Statistical Analysis as a Tool for Building your Family Tree

Bill Celmaster

May 30, 2017

Introduction

Genealogists frequently add ancestors to their family trees even when the evidence is inconclusive. As a result, many genealogies contain a mix of data ranging from highly reliable to highly speculative. Sometimes there are annotations indicating which family connections are particularly questionable, but in general there isn't a systematic approach for deriving or reporting the level of confidence for each entry of a family tree. There also aren't uniform criteria for adding entries to a tree. What I'd like to show in this paper, is that genealogical confidence levels are inherently statistical and can therefore be computed. Moreover, even when records are incomplete or mutually contradictory, statistical confidence can often be high enough to justify adding new entries or changing old entries.

The first section of the paper begins with an examination of several simple examples where I'll introduce statistical methods for drawing conclusions based on incomplete records. These methods will then be expanded to analyze a certain kind of situation that commonly occurs when attempting to establish 18th century Ashkenazi Jewish family connections based on archives from the early 19th century. Typically, a number of names have been derived from civil records. Then the question arises as to what can be inferred about the names of prior generations, noting that in the Ashkenazi tradition, children are usually named after dead ancestors. Also, how much confidence can we have in those inferences?

In the second section of the paper, I will review an actual family tree from the 18th century. Several new assumptions will be required for analyzing that tree. The family in question is that of Rabbi Jonatan of Drobin. According to rabbinic literature, Jonatan was the son of Aaron. However, after exploring many hundreds of vital records of Jonatan's family in Drobin and other towns, I had been struck by a curious absence of the name Aaron. This led me to wonder how likely it is that Jonatan was really the son of Aaron. A thorough answer to this question depends both on the name-analysis of Jonatan's descendants, and also on how much confidence to give to the rabbinic genealogies. For example, suppose we conjecture (as I will try to justify later) that in the absence of other non-rabbinic information, there is a 50% chance of the correctness of the 'Aaron Hypothesis' – that is, the hypothesis that Aaron is the father of Jonatan. Then a statistical analysis can be done of civil records, to see whether that should increase or decrease confidence in the Aaron Hypothesis. The analysis was done and on that basis, I concluded that it is extremely unlikely that Jonatan's father was named Aaron. Notice that if the analysis had started with 100% confidence in the rabbinic record, then that would have amounted to 100% confidence in the Aaron Hypothesis irrespective of any civil records. This illustrates the importance of estimating the statistical validity of anecdotal evidence (in this case, the rabbinic record) as well as the statistical significance of incomplete records.

Basic Statistical Methods for Genealogical Deductions

The key premise of any statistical analysis is this: when there are several alternatives consistent with all known information, then the alternatives are equally probable. The most challenging aspect of this premise, is the difficulty of identifying all the relevant information. Much of the time, that information is subsumed under the label of "general experience" or "context". Even so, careful consideration usually exposes the factors relevant for statistical analysis.

In the examples to follow, I will make up names and family trees that hopefully aren't those of real people or places. The examples will be artificially constrained to allow simple deductions but should suffice to show how statistical principles can be used effectively as genealogical research tools.

Example 1 – Distinguishing between two branches

First scenario – names repeat themselves

Your mother's father was named Noah Plockman and his father Jacob Plockman was born in 1920 in the shtetl of Przrtzrow (there's no such place so don't look for it). You want to find the name of Jacob's father but the publicly available civil records of Przrtzrow are only for the dates 1865 through 1912. Undeterred, you do a thorough examination of all the available Przrtzrow records, and you discover the birth in 1895 of a Noah Plockman, and the birth in 1902 of a Jacob Plockman. Although you find birth, marriage and death records of other Plockmans, no other males appear to have survived past 1912.

Based on this information, you would conclude that your great-great-grandfather (gggf) was named Noah and you have almost 100% confidence in this conclusion. The reasoning is thus: according to the Przrtzrow records, there are only two possible branches for your family – one descending from Noah (born in 1895) and the other descending from Jacob (born in 1902). So, your great-grandfather (ggf) was either the son of Noah, or the son of Jacob. However, your ggf's name was Jacob and according to Ashkenazi naming traditions, the son is not given the same name as his father. Furthermore, and almost as compelling, your ggf's son (i.e. your grandfather) was given the name Noah. That would be consistent with the Ashkenazi tradition of naming your son after your father (if your father is deceased).

