STSCI 4780: Continuous parameter estimation, cont'd

Tom Loredo, CCAPS & SDS, Cornell University

2020-02-06

Plan

Inference with discrete spaces

- **Lec03:** Binary hypothesis space (C, $\overline{C}$), binary data ($+$, $-$)
- **Lec04:** Larger discrete hypothesis space (doors, α_i), discrete data from *multiple* binary outcomes

Inference with continuous spaces

- PDFs vs. PMFs
- Bernoulli trials with *continous* parameter space
- Multinomial distribution: Multiple, discrete outcomes (categorical data)
- Poisson distribution: Inferring rates from count data over intervals

Recap

- PMFs on discrete spaces, PDFs on continuous spaces
- BT holds for PDFs ($\mathrm{d}\theta$'s cancel)
- Estimating the outcome probability, α, for binary outcomes

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 $p(\alpha|D, \mathcal{C})$

Adopt a *flat/uniform prior* as a default expression of initial ignorance about α — two motivations

Posterior (using sequence, binomial, negative binomial data)

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

A *Beta distribution*.

Beta distribution (in general)

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

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

A PDF over possible 2-outcome PMFs

The beta-binomial conjugate model

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

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

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

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

Probability & frequency

Recall $\hat{\alpha} = \frac{n}{N}$, the *relative frequency* of successes;
also $\sigma_\alpha \approx \frac{\sqrt{n}}{N}$ for $N, n \gg 1$

Frequencies arise when modeling repeated trials, or repeated sampling from a population or ensemble.

Finite-sample frequencies are observables

- When available, can be used to *infer* probabilities for next trial
- When unavailable, can be *predicted*

Bayesian/Frequentist relationships

- Relationships between probability and frequency
- Long-run performance of Bayesian procedures in IID settings (no accumulation of information)

Probability & frequency in IID settings

Frequency from probability

Bernoulli's (weak) law of large numbers: In repeated IID trials, given $P(\text{success}|\ldots) = \alpha$, predict

$$\frac{n_{\text{success}}}{N_{\text{total}}} \rightarrow \alpha \quad \text{as} \quad N_{\text{total}} \rightarrow \infty$$

"Bernoulli's swindle" — B. argued this justified estimating a next-trial probability with a (finite-sample) frequency

Probability from frequency

Bayes's "An Essay Towards Solving a Problem in the Doctrine of Chances" $\rightarrow$ First use of Bayes's theorem:

Probability for success in next trial of IID sequence:

$$\mathbb{E}(\alpha) \rightarrow \frac{n_{\text{success}}}{N_{\text{total}}} \quad \text{as} \quad N_{\text{total}} \rightarrow \infty$$

If $P(\text{success}|\ldots)$ does not change from sample to sample, it may be estimated using relative frequency data

Categorical data

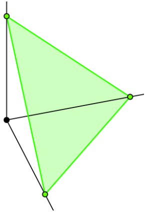
D = Discrete outcomes from N observed trials, $o_1 o_2 o_3 \ldots o_N$:

Roles of a die: 321344622...

Customer choices: AAOBBOOO... (*A*pple, *B*anana, *O*range)

$\mathcal{C}$ = Each outcome in one of K categories; parameters $\alpha \equiv \{\alpha_k\}$ such that $P(o_i = k|\alpha, \ldots, \mathcal{C}) = \alpha_k$ (*categorical distribution*)

Constraint: $\sum_k \alpha_k = 1$; equivalently $\alpha_K = 1 - \sum_{k=1}^{K-1} \alpha_k$
I.e., the K-dimensional α must lie on the $(K-1)$-dimensional standard simplex:



$K = 3$ case

Sequence sampling dist'n/Likelihood function

$$\begin{aligned} p(D|\alpha, \mathcal{C}) &= p(o_1 = k_1|\alpha, \mathcal{C}) \times p(o_2 = k_2|\alpha, \mathcal{C}) \times \cdots \\ &= \prod_k \alpha_k^{n_k} \\ &\equiv \mathcal{L}(\alpha) \end{aligned}$$

