Challenge: Modeling

Caden Hewlett

2024-02-25

Question:

You are collecting lithographs from a series. Each one has a unique serial number, but contrary to standard practice, the artist did not specify the total number of copies in circulation.

Here is the collection so far:

```
collection = c(1, 3, 15)
```

Part 1: Designing a Model

We make the following assumptions:

1.) Serial numbers are natural numbers, i.e., there is no serial number 0.

2.) Serial numbers are in a sequential order, i.e. the existence of lithograph number 15 implies the existence of lithograph number 14, and all others below it (by induction.)

Design a Bayesian model to infer the total number of copies in circulation. Motivate all choices you make.

Introduction

We'll begin with a brief discussion of nomenclature. Let the random variable Ψ be the total number of copies in circulation. Let Y be the serial number of the lithographs in our collection. Directly, we have observed $Y_{1:3} = \{1, 3, 15\}$.

We wish to design a model to tell us $P(\Psi = \psi \mid Y_{1:3} = \{1, 3, 15\})$. However, in order to do this, we have a bit of work to do.

Firstly, we need some likelihood $P(Y = y \mid \Psi = \psi)$, we need a prior $p_\Psi(\psi)$ on Ψ. If we really wanted to be thorough, we could also derive $P_Y(y)$ for normalization constant Z.

Discussion of Distributions

Let's take some time to discuss the distribution of the total number of serialized lithographs in circulation by this particular artist. Since the total number of lithographs in circulation is some positive integer, we can consider the set of Probability Mass Functions for our distribution. Our collection of serial numbers so far is $\{1, 3, 15\}$

Let F be the set of all valid PMFs we can choose from, i.e.,

$$F = \left\{ f : \mathbb{Z}^+ \to [0,1] \;\text{ s.t. } \sum_{x \in \text{supp}(f)} f(x) = 1 \right\}$$

.

From our data, we can begin to place some restrictions on the outcome. Since we have Serial 1, assumed to be the first, we know the PMFs we can choose from (which I will call G) has the following restriction via infimum, and hence is likely a subset of F. Let $\mathcal{S}(g) = \text{supp}(g)$ for any $g \in G$. This nomenclature is used for clarity only.

$$\forall g \in G,\, \inf\big(\mathcal{S}(g)\big) = 1, \;\; \therefore \; G \subset F$$

Directly, we know that G is a strict subset of F because $\exists f \in F$ s.t. $\text{supp}(f) \neq 1$ (for example the discrete uniform distribution from -3 to 0.)

Further, since we have Serial 15 in our collection, we know that:

$$\forall g \in G,\, \sup\big(\mathcal{S}(g)\big) \geq 15$$

From this, I suggest the following additional restriction:

$$\forall g \in G,\, \forall \psi \in \Psi,\, (\psi \leq 14) \implies g(\psi) = 0$$

In other words, since we have serial number 15, it is *impossible* for the total number in circulation (i.e. highest serial number) to be less than 15. Therefore, a sensible prior for this setting should place zero mass on all values below 15, and, in addition, should place some mass on all values *above* 15.

Finally, due to the highly expensive and collectible nature of lithographs, to increase the usefulness of the prior I decided to add one more additional consideration; specifically, that our prior belief on the *true* number of lithographs in circulation does not stray far from the theorized highest value.

With all of this in mind, I chose to use the **shifted geometric distribution** as the prior. It is given by:

$$p_\Psi(\psi) = \mathbb{P}(\Psi = \psi) = p(1-p)^{\psi - s - 1}$$

We will consider the situation in which $s = 15$. However, were we to observe some external information (i.e. that our serial number 15 is a fake,) this value would be subject to change.