Example 1: Scenario 1

Noah Plockman born 1895
Jacob Plockman born 1902

RELATIONSHIP	NAME	CONFIDENCE
Your gggf	Noah Plockman	~100%
Your ggf	Jacob Plockman (born 1920)	100%
Your grandfather	Noah Plockman	100%
Your mother	X	100%
You	Y	100%

Notice that the naming patterns lead to two separate ways of inferring the name of your gggf. That should significantly increase the confidence in your inference. Are there any factors that should reduce your confidence and if so, what would be the impact? Here is a possibility:

- Your gggf wasn't born in Przrtzrow even though he subsequently moved there. In that case, you wouldn't have found his birth record. His name probably wouldn't be Jacob so it's conceivable that your ggf could have been his son.
- The father or grandfather of your gggf was named Noah. Then your grandfather could have been named Noah after a more distant ancestor, and not after his grandfather.

Is this possibility likely? That depends on other information that you may, or may not, have obtained. In subsequent examples, I will show how to assign probabilities to various possibilities. For now, it is sufficient to note that it is rarely possible to infer ancestral names with 100% confidence.

Second scenario – unique names

The second scenario is the same as the first, except that your grandfather was named Abraham and your ggf was named Moshe. All other facts are the same as in the first scenario.

Example 1: Scenario 2

Noah Plockman born 1895
Jacob Plockman born 1902

RELATIONSHIP	NAME	CONFIDENCE
Your gggf	See below (Noah)	See below (75%)
Your ggf	Moshe Plockman (born 1920)	100%
Your grandfather	Abraham Plockman	100%
Your mother	X	100%
You	Y	100%

Now, in the absence of any other information, there is a 50/50 chance that Moshe's father was Noah, and a 50/50 chance that Moshe's father was Jacob. However, there is other information – namely that Jacob was 18 years old and Noah was 25 years old in 1920, when Moshe was born. Assume that in the town of Przrtzrow, 75% of men have their first child after they are 19 years old (this assumption could be tested by examining all birth records from Przrtzrow). Then you would conclude that there is a probability of 0.75 (75%) that Noah was the father of Moshe, and your family tree would show that your gggf was named Noah – with a confidence of 75%.

Example 2 – Missing data and Bayes' theorem

In Example 1, it was assumed that the Przrtzrow civil records completely covered the period of 1865 through 1912. However, there are many cases where 19th century Polish civil records have gaps. Sometimes several years are missing, or certain kinds of records are missing from a specific year. In particular, the earliest Polish Jewish civil records, especially those dating from about 1808 to the early 1820's, are often spotty. Going even further back in time it's possible to obtain, from post-1808 records, information about people born in the 1700's. However, many other people born in the 1700's died before 1808 and are entirely missing from any existing Jewish civil records. Therefore, the pre-1800 data has many gaps.

This kind of situation can be illustrated by examples where you are trying to deduce the name of a distant ancestor about which little is known except for the names of some of his descendants. In fact,

in the examples to be considered in the remainder of this paper, only first names rather than surnames[1], will be assumed known. Surnames weren't widely used in Polish Jewish records until the early 1820's. Prior to that time, people were known by their patronymic names – A, son of (or daughter of) B.

First scenario – grandsons only

In the first scenario to be considered, the ancestor of interest is known to have had three grandsons, one of whom was named Noah. How likely is it that the ancestor was also named Noah? In order to answer this question without introducing too much computational complexity, I'll make a number of artificial (but not unreasonable) assumptions about information that may be known, and about rules for naming children. These assumptions are most easily explained with reference to the table below. In this table, the ancestor of interest is labelled "Gf" (grandfather). His son, Isaac, had three sons but only two of their names are known. Nothing is known about the order of birth of Isaac's sons (in the language of statistics, we say that all birth-orders are equally probable). Assume, also, that all grandfathers and great-grandfathers (Ggf) died before the birth of any of Isaac's sons. (In practice, it's highly unlikely that we would have this kind of specific information. This is one of the artificial assumptions mentioned earlier.)

Example 2 – Scenario 1

Ggf1	wife of Ggf1	Ggf2	wife of Ggf2	Ggf3	wife of Ggf3	Ggf4	wife of Ggf4
Gf (? -1800)		wife of Gf		Jacob (? -1800)		wife of Jacob	
Isaac (1795-1855)				wife			

Noah (born after 1810)	Jacob (born after 1810)	S (born after 1810)

For this example, assume the naming-rule that the oldest son is always named after the paternal grandfather if deceased, the next oldest son is named after the maternal grandfather if deceased, and later sons are named after great-grandfathers. (In general, that naming-rule may be slightly too restrictive.)