The counts (frequencies) are sufficient statistics

Count data sampling dist'n/Likelihood function

Take $D' = \{n_k\}$ (e.g., histogram); then the sampling PMF is a *multinomial dist'n*:

$$\begin{aligned} p(D'|\alpha, \mathcal{C}) &= \frac{N!}{\prod_k n_k!} \prod_k \alpha_k^{n_k} \\ &\propto \mathcal{L}(\alpha) \end{aligned}$$

The factor $N!/\prod_k n_k!$ counts the number of sequences having the stated numbers of outcomes in each category

Uniform prior

Prior PDF over $(K-1)$-D standard simplex:

$$p(\alpha_1, \ldots, \alpha_{K-1}|\mathcal{C}) = \begin{cases} C & \text{for } 0 \leq \alpha_k \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

with $1/C =$ "volume" of the $(K-1)$-D standard simplex satisfying the normalization constraint (one of the boundaries of the K-D corner simplex in the full α space)

Posterior

Posterior PDF over $(K-1)$-D standard simplex (using either D or D'):

$$p(\alpha_1, \ldots, \alpha_{K-1}|D, \mathcal{C}) \propto \begin{cases} \left[\prod_{k=1}^{K-1} \alpha_k^{n_k}\right] \left(1 - \sum_{k=1}^{K-1} \alpha_k\right)^{n_K} & \text{for } 0 \le \alpha_k \le 1 \\ 0 & \text{otherwise} \end{cases}$$

This has the form of a *Dirichlet dist'n* (the multivariate generalization of the beta dist'n)

Symmetrical treatment with delta functions

Write a PDF over a $(K-1)$-D standard simplex as a K-D function *constrained* to lie on the $(K-1)$-D simplex:

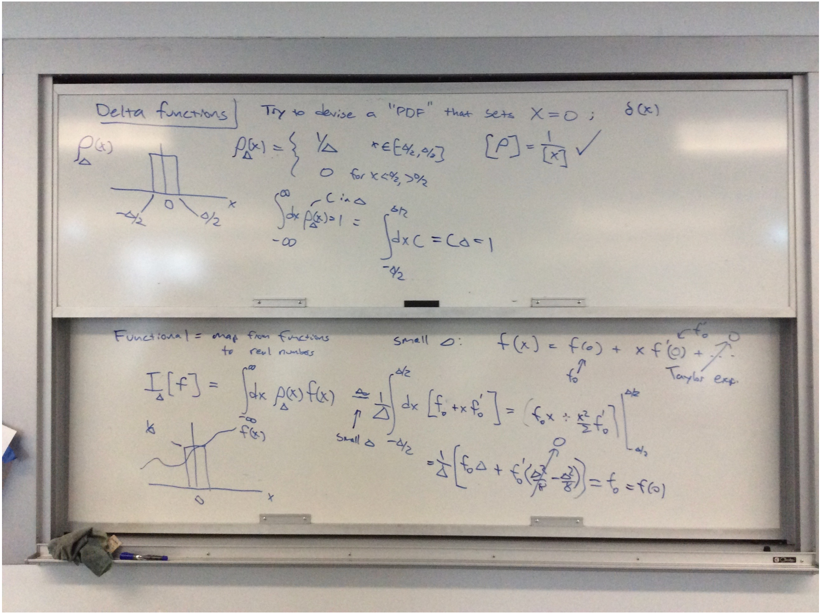
$$
\begin{aligned}
p(\alpha_1, \ldots, \alpha_K | \ldots) = \\
& p(\alpha_1, \ldots, \alpha_{K-1} | \ldots) \times p(\alpha_K | \alpha_1, \ldots, \alpha_{K-1}, \ldots)
\end{aligned}
$$

where $p(\alpha_K | \alpha_1, \ldots, \alpha_{K-1}, \ldots)$:

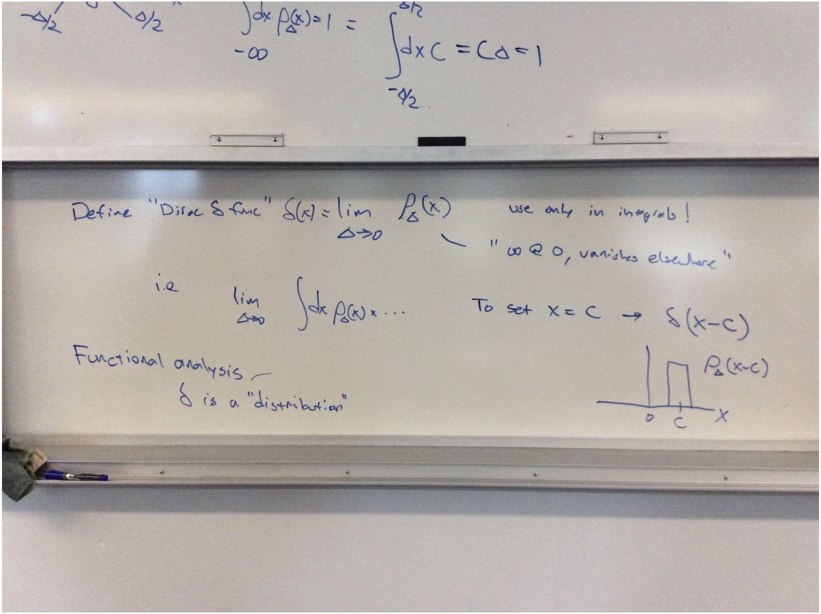
- Must set $\alpha_K = 1 - \sum_{k=1}^{K-1} \alpha_k$
- Must be a proper PDF (normalized!)

The *Dirac delta function*, $\delta(x)$, can accomplish this

Delta function development, whiteboards 1 & 2



Delta function development, whiteboard 3



Normalization constant

Generalized beta integral:

$$\int_0^\infty d\alpha_1 \cdots \int_0^\infty d\alpha_K \; \alpha_1^{\kappa_1 - 1} \cdots \alpha_K^{\kappa_K - 1} \; \delta\left(a - \sum_k \alpha_k\right) = \frac{\Gamma(\kappa_1)\cdots\Gamma(\kappa_K)}{\Gamma(\kappa_0)} \; a^{\kappa_0 - 1}$$

with $\kappa_0 \equiv \sum_{k=1}^K \kappa_k$

With $a = 1$ this is also known as the *multinomial beta function*

$$\Rightarrow p(\alpha|D,\mathcal{C}) = \frac{(N+K-1)!}{n_1! \cdots n_K!} \left[\prod_k \alpha_k^{n_k}\right] \delta\left(1 - \sum_k \alpha_k\right)$$

For $K = 2$ we recover beta posterior from Bernoulli/binomial cases

Marginal PDF for one category

Consider $K = 3$, but suppose we are interested only in α_1:

$$\begin{aligned} p(\alpha_1|D,\mathcal{C}) &= \int d\alpha_2 \int d\alpha_3 \; p(\alpha|D,\mathcal{C}) \\ &= C\alpha_1^{n_1} \int d\alpha_2 \int d\alpha_3 \; \alpha_2^{n_2}\alpha_3^{n_3} \\ &\qquad \times \delta\left[(1-\alpha_1) - (\alpha_2+\alpha_3)\right] \\ &= C'\alpha_1^{n_1}(1-\alpha_1)^{n_2+n_3+1}; \quad \text{note } n_2+n_3 = N - n_1 \end{aligned}$$

The marginal for a single category is a beta PDF, almost as if the data from the other categories were pooled—*but not quite*.