It should be noted that the geometric probability p is not fixed. We will consider a small set of positively-skewed beta distribution for this probability value. In other words,

$$p \sim \text{beta}(\alpha, \beta), \;\text{ where } \alpha \in \{2,3\}, \;\text{ and } \beta \in \{14,15,16\}$$

Trivially,

$$\alpha \sim \text{unif}\big(\{2,3\}\big) \;\text{ and }\; \beta \sim \text{unif}\big(\{9,10,11\}\big)$$

Hence, we have the prior as the shifted distribution:

$$\psi \mid p \sim \text{geomShift}(p, s = 15)$$

Or, in other words,

$$\mathbb{P}(\Psi = \psi \mid p) = p(1-p)^{\psi - 15 - 1}$$

We've done a lot of work discussing the random variable discussing the number of possible lithographs. I don't even think it's really a "prior" anymore.

Let y_i be the i-th lithograph in our collection.

For our data, we have $i \in [1, 3]$ and $\{y_1, y_2, y_3\} = \{1, 3, 15\}$. I will propose simple a discrete uniform likelihood on y_i given ψ.

I will argue in defense of this likelihood choice. It only relies on the assumption that all lithographs in the collection are independent of one another. For example, if I have serial number 1, this does not mean it is more likely for me to have serial number 3, 15, or k for all $k \in \mathbb{N}$.

Let's formalize this proposition a bit. Let $\{\vec{k}_i\}_{i=1}^n$ and $\{\vec{\ell}_j\}_{j=1}^m$ be sets representing the serial numbers of two different collections of lithographs of size m.

Since serial numbers are unique by definition, we have that:

$$\forall n, m \in [1, \psi - 1], (n + m \leq \psi) \implies \big(\{\vec{k}_i\}_{i=1}^n \cap \{\vec{\ell}_j\}_{j=1}^m\big) = \emptyset \tag{1}$$

In other words, any non-overlapping collections of lithographs will have a disjoint intersection of serial numbers. Simply put; no two valid collections can share serial numbers.

This conjecture is important, because it remains valid we fix $n = m = 1$. This means that the intersection of collections of size 1 will remain disjoint!

Let c_i be the i-th possible collection of size 1. Since order doesn't matter, $i \in [1, \psi]$

By the implication above,

$$\forall i, j \in [1, \psi] \text{ s.t. } i \neq j, (c_i \cap c_j) = \emptyset \tag{2}$$

This implies that all collections of size 1 are independent, therefore all c values are independent and identically distributed.

In other words,

$$\forall i, j \in [1, \psi] \text{ s.t. } i \neq j, (c_i \perp c_j) \tag{3}$$

Finally, all $\{c_1, c_2, \dots c_\psi\}$ are equally likely by nature of being *iid*. Hence, the contents of these arbitrary collections of size 1 are also independent.

By this fact, letting $y_i \in [1, \psi]$ be the i-th serial number, we know that that:

$$\forall y, j \in [1, \psi], \ \mathbb{P}(y \in c_j) = \frac{1}{\psi}$$

So, since all y_i are *iid* with probability $1/\psi$, the specific serial number we observe in a given c_i is given by:

$$y_i \mid \psi \sim \text{categorical}(\{1, 2, \dots, \psi\}, \{1/\psi, 1/\psi \dots, 1/\psi\}) \equiv \text{unif}(\{1, 2, \dots, \psi\})$$

Hence, we arrive at the discrete uniform distribution for serial numbers, as required.

Detailing the Full Model

With all of this in mind, our full model is as follows.

We note that a simplified version of this model fixes $\{\alpha, \beta\} = \{2, 10\}$

$$\begin{aligned}
\alpha &\sim \text{unif}\big(\{2, 3\}\big) \\
\beta &\sim \text{unif}\big(\{9, 10, 11\}\big) \\
p \mid (\alpha, \beta) &\sim \text{beta}(\alpha, \beta) \\
\psi \mid p &\sim \text{geomShift}(p, s = 15) \\
y_i \mid \psi &\sim \text{unif}(\{1, 2, \dots, \psi\})
\end{aligned}$$