Based on the scenario described, it would appear that the name of Isaac's father has a 50/50 chance of having been Noah. However, it is worthwhile analyzing the situation carefully so that it can be generalized for more complex scenarios. One extremely valuable tool for these kinds of analyses, is Bayes' theorem (in conjunction with what is known as *The Rule of Total Probability*) (Freund, 1988).

If B_1, B_2,…, and B_k are mutually exclusive events of which one must occur, then

$$P(B_i|A) = \frac{P(B_i)\cdot P(A|B_i)}{\sum_{j=1}^{k} P(B_j)\cdot P(A|B_j)} \qquad (1)$$

[1] Surnames, when available, provide extra information that can significantly alter the statistical results.

for I = 1, 2, … or k and where $P(B|A)$ is the conditional probability of B relative to A (i.e., the probability of B given A).

In the scenario above, the two "mutually exclusive events" are

B_1: Isaac's father was Noah
B_2: Isaac's father was not Noah

The probability of B_1 (i.e., the probability that Isaac's father was Noah) in the absence of any other information, is written as $P(B_1)$. For notational clarity, define $p(Noah) \equiv P(B_1)$. One way to estimate $p(Noah)$ would be to tabulate all first names recorded in Przrtzrow, and then see how often the name Noah appears. Alternatively, if that information isn't available for the town of Przrtzrow, it might be available for some other region nearby. Since Noah is quite a rare name in most communities, it's likely that $p(Noah)$ is less than 0.01. Note that $P(B_2) = 1 - P(B_1) = 1 - p(Noah)$.

The Bayes variable A, in this scenario, represents all known information about Isaac's sons. Then $P(A|B_1)$ is the probability of that information, given that Isaac's father was named Noah. There are two steps in obtaining that probability. First is the probability that Isaac would have named two of his sons Noah and Jacob. Second is the *selection probability*, that is the probability that the two sons whose names happen to be known, are Noah and Jacob (we might instead only have known the names of Noah and the third son, or of Jacob and the third son). It is easy to see that the selection probability is 1/3. According to the assumed naming traditions, if Isaac's father was named Noah and his father-in-law named Jacob, then Isaac would have named his first son Noah and his second son Jacob. So, the probability of naming two sons Noah and Jacob, would be exactly 1. This argument shows that $P(A|B_1) = 1 * \frac{1}{3}$.

Next, compute $P(A|B_2)$, the probability of the information known about Isaac's sons, given that Isaac's father was <u>not</u> named Noah. As before, there are two steps. Isaac would have named his first son after his father, so that son's name would <u>not</u> be Noah. Designate that name as N-N (for 'not Noah'). Isaac's second son would be named Jacob. The probability of naming his third son Noah is approximately the same as $p(Noah)$ (the probability, in the absence of any other information, of naming someone Noah). (This is only an approximation because the third son cannot have the names J or N-N, so that slightly increases the probability of having the name Noah. The correction is described mathematically as $o\left(\left(p(Noah)\right)^2\right)$ or "second-order".) Again, the selection probability is 1/3, since the sons whose names happen to be known, are Jacob and the third son. Thus, this shows that $P(A|B_2) = p(Noah) * \frac{1}{3}$.

Now Bayes' theorem can be applied to finding $P(B_1|A)$, the probability that Isaac's father was named Noah, given the information known about Isaac's sons.

$$P(B_1|A) = \frac{P(B_1) \cdot P(A|B_1)}{P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2)}$$

$$\cong \frac{p(Noah) * \frac{1}{3}}{p(Noah) * \frac{1}{3} + \left(1 - p(Noah)\right) * p(Noah) * \frac{1}{3}} = \frac{1}{2 - p(Noah)} \cong 0.5$$

That result agrees, as it should, with our initial conclusion. However, the method generalizes easily to more complicated examples, as will next be explored in the second scenario.

Second scenario – grandsons and great-grandsons

We continue to study the family of Isaac introduced above, but with extra information about Isaac's grandsons as shown below. Isaac's son Noah (to be referenced as Noah #1) had two sons, one of whose name was Meyer and the other whose name isn't known to us. Again, nothing is known about the birth-order of those two grandsons. Assume, also, that Noah #1's father-in-law died before the birth of any of the sons of Noah #1.

