STSCI 4780/5780
Continuous parameter estimation, cont'd

Tom Loredo, CCAPS & SDS, Cornell University

Recap: Discrete and continuous spaces

We are calculating $P(H_i|\ldots)$, $P(D_{\rm obs}|\ldots)$ over spaces of alternatives labeled by discrete or continuous parameters.

$P()$ is a real-valued function of (logical) *arguments*, e.g., $H_i|\mathcal{C}$.

Discrete spaces

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

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

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

Continuous spaces

Alternatives: H_θ for $\theta \in \mathbb{R}$

$p(\theta)d\theta \equiv P(\theta \in [\theta, \theta + d\theta|\ldots)$ defines a *probability density function* (PDF), $p(\theta)$

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

More terminology

Functions, distributions, & measures

Mapping or **map:** An input/output relationship

PMF, PDF: Input is a label for a *single member* of the hypothesis/sample space (a set); output $\in [0, 1]$

$$p_i; \quad p(\theta)$$

Probability distribution or **measure:** Input is a *subset* in the space

$$\mathcal{P}(S) = \sum_{i \in S} p_i; \qquad \mathcal{P}(S) = \int_{\theta \in S} d\theta p(\theta)$$

Probability distribution function: Input is a label for a *single member defining a subset*; e.g., *cumulative distribution function* (CDF):

$$F_i = \sum_{j \leq i} p_j; \qquad F(\theta) = \int_{\theta' < \theta} d\theta' p(\theta')$$

Recap of inference with Bernoulli/binomial data

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 said to comprise a *conjugate* pair

A Beta prior is a conjugate prior for the Bernoulli process, and for the 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 (uniform/flat prior); 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

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

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

Probability from frequency

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

Compute *posterior probability* for success in next trial of IID sequence:

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

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

Poisson process: A continuous analog of the Bernoulli process

Bernoulli process and binomial distribution

Bernoulli process with success probability α produces binary sequences:

$$0110010011001001101010010001000001 \cdots$$

Report n, the count of 1s in a sequence of length $N \rightarrow$ **binomial distribution**:

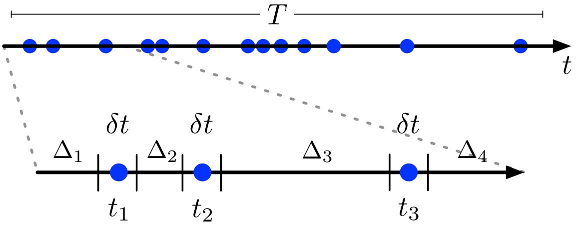
$$\begin{aligned}
\mathcal{L}(\alpha) &\equiv p(n|\alpha, \mathcal{C}) \\
&= \frac{N!}{n!(N-n)!} \alpha^n (1-\alpha)^{N-n}
\end{aligned}$$

Expected number of successes in N trials:

$$\mathbb{E}(n) = \alpha N$$

Poisson point process and Poisson (counting) distribution

Poisson point process with *intensity* λ (rate per unit interval):



Report n, the number of events in an interval of size T $\rightarrow$ **Poisson distribution**:

$$\begin{aligned}\mathcal{L}(\lambda) &\equiv p(n|\lambda, \mathcal{C}) \\ &= \frac{(\lambda T)^n}{n!} e^{-\lambda T}\end{aligned}$$

Expected number of counts in T:

$$\mathbb{E}((n)) = \lambda T$$

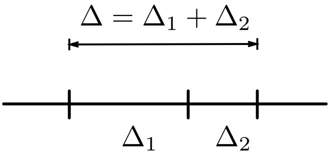
The Poisson distribution for counts from a point process

For occurence/arrival of n events in an interval Δ, let's seek

$$f_n(Delta) \equiv P(n \text{ events in } \Delta|\mathcal{P}),$$

where we'll figure out what we have to assume ($\mathcal{P}$) as we go

Partitioning an empty interval



$$f_0(\Delta) = P(\text{no events in } \Delta_1) \times P(\text{no events in } \Delta_2|\text{no events in } \Delta_1) \; || \, \mathcal{P}$$