For the $K = 3$ case, there's an $N - n_1 + 1$ exponent, instead of the $N - n_1$ we might expect from pooling the data. This hints at a problem with the uniform prior we adopted; see Lab04 and Assignment03.

Dirichlet distributions

A family of "PDFs for PMFs," i.e., densities over possible categorical or multinomial distributions:

$$\mathrm{Dir}(\alpha|\kappa_1,\ldots,\kappa_K) = \frac{\Gamma(\kappa_0)}{\Gamma(\kappa_1)\cdots\Gamma(\kappa_K)}\left[\prod_{k=1}^{K}\alpha_k^{\kappa_k-1}\right]\delta\left(1-\sum_{k=1}^{K}\alpha_k\right)$$

with $\kappa_0 = \sum_k \kappa_k$; the κ_k are *concentration parameters*

Mode: $\hat{\alpha}_k = \frac{\kappa_k-1}{\kappa_0-K}$ for $k = 1\ldots K$

Marginal means: $\mathbb{E}(\alpha_k) = \frac{\kappa_k}{\kappa_0}$

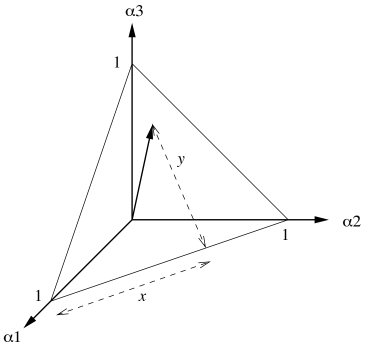
Marginal variances: $\mathrm{Var}(\alpha_k) = \frac{\kappa_i(\kappa_0-\kappa_i)}{\kappa_0^2(\kappa_0+1)}$

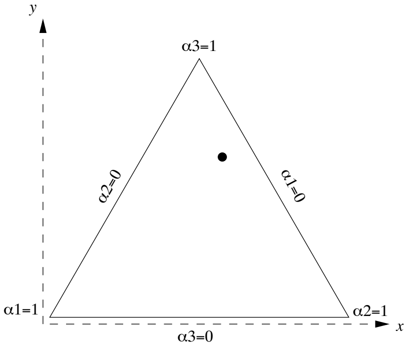
All covariances *negative* (necessarily!)

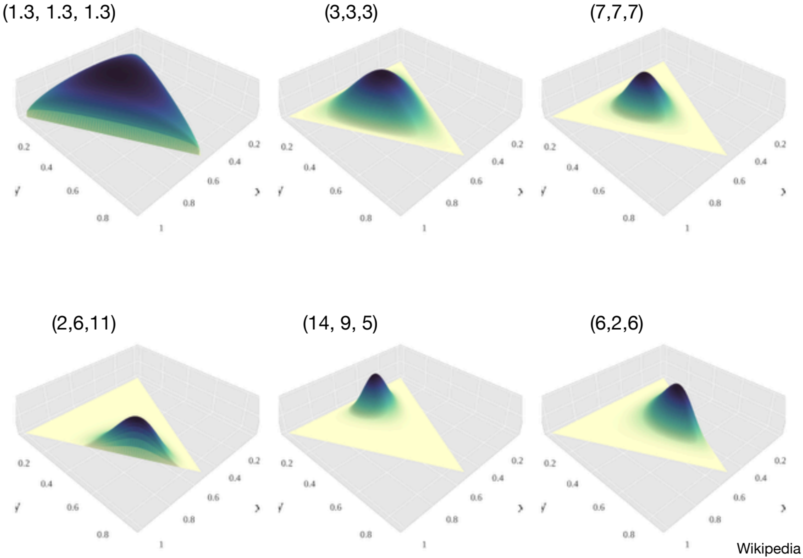
Special case: *Symmetric Dirichlet* with $\kappa_i = \kappa$

Dirichlet distribution priors are *conjugate priors* for categorical and multinomial likelihood functions

Simplex/ternary plots







Wikipedia