Example 2 – Scenario 2

Ggf1	wife of Ggf1	Ggf2	wife of Ggf2	Ggf3	wife of Ggf3	Ggf4	wife of Ggf4
Gf (? -1800)		wife of Gf		Jacob (? -1800)		wife of Jacob	
Isaac (1795-1855)				wife			

Noah #1 (born after 1810)		Jacob (born after 1810)		S (born after 1810)	
Meyer (1830 - ?)	Gs1	Gs2	Gs3	Gs4	Gs5

We proceed by continuing the previous analysis, now accounting for our knowledge of Isaac's grandsons. First consider the case where Isaac's father was named Noah. Since Isaac was alive at the time of birth of his grandson Meyer, Noah #1 would have named his first-born after his (Noah #1's) father-in-law and would have named his second-born Jacob (Noah #1 can't have named his second son Noah since by tradition, sons don't carry the same name as their fathers). The only son we know of, is named Meyer. Therefore, Meyer must be the name of the father-in-law of Noah #1. The probability for this is $p(\ Meyer)$. As before, we must multiply by the selection probability – the probability that the son whose name we know, is Meyer. This probability is ½. In summary, given that Isaac's father was Noah, then the probability of knowing that Noah #1 had a son named Meyer, is $p(\ Meyer) * \frac{1}{2}$. We can now calculate $P(A'|B_1)$ where "*A'*" denotes everything known about Isaac's sons and grandsons.

$$P(A'|B_1) = P(Knowledge\ about\ sons\ of\ Isaac|B_1) * P(Knowledge\ about\ sons\ of\ Noah\#1|B_1) = \frac{1}{3} * \left(p(\ Meyer) * \frac{1}{2}\right)$$

This calculation[2] has used the result previously obtained in the first scenario above where we considered only the sons of Isaac.

Now consider the case where Isaac's father is <u>not</u> Noah (N-N). As before, Noah #1 would have named his first son after his (Noah #1's) father-in-law. His second son would have been named N-N. The name

[2] As before, second-order terms are ignored.

Meyer is consistent with either the first or second son. In either case (hence a factor of 2), the probability is $p(\ Meyer)$ and then accounting for the selection probability, we have

$$P(Knowledge\ about\ sons\ of\ Noah\#1|B_2) = 2 * p(Meyer) * \frac{1}{2}$$

so

$$P(A'|B_2) = P(Knowledge\ about\ sons\ of\ Isaac|B_2) * P(Knowledge\ about\ sons\ of\ Noah\#1|B_2) = p(Noah) * \frac{1}{3} * \left(2 * p(\ Meyer) * \frac{1}{2}\right)$$

Finally, we can apply Bayes' theorem.

$$P(B_1|A') = \frac{P(\ B_1) \cdot P(A'|B_1)}{P(B_1) \cdot P(A'|B_1) + P(B_2) \cdot P(A'|B_2)}$$

$$= \frac{p(Noah) * \frac{1}{3} * \left(p(\ Meyer) * \frac{1}{2}\right)}{p(Noah) * \frac{1}{3} * \left(p(\ Meyer) * \frac{1}{2}\right) + \left(1 - p(\ Noah)\right) * p(Noah) * \frac{1}{3} * \left(2 * p(\ Meyer) * \frac{1}{2}\right)}$$

$$= \frac{1}{3 - 2 * p(\ Noah)} \cong 0.33$$

Example 3 – Unknown number of siblings and the Poisson distribution

In the examples above, the number was known of children in each family, even if not all their names were known. Often, especially when there are gaps in the records, we may not know about "missing" siblings. In that case, progress can still be made by guessing how many sons and how many daughters are in each family. This 'guess' can be based on family-size statistics, if such data has been collected[3].

Obviously, there are statistical variations in family-size. In the absence of good data, it is useful to have a probabilistic model that can be applied. In my research, I have employed a model in which the number of sons follows a Poisson distribution with an expected value of N/5, where N is the number of child-bearing years of the parents. The model derives from the plausible assumption that the years of childbirth follow a Poisson distribution (Derrida, Manrubia, & Zanette, 1999), at least during the best part of child-bearing years (approximately 25 years). Based on a cursory examination of family sizes where civil records are complete, the average number of children is about 10, divided equally into sons and daughters (hence 5 each). In situations where parents are known to have had a shorter than average child-bearing period (owing to early death or late marriage of one of the parents), the expected value should be pro-rated accordingly[4].

Let us apply this model to a variation of the first scenario of Example 2. Instead of 3 sons, suppose that Isaac had an unknown number of sons, of which we know only the names Noah and Jacob. Again, birth-

[3] I'm not aware of any quantitative studies on this topic.

[4] This assumption, as well as others, could potentially be studied empirically by surveying traditional communities in Israel and elsewhere – to obtain contemporary data regarding names, birth and death-dates of children and the people after whom they are named.

order isn't known, and again it is assumed that all grandparents are deceased prior to the first birth (generally, it's rare that we would know this without also knowing the names of the grandparents).