As a simple modeling choice, let's *assume independence*:

$$P(\text{no events in } \Delta_2|\text{no events in } \Delta_1) = P(\text{no events in } \Delta_2) \quad || \, \mathcal{P}$$

Independence implies

$$f_0(\Delta) = P(\text{no events in } \Delta_1) \times P(\text{no events in } \Delta_2) \quad || \mathcal{P}$$
$$\rightarrow \boxed{f_0(\Delta_1 + \Delta_2) = f_0(\Delta_1) \times f_0(\Delta_2)}$$

This is a **functional equation** for $f_0(\cdot)$; it has two solutions:

$$f_0(\Delta) = 0 \qquad \text{so} \quad 0 = 0 \times 0$$
$$f_0(\Delta) = e^{-\lambda\Delta} \quad \text{so} \quad e^{-\lambda(\Delta_1+\Delta_2)} = e^{-\lambda\Delta_1} \times e^{-\lambda\Delta_2}$$

Let's use the *interesting one*:

$$\boxed{f_0(\Delta) = e^{-\lambda\Delta}}$$

Note this requires that *we specify a constant, λ*

Small interval behavior

What is the meaning of λ? Note that

$$P(1 \text{ or more events in } \Delta|\mathcal{P}) \;=\; 1 - f_0(\Delta) \;=\; 1 - e^{-\lambda\Delta}$$

If Δ is small so that $\lambda\Delta \ll 1$, then $e^{-\lambda\Delta} = 1 - \lambda\Delta + O(\Delta^2)$

$$P(1 \text{ or more events in } \Delta|\mathcal{P}) = \lambda\Delta + O(\Delta^2)$$

The probability of seeing *at least* one event in a small interval is $\propto \Delta$ (and λ), and $\lambda \geq 0$

What about 2 events in a small interval?

$$\begin{aligned} P(2 \text{ or more events in } \Delta|\mathcal{P}) &= [1 - f_0(\Delta)] - f_1(\Delta) \\ &= \lambda\Delta - f_1(\Delta) + O(\Delta^2) \end{aligned}$$

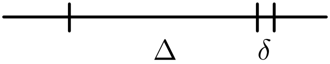
As another simplifying assumption, let's require that *events are simple*, so that this probability vanishes as $\Delta \to 0$ (multiple events can't happen at exactly the same instant/location):

$$\boxed{f_1(\Delta) = \lambda\Delta + O(\Delta^2)}$$

The probability of seeing *exactly* one event in a small interval is $\propto \Delta$, i.e., λ is a *rate parameter* (point process *intensity parameter*)

Extending an interval

To get a handle on the exact $f_n(\Delta)$ for $n > 0$, let's look at how $f_n(\Delta)$ changes if we grow the interval by a small amount δ:



Using the LTP we can write

$$f_n(\Delta + \delta) = f_n(\Delta) f_0(\delta) + f_{n-1}(\Delta) f_1(\delta) + f_{n-2}(\Delta) f_2(\delta) + \cdots$$

Let's exploit what we know about f_0 and small-interval behavior:

$$\begin{aligned} f_n(\Delta + \delta) &= f_n(\Delta) e^{-\lambda\delta} + f_{n-1}(\Delta)\lambda\delta + O(\delta^2) \\ &= f_n(\Delta)(1 - \lambda\delta) + f_{n-1}(\Delta)\lambda\delta + O(\delta^2) \\ f_n(\Delta + \delta) - f_n(\Delta) &= -\lambda\delta f_n(\Delta) + \lambda\delta f_{n-1}(\Delta) + O(\delta^2) \end{aligned}$$

Divide by δ and take $lim_{\delta \to 0}$:

$$\boxed{f_n'(\Delta) = -\lambda f_n(\Delta) + \lambda f_{n-1}(\Delta)}$$

This is a recursive sequence of inhomogeneous differential equations—infinitely many!

