

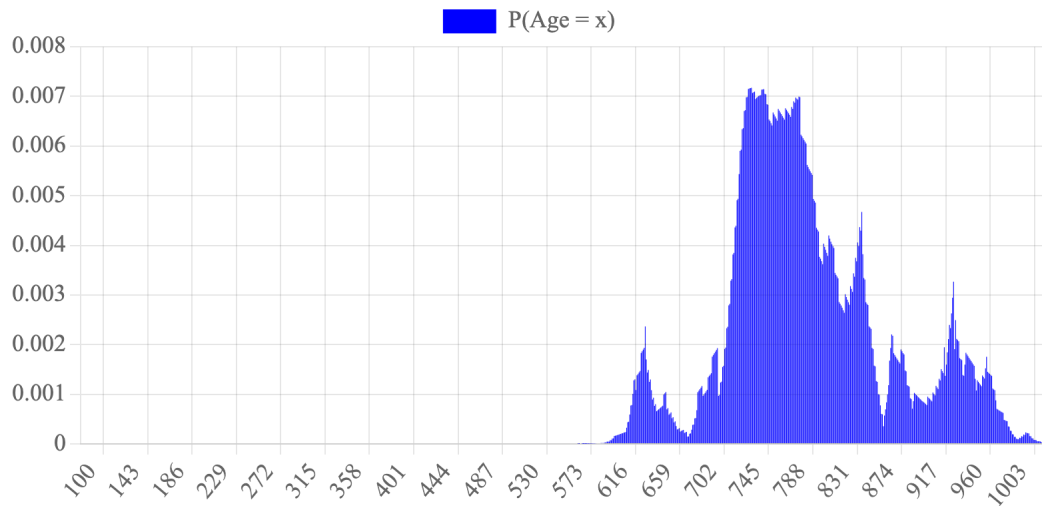
Bayesian Carbon Dating

We are able to know the age of ancient artefacts using a process called carbon dating. This process involves a lot of uncertainty! You observe a measurement of 90% of natural C14 molecules in a sample. What is your belief distribution over the age of the sample? This task requires probabilistic models because we have to think about two random variables together: the age A of the sample, and M the remaining C14 molecules.

Carbon Dating Demo

Imagine you have just taken a sample from your artifact. For the sample size you took, a living organism would have had 1000 molecules of C14. Use this demo to explore the relationship between how much C14 is left and your belief distribution for how old your artifact is.

Remaining C14:



Note: this demo was created in 2023 and the age reported is relative to that year! This chapter only has historical C14 rates from 10,000 years ago and as such is not able to estimate age when there are fewer than 350 molecules of C14 in the sample.



Carbon dating allows us to know the age of things that used to be alive, like dinosaur bones.

Understanding Decay of C14 molecule

All living things have a constant proportion of a radioactive molecule called C14 in them. When living things die those molecules start to decay radioactively. Specifically, the time to decay in years, T , of a single C14 molecule is distributed as an exponential, $T \sim \text{Exp}(\lambda = 1/8267)$ where 8267 is the mean life of C14.

Consider a single C14 molecule. What is the probability that it decays within 750 years?

$$\begin{aligned} P(T \leq 750) &= 1 - e^{-\lambda \cdot 750} && \text{Exp. CDF} \\ &= 1 - e^{-\frac{1}{8267} \cdot 750} \\ &= 0.0867 \end{aligned}$$

That is only for a single molecule. Since C14 molecules decay independently, it is not much harder to think of how many are left out of a larger initial count of C14. A particular sample started with 1000 molecules. What is the probability that exactly 900 are left after 750 years? This is equivalently to the event that 100 molecules have decayed.

$$\begin{aligned} X &\sim \text{Bin}(n = 1000, p = 0.0867) \\ P(X = 100) &= \binom{1000}{100} 0.0867^{100} (1 - 0.0867)^{900} \\ &\approx 0.0144 \end{aligned}$$

Let's generalize. Define M to be a random variable for the number of molecules left, and A to be the age of the sample. The probability $P(M = m | A = i)$ of having m remaining C14 molecules given that the artifact is i years old will be equal to $P(X = n - m)$ where n is the starting number of C14 molecules, $p = 1 - e^{-i/8267}$ and $X \sim \text{Bin}(n, p)$ is the count of decayed C14 molecules.

Inferring Age From C14

You observe a measurement of 900 C14 molecules in a sample. You assume that the sample originally had 1000 C14 molecules when it died. Infer $P(A = i | M = 900)$ where $A = i$ is the event that the sample organism died i years ago. Note that age is a discrete random variable which takes on whole numbers of years. You will need use $P(M = m | A = i)$ from the previous part.

For your prior belief you know that the sample must be between $A = 100$ and $A = 10000$ inclusive and you assume that every year in that range is equally likely.

