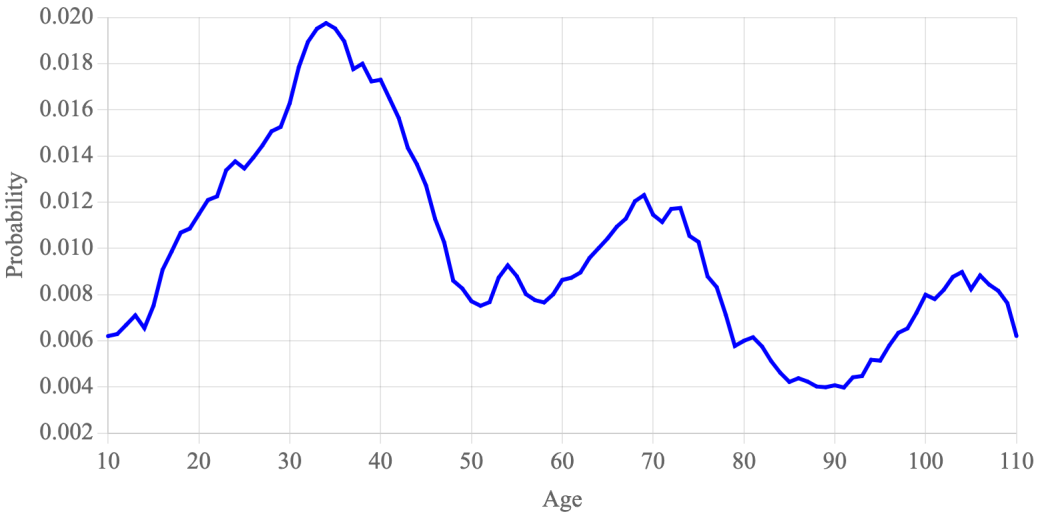
Name to Age

Because of shifting patterns in name popularity, a person's name is a hint as to their age. The United States publishes a data which contains counts of how many US residents were born with a given name in a given year, based off Social Security applications. We can use inference to compute the reverse probability distribution: an updated belief in a person's age, given their name. As a reminder, if I know the year someone was born, I can calculate their age within one year.

Query Name: Katherine ✔



Records with name: 589753

This demo is based on real data from US Social Security applications between 1914 and 2014. Thank you to https://www.kaggle.com/kaggle/us-baby-names for compiling the data. Download Data.

Computation

The US Social Security applications data provides you with a function: `count(year, name)` which returns the number of US citizens, born in a given year with a given name. You also have access to a list `names` which has each name ever given in the US and `years` which has all the years. This function is implicitly giving us the joint probability over names and birth year. The probability of a joint assignment to name and birth year can be estimated as the count of people with that name, born on that year, over the total number of people in the dataset. Let B be the year someone is born, and let N be their name. We will use k to denote the number of entries in the dataset:

$$P(B = b, N = n) \approx \frac{\text{count}(b, n)}{k}$$

The question we would really like to answer is: what is your belief that a resident was born in 1950, given that their name is Gary?

We can get started by applying the definition of conditional probability for random variables:

$$\text{P}(B = 1950 | N = \text{Gary}) = \frac{\text{P}(N = \text{Gary}, B = 1950)}{\text{P}(N = \text{Gary})}$$

Note: Bayes' Theorem is a more typical choice for inference tasks like this one. However, in this case its was necessary because it is easier to compute $P(B = b, N = n)$ than $P(N = n | B = b)$. That is why we used the definition of conditional probability instead. The conditional probability approach this leaves one term to compute: $\text{P}(N = \text{Gary})$ which we can compute using marginalization:

$$\begin{aligned}\mathrm{P}(N = \text{Gary}) &= \sum_{y \in \text{years}} P(B = y, N = \text{Gary}) \\ &\approx \sum_{y \in \text{years}} \frac{\text{count}(y, \text{Gary})}{k}\end{aligned}$$

Putting this all together we have:

$$\begin{aligned}\mathrm{P}(B = 1950 | N = \text{Gary}) &= \frac{\mathrm{P}(N = \text{Gary}, B = 1950)}{\mathrm{P}(N = \text{Gary})} \\ &\approx \frac{\left(\frac{\text{count}(1950, \text{Gary})}{k}\right)}{\left(\frac{\sum_{y \in \text{years}} \text{count}(y, \text{Gary})}{k}\right)} \\ &\approx \frac{\text{count}(1950, \text{Gary})}{\sum_{y \in \text{years}} \text{count}(y, \text{Gary})}\end{aligned}$$