Example 3

Ggf1	wife of Ggf1	Ggf2	wife of Ggf2	Ggf3	wife of Ggf3	Ggf4	wife of Ggf4
Gf (? -1800)		wife of Gf		Jacob (? -1800)		wife of Jacob	
Isaac (1795-1855)				wife			

Noah	Jacob	S_3	…	S_N

There are *N* sons but the value of *N* is unknown. The values $P(A|B_i)$ are then obtained by calculating the partial probabilities $P(\ A, N|B_i)$ corresponding to exactly N sons, then multiplying by the Poisson probability that there are exactly N sons (note that $N\ \geq 2$), and then taking the sum over all N.

$$P(A|B_i) = \sum_{N=2}^{\infty} P(A, N|B_i) * f_{Poisson}\,(N, mean) \qquad (2)$$

The value of *mean* is the average number of sons expected. This equation is based on the observation that each of the partial probabilities $P(A, N|B_i)$ correspond to an independent event (e.g. if Isaac has 3 sons, then he doesn't have 4 sons).

In the first scenario of Example 2, we computed the value for 3 sons, $P(B_1|A, 3) = 0.5$. Generalize this to other values of N. Begin by computing $P(A, N|B_1)$ – the probability that two out of Isaac's N sons are named Noah and Jacob – given that Isaac's father was named Noah. As before, the eldest two sons will be named Noah and Jacob. Again, there is a selection factor, representing the probability of knowing the names Isaac and Jacob, out of a total of N sons. Said differently, the calculation of conditional probability needs to include a factor for the probability of selecting two objects out of N objects. The value of this selection factor[5] is $\frac{1}{\binom{N}{2}}$ (it is easy to see that in the case $N = 3$ the selection factor is $\frac{1}{3}$ as computed previously) so $P(A, N|B_1) = \frac{1}{\binom{N}{2}}$. Then go on to compute $P(A, N|B_2)$, for the case when Isaac's father is N-N (not Noah). The eldest son is named N-N and the next son is named Jacob. We know that one of Isaac's later sons is named Noah. If the third son is named Noah, then he is named after a relative named Noah, whose name had a probability $p(\ Noah)$ of being Noah. There is a selection probability of $\frac{1}{\binom{N}{2}}$ that the two sons we know of, were the second and third sons. Or alternatively, the 4th son could have been named Noah, again with the same probability $p(\ Noah) * \frac{1}{\binom{N}{2}}$. This argument can be repeated until son number *N*. Then each of those probabilities must be added. The result is $P(A, N|B_2) = \frac{1}{\binom{N}{2}} * p(Noah) * (N - 2)$.

[5] We follow the notation $\binom{m}{n} \equiv \frac{m!}{n!(m-n)!}$

Next, calculate sums over the Poisson distribution using Equation (2). These results are easily computed within Microsoft Excel, using the Visual Basic feature and the built-in Poisson function. The result, for a mean of 5 sons, is

$$P(A|B_1) \cong 0.20$$
$$P(A|B_2) \cong 0.25 * p(Noah)$$

Then, from Bayes' theorem (Equation (1)),

$$P(B_1|A) = \frac{P(B_1) \cdot P(A|B_1)}{P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2)}$$

$$\cong \frac{p(Noah) * 0.20}{p(Noah) * 0.20 + (1 - p(Noah)) * p(Noah) * 0.25} = \frac{0.20}{0.45 - 0.25 * p(Noah)} \cong 0.44$$

In summary, if on average, families have 5 sons, then the information about Isaac's family provides 44% confidence that Isaac's father was named Noah. For comparison, if it were assumed that families have an average of 3 sons, then the result would be a 65% confidence that Isaac's father was named Noah. (The reason the value is higher than the 50% calculated when assuming exactly 3 sons, is that the Poisson distribution has a significant contribution for the case of 2 sons, where there would be almost 100% certainty that Noah was Isaac's father.)

Beyond Basic Statistical Methods: Real Genealogies

In the previous section, certain simplifications were made for purposes of clarity. In particular, the examples had only a small number of known ancestors and descendants, and the naming rules were more restrictive than the conventions that were traditionally used in Ashkenazi families. In this section, I examine the genealogy of an 18th and 19th century rabbinical family from the shtetl of Drobin in Poland, and apply more realistic naming rules. I then conclude with some general remarks on prospects for statistical research of the kind discussed in this paper.