$$f_n'(\Delta) = -\lambda f_n(\Delta) + \lambda f_{n-1}(\Delta)$$

Let's check $n = 0$, where there is no inhomogeneous term:

$$f_0'(\Delta) = -\lambda f_0(\Delta)$$

The solution is $Ce^{-\lambda\Delta}$, but since we know $f_0(0) = 1$, we know $C = 1$

For $n = 1$,

$$f_1'(\Delta) = -\lambda f_1(\Delta) + \lambda e^{-\lambda\Delta}$$

As an inspired guess (or using *variation of parameters*), try

$$\begin{aligned} f_1(\Delta) &= \lambda\Delta e^{-\lambda\Delta} \\ \rightarrow f_1'(\Delta) &= -\lambda^2\Delta e^{-\lambda\Delta} + \lambda e^{-\lambda\Delta} \end{aligned}$$

which satisfies the differential eq'n

Iterating, we find:

$$\boxed{f_n(\Delta) = \frac{(\lambda\Delta)^n}{n!} e^{-\lambda\Delta}}$$

The Poisson distribution

If we model events distributed in an interval Δ such that:

- A single parameter, λ, governs the process
- With λ specified, probabilities for event counts in non-overlapping intervals are independent
- The events are simple (no two are at the same time/location)

then denoting these assumptions by $\mathcal{P} = \lambda, \mathcal{C}$ (and including the interval size in $\mathcal{C}$)

$$p(n|\lambda, \mathcal{C}) = \frac{(\lambda\Delta)^n}{n!} e^{-\lambda\Delta}$$

with λ corresponding to the event rate

We can show:

$$\mathbb{E}(n) = \lambda\Delta, \qquad \mathrm{Var}(n) = \lambda\Delta$$

"$\lambda, \mathcal{C}$" is analogous to "α, N IID trials" for binomial

Infer a Poisson rate from counts

Problem:

Observe n counts in T; infer rate (intensity), r

Likelihood

Poisson distribution:

$$\begin{aligned}\mathcal{L}(r) &\equiv p(n|r, \mathcal{C}) \\ &= \frac{(rT)^n}{n!} e^{-rT}\end{aligned}$$

Prior

Two simple "uninformative" standard choices:

- r known to be *nonzero*: it is a scale parameter; scale invariance $\rightarrow$

$$p(r|\mathcal{C}) = \frac{1}{\ln(r_u/r_l)}\,\frac{1}{r}$$

  This corresponds to a flat prior on $\lambda = \log r$

- r may *vanish*; require prior predictive $p(n|\mathcal{C}) \sim$ Const:

$$p(r|\mathcal{C}) = \frac{1}{r_u}$$

The *reference prior* ("uninformative" in an asymptotic, information-theoretic sense) is $p(r|\mathcal{C}) \propto 1/r^{1/2}$

Prior predictive

Adopting a flat (uniform) prior,

$$
\begin{aligned}
p(n|\mathcal{C}) &= \frac{1}{r_u}\frac{1}{n!}\int_0^{r_u} dr(rT)^n e^{-rT} \\
&= \frac{1}{r_u T}\frac{1}{n!}\int_0^{r_u T} d(rT)(rT)^n e^{-rT} \\
&\approx \frac{1}{r_u T} \quad \text{for} \quad r_u \gg \frac{n}{T}
\end{aligned}
$$

Posterior

A *gamma distribution*:

$$
p(r|n,\mathcal{C}) = \frac{T(rT)^n}{n!}e^{-rT}
$$

Gamma Distributions

A 2-parameter family of distributions over nonnegative x, with shape parameter α and scale parameter λ (or inverse scale ϵ):