More generally, for any name, we can compute the conditional probability mass function over birth year B:

$$\mathrm{P}(B = b | N = n) \approx \frac{\text{count}(b, n)}{\sum_{y \in \text{years}} \text{count}(y, n)}$$

From Birth Year to Age

Of course, if B is the birth year of a person, their age, A is approximately the current year minus B. This could be off by one if someone has a birth day later in the year, but we will ignore this small deviation for now. So for example, if we think that a person was born in 1988, since the current year is 2024 then their age is 2024 - 1988 = 36

Assumptions

This problem makes many assumptions which are worth highlighting. In fact, any time we make generalizations (especially about demographics) based on sparse information we should tread lightly. Here are the assumptions that I can think of:

1. This data only is accurate for names of people in the US. The probability of age given names could be very different in other countries.
2. The US census is not perfect. It does not capture all people who are resident in the US, and there are demographics which are underrepresented. This will also skew our results.

Names that Give Away Your Age

Some names have certain years where they were exceptionally popular. These names provide quite a lot of information about birth year. Let's look at some of the names with the highest max probability.

Medium Popularity (>10,000 people with the name)

Name	Age with max prob	Prob of most likely age
Katina	49	0.245
Marquita	38	0.233
Ashanti	19	0.250
Miley	13	0.250
Aria	7	0.247

High Popularity (>100,000 people with the name)

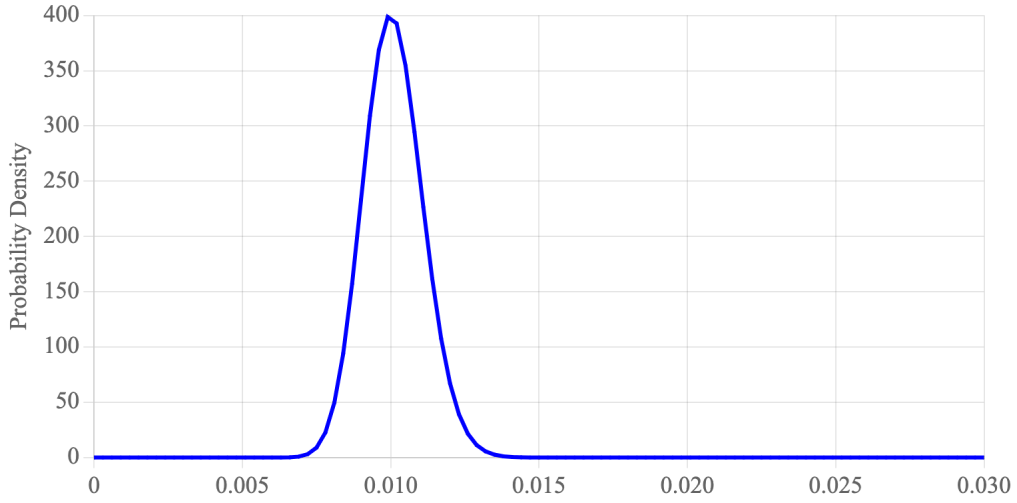
Name	Age with max prob	Prob of most likely age
Debbie	62	0.104

Name	Age with max prob	Prob of most likely age
Whitney	35	0.098
Chelsea	29	0.103
Aidan	18	0.098
Addison	14	0.112

A search for "Katina 1972" brought up this interesting article about a baby named Katina in a 1972 CBS Soap Opera. Marquita's popularity was likely from a 1983 toothpase add. Ashanti Douglas and Miley Cirus were popular singers in 2002 and 2008 respectively.

Futher Reading

Some names don't seem to have enough data to make good probability estimates. Can we quantify our uncertainty in such probability estimates? For example, if a name has only 10,000 entries in the database, of which only 100 were born in the year 1950, how confident are we that the true probability for 1950 is $\frac{100}{10000} = 0.01$? One way to express our uncertainty would be through a Beta Distribution. In this scenario we could represent our belief in the probability for 1950 as $X \sim \text{Beta}(a = 101, b = 9901)$ reflecting that we have seen 100 people born in 1950 and 9900 people who were not. We can plot that belief, zoomed into the range [0, 0.03]:



We can now ask questions such as, what is the probability that X is within 0.002 of 0.01?

$$\begin{aligned} &P(0.008 < X < 0.012) \\ &\quad = P(X < 0.012) - P(X < 0.008) \\ &\quad = F_X(0.012) - F_X(0.008) \\ &\quad = 0.966 - 0.013 \\ &\quad = 0.953 \end{aligned}$$

Semantically this leads to the claim that, after observing 100 births with a name in 1950, out of 10,000 births with that name over the whole dataset, there is a 95% chance that the probability of someone being born in 1950 is 0.010 ± 0.002.