The family of Rabbi Jonatan of Drobin

The genealogy studied here, is that of Rabbi Jonatan Eybeschuetz of Drobin. Jonatan's wife – according to vital records from Drobin – was named Dwojra and her father was named Jacob. Jonatan – known in the vital records as Jonas – lived from approximately 1720 to approximately 1768. As it happens, there are several family histories (typically found in the prefaces to 19th century rabbinical works) that mention Rabbi Jonatan. Those references claim that Jonatan was the son of Aaron Fird, brother of the famous rabbi, Jonatan Eybeschuetz of Prague.[6] However, such histories and deductions have occasionally been known to be inaccurate. In this case, how reliable is the conclusion that Aaron Fird

[6] Most of the rabbinic information, has been provided to me by Dr. Heshel Teitelbaum. An example (amongst many) can be found in (Blokherovitsh, 1939) page 16

ט) חתן הרה"נ מוח"ר יונתן אייבעשיץ אבר"ק דרובנין.

י) בהרה"ג 'חמופלנ מוה"ר אהרן פירר אייבעשיץ אבד"ק שטעדבורז

really was the father of Jonas (from now on, I will use the name Jonas to distinguish the rabbi from Drobin, from the better-known rabbi from Prague)? Aaron Ford is mentioned in several independent sources as being the brother of Rabbi Jonatan Eybeschuetz of Prague. However, the documents mentioning Jonas and his relationship to Aaron, were written more than 150 years after Aaron's death. It seems plausible that Jonas was known to be a nephew of Rabbi Jonatan Eybeschuetz of Prague and that later rabbinical historians – knowing of no other siblings of Jonatan and Aaron – concluded that Aaron must be the father of Jonas. Based on this reasoning, I have somewhat arbitrarily proposed a 50% probability, based solely on the rabbinic writings, that the claim is valid (by which I mean this: whether or not it turns out that Aaron Ford really was the father of Jonas, that conclusion cannot be drawn with any more than 50% confidence from the aforementioned documents). The question then investigated using statistical analysis is "what is the probability, given information derived from both civil records and rabbinic writings – that the name of Jonas's father was Aaron?". The methods used are similar to those described previously. In particular, Bayes' theorem (equation (1)) is applied, with $p(Aaron) = 0.5$.[7]

Continuing with anecdotal information relevant to Jonas, there are sources stating that Aaron Ford's father-in-law was named Meyer. That information is used in the analysis by setting a 100% probability that Meyer was Jonas's father's father-in-law provided that Aaron was Jonas's father. Separately, based on anecdotal information including some gravestone inscriptions, we also set a 100% probability that Jonas had a grandfather named Nuta and two great-grandfathers named Eyzyk and Mosze[8]. All of the aforementioned information, including the names of Jonas' wife and father-in-law is summarized in the following chart. The X_i are placeholders for ancestors whose names we don't know.

Documented Ancestors of Jonas and Dwojra (assuming Aaron is the father of Jonas)

Eyzyk + wife	Mosze + wife	X_3 + wife	X_4 + wife	X_5 + wife	X_6 + wife	X_7 + wife	X_8 + wife
Nuta	wife of Nuta	Meyer	wife of Meyer	X_1	wife of X_1	X_2	wife of X_2
	Aaron		wife of Aaron	Jacob		wife of Jacob	
			Jonas	Dwojra			

Jonas' descendants are recorded in vital records from Drobin and elsewhere, dating from approximately 1808 onwards. Here is what I currently know, as of May 2017.

- Jonas (~1720 – ~1768) + Dwojra
  - Jacob (1743 –1811) + wife
    - Nachman-Wulf (1778 –?)
    - Meyer-Nuta (1791 – before 1866)
  - Daughter + Jacob-Lipman (1743 –?) (son of Haim)
    - Israel (? – ?)

[7] Note the importance of taking account, as we have done, of the anecdotal information about Aaron. Otherwise, it would be necessary to make probabilistic analyses based on the assumed frequency of the first name Aaron in the general Jewish population. That number can by hypothesized to be about 2%, from a study of Ellis Island records as published by Yannay Spitzer in 2012 at https://yannayspitzer.net/2012/07/24/most-common-jewish-names/ with similar results at http://www.jewishgen.org/databases/USA/1890nyNames.htm.

[8] Except for the name Nuta, there isn't a compelling reason to set 100% probability for the other names of Jonas's ancestors. However, those names may be reasonably reliable since they have to do with the ancestry of either Aaron Ford or Rabbi Jonatan Eybeschuetz of Prague – both considerably better known than Jonas.

