

STSCI 4780:
Continuous parameter estimation
with discrete data

Tom Loredo, CCAPS & SDS, Cornell University

2020-02-04

Plan

Previous lectures

- **Lec03:** Binary hypothesis space (C, \overline{C}), binary data ($+, -$)
- **Lec04:** Larger discrete hypothesis spaces (doors, α_i for coin), larger discrete data spaces (binary sequences, counts)

Lec05 – 07 — Continuous parameter est'n with discrete data

- Bernoulli trials with continuous success probability, α
- Multinomial distribution: Multiple, discrete outcomes (categorical data)
- Poisson distribution: Inferring rates from count data over intervals

Then we'll do the normal (Gaussian) continuous/continuous case

Terminology

We'll be calculating $P(H_i|\dots)$, $P(D_{\text{obs}}|\dots)$ over discrete and continuous spaces. We'll use them to calculate other probabilities: $P(H_1 \vee H_2|\dots)$, $P(H_i, D_{\text{obs}}|\dots)\dots$

Discrete spaces

Alternatives: $H_1, H_2 \dots$ for $i \in \mathbb{Z}$

$p_i \equiv P(H_i|\dots)$ is a *probability mass function* (PMF)

May use other similar symbols: $p(i)$, f_i , $g(i)$

Continuous spaces

Let θ_* denote the (unknown!) true value of the parameter
Basic (“point”) hypotheses: $H_\theta \equiv “\theta_* = \theta,”$ for candidate
 $\theta \in \mathbb{R}$

Interesting composite hypotheses:

- $\theta_* > 0$, or $\theta_* > \theta$ as function of θ
- $\theta_* < 5$, or $\theta_* < \theta$
- $\theta_* \in [1, 2]$ or $\theta_* \in [\theta_l, \theta_u]$

LTP \rightarrow We can compute probabilities for these by summing
(*integrating*) probabilities for the associated point hypotheses

Define a *probability density function* (PDF) by:

$$p(\theta | \dots) d\theta \equiv P(\theta_* \in [\theta, \theta + d\theta] | \dots) \quad \text{for small } d\theta$$

Note that $p(\theta | \dots)$ *has dimensions* $[1/\theta]$

May use other similar symbols: $f(\theta)$, $g(\theta)$, $\pi(\theta), \dots$

Bayes's theorem for θ

Abbreviate $\theta_* \in [\theta, \theta + d\theta]$ as $\theta_* \in d\theta$

Bayes's theorem:

$$P(\theta_* \in d\theta | D) = \frac{P(\theta_* \in d\theta)P(D|\theta_* \in d\theta)}{P(D)} \quad || \mathcal{C}$$

- Use $p() d\theta$ for the θ_* probabilities — $d\theta$'s cancel!
- Assume the sampling distribution is continuous WRT θ , so $P(D|\theta_* \in d\theta) = P(D|\theta_* = \theta) \equiv \mathcal{L}(\theta)$

$$\rightarrow p(\theta|D) = \frac{p(\theta)\mathcal{L}(\theta)}{P(D)} \quad || \mathcal{C}$$

with $P(D) = \int d\theta p(\theta)\mathcal{L}(\theta)$

BT holds for PDFs!

“The labyrinth of the continuum”

There are two famous labyrinths where our reason very often goes astray: one concerns the great question of the Free and the Necessary, above all in the production and the origin of Evil; the other consists in the discussion of continuity and of the indivisibles which appear to be the elements thereof, and where the consideration of the infinite must enter in. The first perplexes almost all the human race, the other exercises philosophers only.

—G. W. Leibniz

The continuum is *tricky*! When θ_* may lie in a finite interval:

- $P(\theta_* = C | \dots) = 0$ (this corresponds to $d\theta = 0$)
- $P(\theta_* \in \mathbb{Z} | \dots) = 0$, where \mathbb{Z} = set of integers
- $P(\theta_* \in \mathbb{Q} | \dots) = 0$, where \mathbb{Q} = set of rationals
- There are strange subsets (e.g., dense dust-like sets; see Banach-Tarski paradox: Wikipedia, YouTube)

Measure theory adopts careful, technical terminology and notation arises to address potential pathologies associated with the

Binary Outcomes:

Estimating the outcome probabilities

Setup

\mathcal{C} specifies existence of two outcomes, S and F , in each of N cases or trials; for each case or trial, the probability for S is α ; for F it is $(1 - \alpha)$

The trial probabilities are *IID* (independent and identically distributed)

H_i = Statements about α , the probability for success on the next trial \rightarrow seek $p(\alpha|D, \mathcal{C})$

D = Sequence of results from N observed trials:

FFSSSSFSSSFS ($n = 8$ successes in $N = 12$ trials)

Likelihood (Bernoulli process)

$$\begin{aligned} P(D|\alpha, \mathcal{C}) &= P(\text{failure}|\alpha, \mathcal{C}) \times P(\text{failure}|\alpha, \mathcal{C}) \times \cdots \\ &= \alpha^n (1 - \alpha)^{N-n} \end{aligned}$$

