Chris Piech
CS 109

Section #4
Oct 23, 2025

Section 4

With questions by Chris

1 Beta ±

You have observed 30 successes and 20 fails for an event with unknown probability. Based on this information you can model the probability of success, X, as a Beta random variable. You want to make a claim of the form: "`the probability of success is m ± b.`"

Select m to be the expectation of X. Select b to be the smallest value, *rounded to two decimal places*, such that $P(m - b < X < m + b)$ is greater than or equal to 95%. Provide your answer as pseudocode that prints out the values m and b.

Incremental solution:

```
def main():
    X = stats.beta(A, B)
    mean = A / (A + B)
    x = 0
    while True:
        lower = mean - x
        upper = mean + x
        pr = X.cdf(upper) - X.cdf(lower)
        if pr > 0.95:
            break
        x += 0.01
    print(f'{mean:.2f} +- {x:.2f}')
```

2 Bayesian RNA Quantification

Let N_A and N_B be the count of type-A RNA and type-B RNA in a person's body. Both are Poisson distributed:

$$N_A \sim \text{Poi}(12 \cdot X) \qquad N_B \sim \text{Poi}(12 \cdot (1 - X))$$

N_A and N_B have rates that come from an unknown continuous variable X. X takes on values between 0 and 1 and is an indicator of a person's overall health. N_A and N_B are independent, conditioned on knowing X.

a. For part (a) only, assume $X = 1/4$. What is $P(N_A > 10)$?

$N_A \sim \text{Poi}(3)$, so $P(N_A > 10) = 1 - \sum_{k=0}^{10} \frac{3^k e^{-3}}{k!}$.

b. Our prior belief is that $X \sim \text{Beta}(a = 2, b = 2)$. An experiment observes $N_A = 12$ and $N_B = 16$. Based on these observations, what is your updated distribution for X? In other words, what is $f(X = x|N_A = 12, N_B = 16)$? You can leave your answer in terms of a normalization constant.

Posterior is proportional to prior × likelihood. The likelihood is Poisson(Xk) for N_A and Poisson($(1 - X)k$) for N_B. Combined with Beta(2,2) prior yields a known Beta update if we treat $N_A + N_B$ as total counts. The prior is updated based on the observations of likelihood N_A and N_B. Therefore, our posterior beta is:

$$f(X = x|N_A = 12, N_B = 16) = \frac{P(N_A = 12, N_B = 16|X = x)f(X = x)}{P(N_A = 12, N_B = 16)}$$

$$= \frac{(12x)^{N_A}e^{-12x}}{N_A!} \cdot \frac{(12(1-x))^{N_B}e^{-12(1-x)}}{N_B!} \cdot B \cdot x^{a-1}(1-x)^{b-1}$$

$$= K \cdot x^{a+N_A-1}(1-x)^{b+N_B-1}$$

This simplifies to the beta $\text{Beta}(2 + 12, 2 + 16) = \text{Beta}(14, 18)$.

c. What is the variance of your updated belief distribution for X, found in part (b)?

Use the variance formula for the resulting Beta distribution, $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{14 \cdot 18}{32^2 \cdot 33}$.

3 Midterm Review (Optional)

m strings are hashed (not necessarily uniformly) into a hash table with n buckets. Each string's hash is an independent trial, and the probability that a string hashes to bucket i is p_i, with $\sum_{i=1}^{n} p_i = 1$.

a. Let E be the event that bucket 1 has ≥ 1 string hashed into it. What is $P(E)$?

Define S_i = event that string i hashes to bucket 1. Then S_i^C = string i does not hash to bucket 1.
We want E = "at least one string goes to bucket 1." Using DeMorgan's Law:

$$P(E) = 1 - P(E^C) = 1 - P(\text{no string goes to bucket 1}).$$

Each string avoids bucket 1 with probability $(1 - p_1)$, so:

$$P(E) = 1 - (1 - p_1)^m.$$

b. Let E be the event that at least one of buckets 1 to k has ≥ 1 string hashed into it. What is $P(E)$?

Define F_i = event that bucket i has at least one string hashed into it. We want $P(E) = P(F_1 \text{ or } F_2 \text{ or } \ldots \text{ or } F_k)$. F_i bucket events are dependent, so we cannot just add these! By DeMorgan's Law:

$$P(E) = 1 - P(F_1^C F_2^C \ldots F_k^C),$$

where F_i^C = "bucket i is empty." A string avoids being hashed to any bucket 1 through k with probability $(1 - p_1 - p_2 - \cdots - p_k)$, so:

$$P(E) = 1 - (1 - p_1 - p_2 - \cdots - p_k)^m.$$

c. E = each of buckets 1 to k has ≥ 1 string hashed into it. What is $P(E)$?

Let F_i be the event that bucket i has at least one string hashed into it. We want $P(F_1, F_2, \ldots, F_k)$, the probability that all of the first k buckets get at least one string. It helps to think in terms of the complement F_i^C, the event that bucket i is empty. The event we want, $(F_1, F_2, \ldots, F_k)$, means none of these buckets are empty. Its complement is that *at least one* of them is empty:

$$(F_1, F_2, \ldots, F_k)^C = (F_1^C \text{ or } F_2^C \text{ or } \ldots \text{ or } F_k^C).$$

So:

$$P(F_1, F_2, \ldots, F_k) = 1 - P(F_1^C \text{ or } F_2^C \text{ or } \ldots \text{ or } F_k^C).$$

Let's see what this looks like for small k.

Case $k = 2$.

$$P(F_1, F_2) = 1 - P(F_1^C \text{ or } F_2^C).$$

This complement includes *any* outcome where at least one bucket is empty, not just when both are empty. We use inclusion exclusion to combine these possibilities:

$$P(F_1, F_2) = 1 - [P(F_1^C) + P(F_2^C) - P(F_1^C, F_2^C)].$$

Let's break down each term:
$P(F_1^C) = (1 - p_1)^m$ (all m strings avoid bucket 1)
$P(F_2^C) = (1 - p_2)^m$ (all strings avoid bucket 2)
$P(F_1^C, F_2^C) = (1 - p_1 - p_2)^m$ (all strings avoid both 1 and 2)
So:

$$P(F_1, F_2) = 1 - (1 - p_1)^m - (1 - p_2)^m + (1 - p_1 - p_2)^m.$$

Case $k = 3$. We now have three ways for buckets to be empty. The pattern continues:

$$\begin{aligned} P(F_1, F_2, F_3) = 1 &- [(1 - p_1)^m + (1 - p_2)^m + (1 - p_3)^m] \\ &+ [(1 - p_1 - p_2)^m + (1 - p_1 - p_3)^m + (1 - p_2 - p_3)^m] \\ &- (1 - p_1 - p_2 - p_3)^m. \end{aligned}$$

General case. We add and subtract across all possible subsets S of $\{1, \ldots, k\}$. For any subset S, the event that all buckets in S are empty has probability $\left(1 - \sum_{i \in S} p_i\right)^m$. So in general:

$$P(F_1, \ldots, F_k) = \sum_{S \subseteq \{1, \ldots, k\}} (-1)^{|S|} \left(1 - \sum_{i \in S} p_i\right)^m.$$

Fun fact: this is a famous probability problem - also known as the Coupon Collector's problem.