Inference using the potential population sizes you might be drawn from

Bostrom refers to this as 'indexical' vs. 'non indexical' probability. I think of the question more as something like: "When, and how, can we use the different population sizes that we might be part of to infer whether or not we *are* part of them?"

Example A: Motivation

This is a simple, and well defined case, easily analysed using Bayes Rule: We decide to run a lab experiment (possibly repeated many times). There are N subjects. They all know all details of the experiment, which are

- A fair coin is flipped, outcome unknown to the subjects
- If the coin lands heads then 1 subject is chosen at random, and asked to come into interview. The subjects have no information at all about how many or who has been chosen for interview, they know only whether they were, or were not, chosen for interview.
- If the coin lands tails then all N subjects are invited in for interview. Again, they know nothing other than their own invited/not-invited status.

The question is: what probability would a subject assign to the outcome that the coin landed heads? Does it change with N? I would argue that it does, and

$$P(\text{heads} \mid \text{invited}) = \frac{1}{N+1} \tag{1}$$

I wrote the above example having read the following and tried to understand where this guy was coming from (rather than reading his papers, as I'm a bit like that)

Example B: Improbable Conclusion

This is from the paper

$$\text{http://www.anthropic-principle.com/preprints/beauty/synthesis.pdf} \tag{2}$$

by Nick Bostrom, which he refers to as the **Presumptuous Philosopher**.

It is the year 2100 and physicists have narrowed down the search for a theory of everything to only two remaining plausible candidate theories, T_1 and T_2 (using considerations from super-duper symmetry). According to T_1 the world is very, very big but finite and there are a total of a trillion trillion observers in the cosmos. According to T_2, the world is very, very, very big but finite and there are a trillion trillion trillion observers. The super-duper symmetry considerations are indifferent between these two theories. Physicists are preparing a simple experiment that will falsify one of the theories. Enter the presumptuous philosopher: Hey guys, it is completely unnecessary for you to do the experiment, because I can already show to you that T_2 is about a trillion times more likely to be true than T_1! (Whereupon the presumptuous philosopher explains the Self-Indication Assumption: (See e.g. Bostrom 2002; Bostrom 2003, http://www.anthropic-principle.com/preprints/olum/sia.pdf)

Having considered this I wondered whether by setting things up a little differently from the motivating example I could get a different conclusion, I found that I could - and I think it highlights some of the implicit assumptions being made:

Better defined experiment, to promote discussion

We need to be slightly more powerful beings for this - capable of creating life at will. I guess we could do this using combinations of women and men (and possibly alcohol), if we were prepared to wait... Anyway, practicalities aside - we decide to run a lab experiment (possibly repeated many times). We will generate beings according to the flips of two coins. The beings will all know and understand all details of the experiment, which are

- We have two rooms, as big as we need. To avoid mutiny each room is split into hermetically sealed (but life supporting) units, so that inhabitants have absolutely no idea how many others there might be around them.
- We have two fair coins, Coin A (C_A), and Coin B (C_B), with independent (random) outcomes.
- C_A is flipped:
  - If C_A = heads we fill room A with N_A individuals. In particular, think about the case $N_A = 1$
  - If C_A = tails we don't put anyone in room A.
- C_B is flipped:
  - If C_B = heads we fill room B with N_B individuals. In particular, think about the case $N_B >> 1$
  - If C_B = tails we don't put anyone in room B.
- If there are any subjects to interview, we interview each of them - asking them what probability they assign the event "I am in room A"

So, as a participant, created in the course of the experiment, what should I say? First: what are the events that we want to think about, and what are events that we know how to define, and are clear on?

Want to think about

- The probability that I am in room A
- The probability that either of the coins landed heads given that I am here to ask the question

Questions we can define well

- The probability that the coins come down heads or tails, without conditioning on anyone in the rooms to wonder which room they're in.
- The probability that there will be some people in rooms A, B respectively

So we know how to reason from the perspective of the experimenter, but not from the perspective of the subjects. This is probably what Bostrom means by "Indexical vs. Non-Indexical" reasoning.

Does it help to think about bets that one should accept or not?

Suppose I am asked whether I will accept a bet about whether or not I'm in room A. Let's imagine that $N_B >> N_A$, and I'm offered $4:1$ odds. Well, depending on whether I should think like Bostrom or prefer to use "Proof by symbolic manipulation" (below) I appear to get different answers. First though we should be more clear about the set-up: As a participant in this study I get to know all the details of questions asked of me, and hence I know exactly the structure of this betting procedure. There are multiple possibilities:

1. Every participant is asked whether or not they want to participate in the bet
2. Only one person from each room is asked whether they want to participate in the bet
3. The experimenter only asks people in room B (he wants to make some cash, sneaky devil)

and so on. Obviously in case 3 we should not accept the bet. In case 2 we do want to accept the bet, indeed conditional on being asked we're now *evenly* split between thinking that we're in room A or room B. So, what about case 1? The following argument says that we should *not* accept the bet, hence that we're actually pretty likely to be in room B.

Balance argument

If the experimenter is going to ask everyone to bet (and let's assume for now that they all say yes), then when coin C_B = heads, the experimenter is going to make \$$N_B$. When C_B = tails then the experimenter will lose either \$0 (C_A is also tails) or \$$4N_A$ (C_A is heads). Overall the experimenters expectation, E_e is:

$$\begin{aligned} E_e &= \frac{1}{4}\left(N_B - 4N_A + 0 + (N_B - 4N_A)\right) \\ &= \frac{1}{2}\left(N_B - 4N_A\right) \qquad (3) \end{aligned}$$

From this we see that the experimenter gets a positive expectation from these bets whenever $N_B > 4N_A$. Note that we could replace this 4 with any n, and in order for the experimenter to profit he would simply need to ensure that $N_B > nN_A$.

The purpose of this is that it allows us to go from the easy perspective of the experimenter to the harder perspective of the participants. If we run the experiment many times, we see that on average, each bet the experimenter makes is weighted heavily in his favour (as we imagine that $N_B >> 4N_A$). Thus, on average, the individual bets of the participants must have expectation less than zero.

We can now rigorously show that not wanting to take this bet means that we're less likely than $\frac{1}{3}$ to be in room A. In this set-up a participant should take the bet if and only if her expectation is at least zero. If we let W_A be the amount she wins if she is in room A, and W_B be the amount (negative) that she wins if she's in room B, then the formula for the participants expectation, E_p is:

$$\begin{aligned} E_p &= P(R_A)W_A + P(R_B)W_B \\ &= 4P(R_A) - 1(1 - P_A) \\ &= 5P_A - 1 \end{aligned} \tag{4}$$

So, when the expectation on this bet for a participant is negative, then $5P_A - 1 < 0$, hence

$$P_A < \frac{1}{5} \tag{5}$$

Hence, as N_B gets bigger and bigger compared to N_A the experimenter can afford to give odds less and less in his favour, more extreme than $1 : n$ for any n (shown above), hence by the above argument the participants can be put into a situation where $P(R_A) < \frac{1}{n}$ for any n.

Which of these examples are comparable?

Are there intrinsic probabilities, or just probability models?

The question:"with what probability should I believe this statement with?" is the on of interest here. Above we think about arguments involving bets, and we do the same in the sleeping beauty case.

However, different betting arguments lead to different conclusions. If I am the only person from a room to be offered a bet that changes my position hugely from the situation where everybody is asked. However in this case we can explain this with the idea that being picked gave us extra data.