Prior

Starting with no information about α beyond its definition, use as an “uninformative” prior $p(\alpha|\mathcal{C}) = 1$

Justifications:

- *Intuition*: Don't prefer any α interval to any other of same size
- *Prior predictive ignorance*: Bayes's suggested “ignorance” here can mean that before doing the N trials, we have no preference for how many will be successes:

$$P(n \text{ successes} | \mathcal{C}) = \frac{1}{N+1} \quad \rightarrow \quad p(\alpha | \mathcal{C}) = 1$$

Consider the uniform prior a *convention*—an assumption added to \mathcal{C} to make the problem well posed

Prior Predictive

$$\begin{aligned} p(D|\mathcal{C}) &= \int d\alpha \alpha^n (1 - \alpha)^{N-n} \\ &= B(n+1, N-n+1) = \frac{n!(N-n)!}{(N+1)!} \end{aligned}$$

A Beta integral, $B(a, b) \equiv \int dx x^{a-1} (1-x)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

Posterior

$$p(\alpha|D, \mathcal{C}) = \frac{(N+1)!}{n!(N-n)!} \alpha^n (1-\alpha)^{N-n}$$

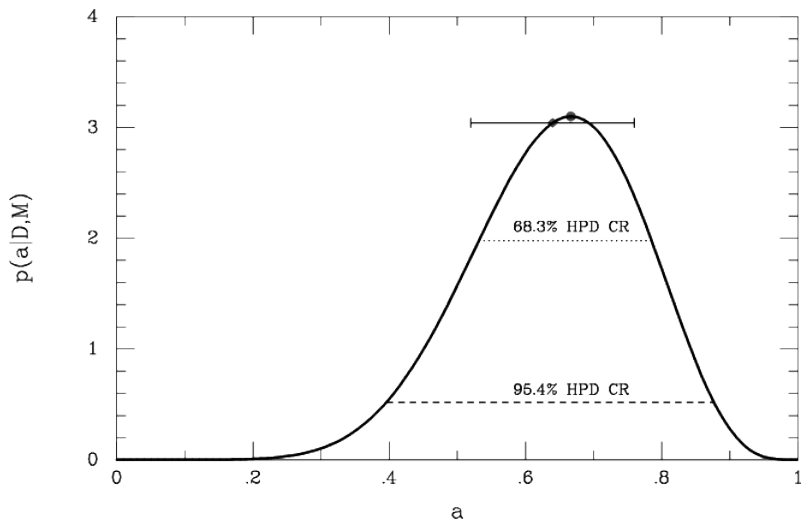
A *Beta distribution*. Summaries:

- Point estimates (“best-fit” values):
mode $\hat{\alpha} = \frac{n}{N} = 2/3$
posterior mean $\langle \alpha \rangle = \frac{n+1}{N+2} \approx 0.64$
- Uncertainty: std dev’n $\sigma_{\alpha} = \sqrt{\frac{(n+1)(N-n+1)}{(N+2)^2(N+3)}} \approx 0.12$
(For large N and n , and $N \gg n$, $\sigma_{\alpha} \approx \sqrt{n}/N$)

Find *highest posterior density credible regions* (HPD regions) numerically, or with *incomplete beta function*

Note that the posterior depends on the data only through n , not the N binary numbers describing the sequence

n is a (minimal) *sufficient statistic*



Beta distribution (in general)

A two-parameter family of distributions for a quantity α in the unit interval $[0, 1]$:

$$p(\alpha|a, b) = \frac{1}{B(a, b)} \alpha^{a-1} (1 - \alpha)^{b-1}$$

Summaries:

- Mode: $\hat{\alpha} = \frac{a-1}{(a-1)+(b-1)}$
- Mean: $\mu \equiv \mathbb{E}(\alpha) \equiv \langle \alpha \rangle = \frac{a}{a+b}$
- Variance: $\sigma^2 \equiv \text{Var}(\alpha) = \frac{ab}{(a+b)^2(a+b+1)}$
- Cumulative distribution (and interval probabilities) via incomplete beta function

(See Wikipedia for more properties)

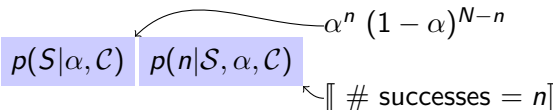
Binary Outcomes: Binomial Distribution

Suppose data is now $D' = n$ (number of heads in N trials), rather than the actual sequence. What is $p(\alpha|n, \mathcal{C})$?

Likelihood

Let \mathcal{S} = a sequence of flips with n heads.

$$\begin{aligned} p(n|\alpha, \mathcal{C}) &= \sum_{\mathcal{S}} p(\mathcal{S}|\alpha, \mathcal{C}) p(n|\mathcal{S}, \alpha, \mathcal{C}) \alpha^n (1 - \alpha)^{N-n} \\ &= \frac{N!}{n!(N-n)!} \alpha^n (1 - \alpha)^{N-n} \end{aligned}$$



The *binomial distribution* for n given α , N
(See Wikipedia for properties)

Marginal likelihood

$$\begin{aligned} p(n|\mathcal{C}) &= \frac{N!}{n!(N-n)!} \int d\alpha \alpha^n (1-\alpha)^{N-n} \\ &= \frac{1}{N+1} \end{aligned}$$

(Recall Bayes's motivation for the uniform prior...)