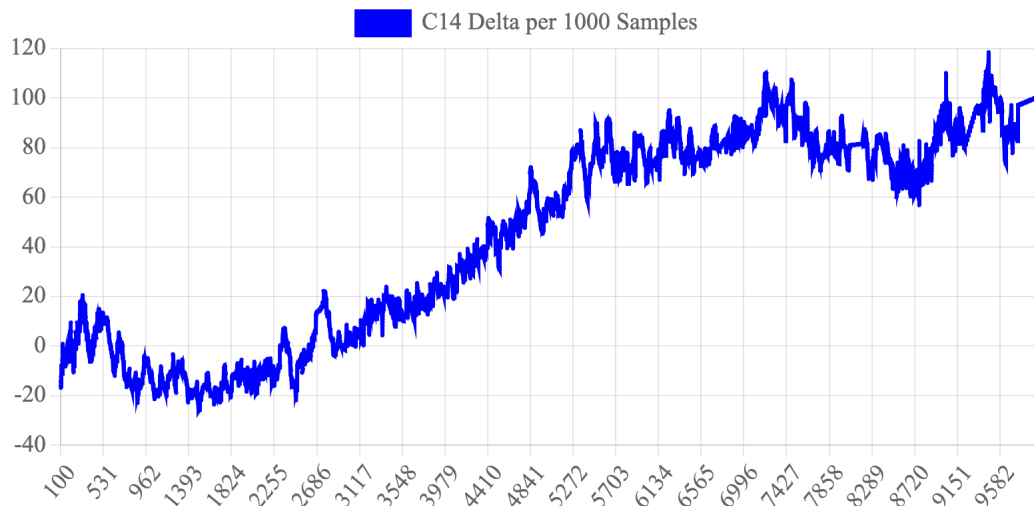
This is a perfect case for Bayes theorem. However instead of updating our belief in an event, like we did in [Part 1](#), we are updating the belief over all the values that a random variable can take on, a process called [inference](#). Here is the generalized version of Bayes' theorem for inferring age, A :

$$\begin{aligned} P(A = i | M = 900) &= \frac{P(M = 900 | A = i) P(A = i)}{P(M = 900)} \\ &= P(M = 900 | A = i) \cdot \frac{P(A = i)}{P(M = 900)} \\ &= P(M = 900 | A = i) \cdot K \end{aligned}$$

The critical part of the last line was to recognize that $\frac{P(A=i)}{P(M=900)}$ is a constant with respect to i . The term $P(A = i)$ is constant as our prior over A was uniform. We could compute the value of $P(M = 900)$ explicitly using the law of total probability. In code this is most easily implemented by computing all values of $P(M = 900 | A = i)$ and normalizing as K will be the value that makes all of the values $P(A = i | M = 900)$ sum to 1.

Fluctuating History

The amount of C14 in the atmosphere fluctuates over time; it is not a constant baseline! Here is the delta C14 (per 1000 molecules) that you would have found if the object died different number of years ago. To incorporate this information we simply start our binomial with 1000 molecules plus the delta for the year, downloaded from a public dataset [1]:



This offset is archeology theory not probability theory. We include it in this chapter because otherwise our code will give an incorrect preiction. Also, it gives the posterior a really interesting shape (see the demo).

Python Code

The math, derived above, leads to the following python code for a function `inference(m)` which returns the probability mass function for age A , given an observation of m C14 molecules in a sample that should have 1000 molecules, were it alive today. Notice the use of normalization to avoid explicitly computing the prior or $P(M = m)$ from Bayes Theorem.

```

from scipy import stats
import math

C14_MEAN_LIFE = 8267

def inference(m = 900):
    """
    Returns a dictionary A, where A[i] contains the
    corresponding probability, P(A = i | M = m).
    m is the number of C14 molecules remaining and i
    is age in years. i is in the range 100 to 10000
    """
    A = {}
    for i in range(100, 10000+1):
        A[i] = calc_likelihood(m, i) # P(M = m | A = i)
    # implicitly computes the normalization constant
    normalize(A)
    return A

def calc_likelihood(m, age):
    """
    Computes P(M = m | A = age), the probability of
    having m molecules left given the sample is age
    years old. Uses the exponential decay of C14
    """
    n_original = 1000 + delta_start(age)
    n_decayed = n_original - m
    p_single = 1 - math.exp(-age/C14_MEAN_LIFE)
    return stats.binom.pmf(n_decayed, n_original, p_single)

def normalize(prob_dict):
    # first compute the sum of the probability
    sum = 0
    for key, pr in prob_dict.items():
        sum += pr
    # then divide each probability by that sum
    for key, pr in prob_dict.items():
        prob_dict[key] = pr / sum
    # now the probabilities sum to 1 (aka are normalized)

def delta_start(age):
    """
    The amount of atmospheric C14 is not the same every
    year. If the sample died "age" years ago, then it would
    have started with slightly more, or slightly less than
    1000 C14 molecules. We can look this value up from the
    IntCal database. See the next section!
    """
    return historical_c14_delta[age]

```

Industry Strength Bayesian Carbon Dating

There are other sources of uncertainty in Carbon Dating which we have not considered. For example a common source of uncertainty is laboratory error: if the same sample were sent to a lab several times, different results would come back.

Perhaps the most fascinating extension is modelling "stratigraphic" relationships. Often in archeological sites, you can know relative age of artifacts based on their position in sediment. This requires a joint model of the age of each artifact with the constraint that you *know* some are older than others. Inference can then be performed using a General Inference technique (often MCMC) and will be much more accurate.

Binomial Approximations?

Could we have used an approximation for the binomial PMF calculation? In Bayesian Carbon dating *both* a normal and a poisson approximation are appropriate. The decay binomial $X \sim \text{Bin}(n, p)$ is well approximated by either a Poisson with $\lambda = n \cdot p$ or a Gaussian with $\mu = n \cdot p$ and $\sigma^2 = n \cdot p \cdot (1 - p)$. This could be used to speed up calculations. Let's rework the example where we had $X \sim \text{Bin}(n = 1000, p = 0.0867)$. We computed that $P(X = 100) = 0.0144$

Poisson Approximation:

$$\begin{aligned} Y &\sim \text{Poi}(\lambda = 86.7) \\ P(X = 100) &\approx P(Y = 100) \\ &= \text{scipy.stats.poi.pmf}(100, 86.7) \\ &\approx 0.0151 \end{aligned}$$

Normal Approximation:

$$\begin{aligned} Y &\sim N(\mu = 86.7, \sigma^2 = 79.2) \\ P(X = 100) &\approx P(Y > 100.5) - P(Y > 99.5) \\ &\approx 0.0146 \end{aligned}$$

[1] [IntCal Historical Atmospheric C14](#)