots\beta})}} > \alpha \gtrsim 1,$$

then that particular instance of $X^{\alpha\cdots\beta}$ is rare enough that it should be counted as "statistically significant". Estimate the typical number of statistically significant $X^{\alpha\cdots\beta}$ variables, given our set of random data.

(c) Show that if we do not want any of the fake correlations found in part (b) to be considered as statistically significant, we must take

$$\alpha \gtrsim \sqrt{M}.$$

[1] You should have to do very little work to provide the answer!