Posterior

$$\begin{aligned} p(\alpha|n, \mathcal{C}) &= \frac{\frac{N!}{n!(N-n)!} \alpha^n (1-\alpha)^{N-n}}{p(n|\mathcal{C})} \\ &= \frac{(N+1)!}{n!(N-n)!} \alpha^n (1-\alpha)^{N-n} \end{aligned}$$

Same result as when data specified the actual sequence

Another Variation: Negative Binomial

Suppose $D = N$, the number of trials it took to obtain a predefined number of successes, $n = 8$. What is $p(\alpha|N, C')$?

Likelihood

$p(N|\alpha, C')$ is probability for $n - 1$ successes in $N - 1$ trials, times probability that the final trial is a success:

$$\begin{aligned} p(N|\alpha, C') &= \frac{(N-1)!}{(n-1)!(N-n)!} \alpha^{n-1} (1-\alpha)^{N-n} \alpha \\ &= \frac{(N-1)!}{(n-1)!(N-n)!} \alpha^n (1-\alpha)^{N-n} \end{aligned}$$

The *negative binomial distribution* for N given α , n

Posterior

$$p(\alpha|N, \mathcal{C}') = f(n, N) \frac{\alpha^n (1 - \alpha)^{N-n}}{p(N|\mathcal{C}')}$$

$$p(N|\mathcal{C}') = f(n, N) \int d\alpha \alpha^n (1 - \alpha)^{N-n}$$

$$\rightarrow p(\alpha|D, \mathcal{C}') = \frac{(N+1)!}{n!(N-n)!} \alpha^n (1 - \alpha)^{N-n}$$

Same result as other cases

Final Variation: “Meteorological Stopping”

Suppose $D = (N, n)$, the number of samples and number of successes in an astronomy observing run whose total number was determined by the weather at the observatory. What is $p(\alpha|D, M)$?

(M adds info about weather to \mathcal{C} .)

Likelihood

$p(D|\alpha, M)$ is the binomial distribution times the probability that the weather allowed N samples, $W(N)$:

$$p(D|\alpha, M) = W(N) \frac{N!}{n!(N-n)!} \alpha^n (1-\alpha)^{N-n}$$

Let $f(n, N) = W(N) \binom{N}{n}$. We get the *same result* as before!

Likelihood principle

To define $\mathcal{L}(H_i) = p(D_{\text{obs}}|H_i, I)$, we must contemplate what other data we might have obtained. But the “real” sample space may be determined by many complicated, seemingly irrelevant factors; it may not be well-specified at all. Should this concern us?

Likelihood principle: The results of inference should depend only on how $p(D_{\text{obs}}|H_i, I)$ varies w.r.t. hypotheses. We can ignore aspects of the observing/sampling procedure that do not affect this dependence.

In Bayesian inference this happens because no sums of probabilities for hypothetical data appear in results; Bayesian calculations *condition on the observed data, D_{obs}* .

This is a sensible property that many frequentist methods do not share. Frequentist probabilities are “long run” rates of performance, and may depend on details of the sampling distribution that are irrelevant in a Bayesian calculation.

Likelihood principle from axioms

Birnbaum's theorem: If we agree that:

- Inferences should depend only on sufficient statistics when they are available (“sufficiency principle”)
- If I measure a quantity with one of two experiments—one precise, one less so—chosen at random, my uncertainty should depend only on the experiment actually used (“conditionality principle”)

Then the LP follows!

The vast majority of statisticians agree with both principles, but most use frequentist methods that don't obey the LP

“Conditional frequentist” methods try to *partly* condition on the data to avoid some problems that would arise in a rigorously frequentist analysis (an open research area; difficult & methods not unique)

The beta-binomial conjugate model

Generalize from the flat prior to a $\text{Beta}(\alpha|a, b)$ prior for α

$$\begin{aligned} p(\alpha|n, M') &\propto \text{Beta}(\alpha|a, b) \times \text{Binom}(n|\alpha, N) \\ &\propto \alpha^{a-1}(1-\alpha)^{b-1} \times \alpha^n(1-\alpha)^{N-n} \\ &\propto \alpha^{n+a-1}(1-\alpha)^{N-n+b-1} \end{aligned}$$

\Rightarrow the posterior is $\text{Beta}(\alpha|n+a, N-n+b)$

When the prior and likelihood are such that the posterior is in the same family as the prior, the prior and likelihood are a *conjugate* pair

A Beta prior is a conjugate prior for the Bernoulli process, binomial, and negative binomial sampling distributions

Conjugacy \rightarrow it's easy to chain inferences from multiple experiments