        - Jonas (1774 –?)
    - Abraham (? – before 1810) + (1st wife) Laja (? – ~1770)
        - Leyzor (1749 – 1830)
        - Jonas (~1769 – 1821)
    - Abraham[9] (as above) + (2nd wife) Szajndel (? – ?)
        - Mosze-Haim (1770 – 1833)
        - Chaskel (1780 – 1831)
    - Fajga (1743 – 1823) + Lewek (1743 – 1819) (son of Jacob)
        - Jacob (1771 –?)
        - Jonas (1799 – 1863)
    - Eyzyk (1768 – 1827) + Hudes (1780 – 1826) (daughter of Haim Meyer (son of Jacob))
        - Jonas (1809 – 1810)
        - Hersz (1811 –?)
        - Jacob (1817 –?)
        - Mathias (1819 –?)
        - Eliasz-Lewek (1823 –?)
    - Unknown daughter + Tanchum (? – ?)
        - Abraham (1778 – 1847)
        - Hertz (1779 – bef. 1843)
        - Lemel (1784 - ?)
    - Mosze (? – ?) + wife (*Assume equal probability that Jonas's child is either Mosze or Mosze's wife)*
    - Hertz (aka Naftali Hertz) (1736 – ?) + Hinda
        - Szmuel (1754 – 1840)

This family tree is then analyzed based on assumptions similar to those described in the previous section. The assumptions are listed here for reference. [10]

A1. Names of sons are given only for deceased ancestors (generations 1 and earlier).

A2. When sons are given single names, assume that names are chosen in order so that near-generations are exhausted before the next ancestor-generation is used – but that within a generation, names are chosen at random. (This assumption is different, but more realistic, than the naming-rule used previously.)

A3. The name Mosze-Haim above (son of Abraham) is treated as the single name Mosze (noting that the name Haim is occasionally added after birth).

A4. The two known sons of Jacob (born in 1743) have double names, namely Nachman-Wulf and Meyer-Nuta. Assume the following naming rules for the children of Jacob:

a. Names are chosen in order so that near generations are exhausted before the next ancestor generation is used – but that within a generation, names are chosen at random.

[9] There is some question about whether this Abraham (husband of Szajndel) was the same as the husband of Laja. If not, then Szajndel is likely to have been the daughter of Jonas and Dwojra and the statistical conclusions aren't much altered. (In fact, it could even turn out that Laja, and not Abraham, was the child of Jonas and Dwojra.)

[10] Although all assumptions are stated in terms of sons, precisely the same assumptions could be made for daughters.

b. When it's possible to form a double name from the names of Jacob's ancestors within a single generation, then Jacob's son is given that double name, otherwise the son is given a single name.

A5. The number of sons follows a Poisson distribution with an expected value of 5 for Jonas' sons as well as for the sons of each of Jonas' children – with the exception of Abraham. Abraham has two wives and for each of them, we take the expected number of sons to be 4.

A6. Jonas was close to being Laja's last son (note that Rabbi Jonatan of Drobin (JD) died a year earlier), but all other sons are named in the *standard* order (the assumption being that all other grandparents, etc. all predeceased the birthdates of Laja's other sons). This assumption is implemented by dividing Laja's child-bearing years into period $P(1)$ of 18 years prior to the death of JD and period $P(2)$ of 2 years subsequent to the death of JD. Poisson distributions are applied to both (where the mean is the number of years divided by 5). Jonas was born in that second period.

A7. Ancestral names – in the absence of genealogical information to the contrary – are all equally probable and are equal 0.02. That is, *p(Jacob) = p(Abraham) = …* = 0.02.

A8. Observations (finding civil records of confirmed sons) are a random selection of the sons[11]. (As mentioned before, some records may be missing from city archives, owing to infant deaths, emigration, etc.) More precisely, for each descendant-generation, the known names have been selected entirely randomly from the collection of descendants of that generation – irrespective of chronology. (This assumption is what led to the selection probabilities used in the previous section.)

A9. All ancestors are deceased prior to the births of any of the sons.

The result of applying statistical analysis to Jonas' family tree, using the above 10 assumptions is:

$$P(Aaron|Jonas\ tree) = \frac{P(\ Aaron) \cdot P(Jonas\ tree|Aaron)}{P(\ Aaron) \cdot P(Jonas\ tree|Aaron)\ \ + P(\ Not\ Aaron) \cdot P(Jonas\ tree|Not\ Aaron)} \cong \frac{p(\ Aaron) * 0.36}{p(\ Aaron) * 0.36 + \big(1 - p(\ Aaron)\big) * 7.7}$$

In the above equation, $p(\ Aaron)$ is the *a priori* probability (the probability in the absence of any other family tree information, based on our confidence in the rabbinical genealogies) that Aaron is the father of Jonas. The terms representing conditional probabilities are to be interpreted as usual. For example, $P(Aaron|Jonas\ tree)$ is the probability, given the information about Jonas's family tree, that Aaron is the name of Jonas' father.

