Reverse cocaine sting

In 1981 in Florida, police raided a house and found 496 bags containing white powder. They tested four random bags and the results were positive for cocaine. The bags were then either destroyed or taken into evidence, depending on which account you read.

Two of the remaining 492 bags were used in a sting operation and were sold to an alleged dealer. Let's say his name was Tony. The bags were sold, but before Tony was taken into custody, the two sold bags went missing. It is not clear what happened to the other 490 bags. Presumably they were destroyed as well.

The two bags that were sold to Tony were never tested and never recovered, so what is the probability that Tony actually bought some legal substance that looks like cocaine, e.g. baking soda?

Let X represent the number of bags containing cocaine that were seized in the original raid and let Y represent the number of bags containing baking soda that were purchased by Tony. The probability that Tony is innocent of purchasing bags that contained cocaine is

$$P(X = 4 \cap Y = 2) = P(X = 4) \cdot P(Y = 2 | X = 4)$$

The probabilities $P(X = 4)$ and $P(Y = 2|X = 4)$ depend on how many bags out of the original 496 actually contained cocaine. Say this number is K. Since four bags tested positive for cocaine, $P(A) = 0$ when $K < 4$. If $K = 4$, exactly four bags contained cocaine and $P(Y = 2|X = 4) = 1$, but $P(X = 4)$ is very close to 0, so the probability of "innocence" is also close to 0.

At the other end of the scale, if $K \geq 495$, then $P(Y = 2|X = 4) = 0$. As K varies from 4 to 494, $P(X = 4)$ increases and $P(Y = 2|X = 4)$ decreases, creating a maximum probability of innocence somewhere in $[4, 494]$.

Hypergeometric distributions

Draw n times from a population of size N without replacement. If the population contains K successes, then the probability of obtaining k successes in the n draws is

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

In the special case where $n = k$, the formula reduces to

$$P(X = k) = \frac{\binom{K}{k}}{\binom{N}{k}}$$

Reverse cocaine sting solution

$$P(X = 4) = \frac{\binom{K}{4}}{\binom{496}{4}}$$

$$P(Y = 2|X = 4) = \frac{\binom{496-K}{2}}{\binom{492}{2}}$$

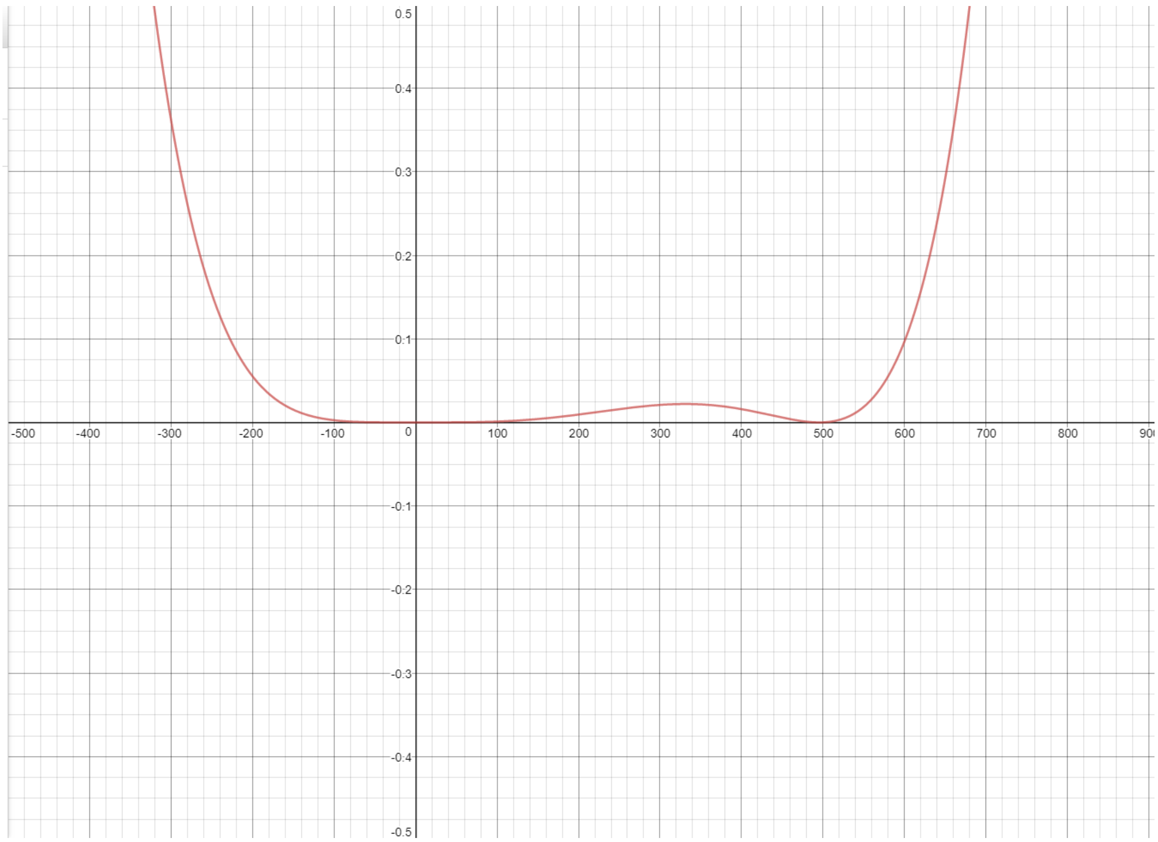
$$P(Y = 2 \cap X = 4) = \frac{\binom{K}{4}\binom{496-K}{2}}{\binom{496}{4}\binom{492}{2}}$$

$$= \frac{\dfrac{K!}{4!(K-4)!} \cdot \dfrac{(496-K)!}{2!(494-K)!}}{\dfrac{496!}{4!492!} \cdot \dfrac{492!}{2!490!}}$$

$$= \frac{(K)(K-1)(K-2)(K-3)(496-K)(495-K)}{(496)(495)(494)(493)(492)(491)}$$

The probability of innocence is polynomial in K with zeros at $K = 0, 1, 2, 3, 495, 496$ and is positive for $4 \le K \le 494$. Therefore, there is a maximal probability of innocence in $[4, 494]$ (bonus: name the theorem that guarantees the maximum).

Here's a graph of that polynomial for $-\infty < K < \infty$.



R Code

Determining the value of K for which the probability of innocence is maximal would be tedious without a computer even if you take logs first. But it's super simple in R.

The R function dhyper gives the values from a hypergeometric distribution. The signature is

$$\text{dhyper}(k, K, N - K, n, \log = \text{FALSE})$$

where N is the number of successes in the population, N is the population size, n is the number of trials, and k in the number of successes in n trials. (If you set log to TRUE, the function returns the log of the probability.)

Thus,

$$P(X = 4) = \text{dhyper}(4, K, 496 - K, 4)$$

and

$$P(Y = 2|X = 4) = \text{dhyper}(2, 496 - K, K - 4, 2)$$

for $K = 4, 5, \ldots 496$.

There is a Jupyter notebook containing the R code at https://github.com/AaronGladish/stats_rcso. If you have Jupyter notebooks on your desktop, you can download the .ipynb file and run it.