If we take $p(\ Aaron) = 0.5$, then we conclude that $P(Aaron|Jonas\ tree) = 0.045$,i.e. a 5% likelihood that Aaron is the name of Jonas's father. On the other hand, if we give much higher credence

[11] These assumptions are reasonably valid when studying Polish Jewish records of generations born prior to 1808, but whose names appear in records later than 1808. For case-studies concerning families whose members of interest were born post-1808, birth-order tends to be much better known.

to the rabbinical reports, then as expected, the likelihood increases. For example, with $p(\ Aaron) = 0.8$, we would conclude that $P(Aaron|Jonas\ tree) = 0.16$.

Conclusions

In this paper, I set out to develop a systematic statistical approach for deducing unknown ancestral names from existing genealogical records. The basic idea is that in Ashkenazi families, children are named after deceased ancestors. Thus, if a certain given name appears several times amongst the descendants of person X (whose name might not appear in any record), it may be reasonable to conclude that this is also the name of person X. Or conversely, if amongst many descendants of person X, a certain name never appears, then it may be reasonable to conclude that this is not the name of person X. Details depend on precisely what is known about the family being investigated, and on precisely what traditions were used for naming children. What also matters, are the assumptions we make about the information we have. Are some records missing and if so, how many? Are some names more commonly encountered in the general population than other names? Is there any anecdotal information that might suggest a particular name for person X and if so, how credible is that information?

What we have seen, is that one can often reasonably deduce the names of some ancestors – but with a confidence level that may be less than 100%. In point of fact, many family trees have entries that were added with less than 100% confidence. The reasons can vary, but generally speaking it is a common consequence of the fact that civil records are usually incomplete. I have proposed that genealogists should try to annotate entries for which they lack 100% confidence. Some effort should be made, perhaps using the methods elaborated in this paper, to quantify the confidence of those entries.

The value and reliability of the statistical approach for 18th through 20th century Jewish genealogy, could be improved by systematic research on various kinds of statistics pertaining to records from that time period. For example, it should be possible to examine marriage records and to develop a statistical distribution of the ages of when men and women wed. This information could be compared from town to town and from era to era. As was shown earlier, that kind of information could be useful in drawing conclusions about whether someone was an ancestor (e.g. if Person X was 15 years old at the time your grandfather was born, then your grandfather was extremely unlikely to be the son of Person X).

These statistical methods were applied both to some simple examples analyzed in the first section of this paper, and also to the family of Rabbi Jonatan of Drobin (JD). The key result of the analysis of JD's family, is that the name of JD's father is unlikely to be Aaron – despite some rabbinical genealogies that suggest otherwise. That result then opens the door to research on some alternative scenarios for the ancestry of JD.

In this paper, I did not show the calculations required for the analysis of JD's family tree. This was because the analysis involves far more combinatoric complexity than what was previously encountered when we examined simple examples. The increased complexity is a consequence of two things: (1) JD's family tree is much larger than that of any of the previous examples, and (2) assumptions A1 - A9 lead to many more computational branches than the assumptions previously presented (even though there are many similarities between those sets of assumptions). One concern that became apparent to me when doing the JD calculation, is that the entire approach can be extremely error prone owing to the massive amount of combinatoric formulas and the ensuing arithmetic. I was able to manage some of

this by writing Simple Basic programs to be used within Excel.[12] However, if statistical methods are to be more accessible to genealogists, it would be better to employ computer programs that simply count all the different family permutations consistent with the known data and the listed assumptions. Although this approach would be too onerous to be executed by hand, or even by Excel, it should be well within the capability of desktop computers.

References

Blokherovitsh, Y. A. (1939). *Zikne Mahaneh Yehudah.* Peitrikow: see hebrewbooks.org.

Derrida, B., Manrubia, S. C., & Zanette, &. D. (1999). Statistical Properties of Genealogical Trees. *Physical Review Letters 82, 9*, 1987-90.

Freund, J. E. (1988). *Modern Elementary Statistics, Seventh Edition.* Englewood Cliffs, New Jersey 07632: Prentice-Hall.

[12] I have all of the calculations and programs available to share for anyone who wishes to look at them.