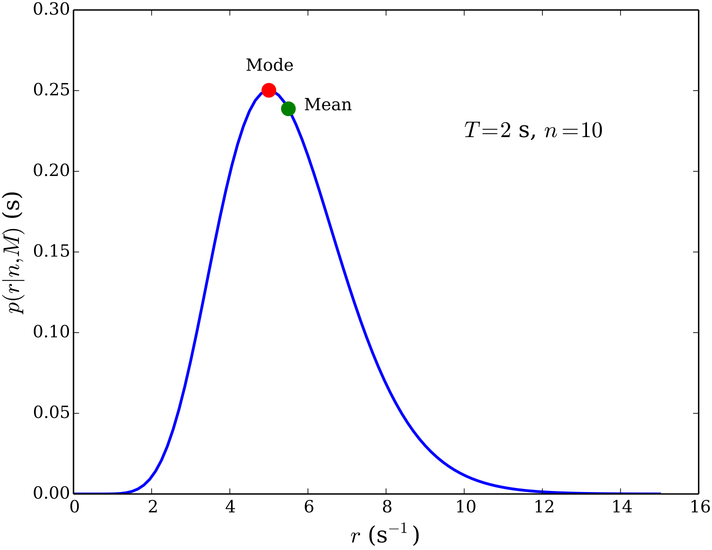
$$
\begin{aligned}
p_\Gamma(x|\alpha,\lambda) &\equiv \frac{1}{\lambda\Gamma(\alpha)}\left(\frac{x}{\lambda}\right)^{\alpha-1} e^{-x/\lambda} \\
&\equiv \frac{\epsilon}{\Gamma(\alpha)}(x\epsilon)^{\alpha-1}e^{-x\epsilon}
\end{aligned}
$$

Moments:

$$
\mathbb{E}(x) = \alpha\lambda = \frac{\alpha}{\epsilon} \qquad \mathrm{Var}(x) = \lambda^2\alpha = \frac{\alpha}{\epsilon^2}
$$

Our posterior corresponds to $\alpha = n + 1$, $\lambda = 1/T$.

- Mode $\hat{r} = \frac{n}{T}$; mean $\langle r \rangle = \frac{n+1}{T}$ (shift down 1 with $1/r$ prior)
- Std. dev'n $\sigma_r = \frac{\sqrt{n+1}}{T}$; credible regions found by integrating (can use incomplete gamma function)



0.30
0.25
0.20
0.15
0.10
0.05
0.00
$p(r|n,M)$ (s)
Mode
Mean
T=2 s, n=10
0
2
4
6
8
10
12
14
16
r (s^{-1})

Conjugate prior

Note that a gamma distribution prior is the conjugate prior for the Poisson sampling distribution:

$$
\begin{aligned}
p(r|n, M') &\propto \mathrm{Gamma}(r|\alpha, \epsilon) \times \mathrm{Pois}(n|rT) \\
&\propto r^{\alpha-1} e^{-r\epsilon} \times r^n e^{-rT} \\
&\propto r^{\alpha+n-1} \exp[-r(T+\epsilon)]
\end{aligned}
$$

Useful conventions

- Use a flat prior for a rate that may be zero
- Use a log-flat prior ($\propto 1/r$) for a nonzero scale parameter
- Use proper (normalized, bounded) priors
- Plot posterior with abscissa that makes prior flat (use log r abscissa for scale parameter case)

Supplementary material

Binomial for rare events

How many Cornell students share your birthday?

- $N \approx 24,000 \gg 1$
- $\alpha \approx \frac{1}{365} \ll 1$
- Expected number $\mu \equiv \mathbb{E}(n) = \alpha N \approx 66 \gg 1$

1000 bacteria are mixed in a liter of water. How many are in a 0.1 ml sample?

- $N = 1000 \gg 1$
- $\alpha = 10^{-4}$
- Expected number $\mu \equiv \mathbb{E}(n) = \alpha N = 0.1 \ll 1$

Seek an approximation for $p(n| \ldots)$ for small α, but not necessarily small μ (or n): A rare event can happen many times in a very large sample

Recall the binomial sampling distribution for n successes in N trials, given success probability α:

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

Expected number of successes $\mu \equiv \mathbb{E}(n) = \alpha N$

Recursion relation:

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

Consider the limit where $N \to \infty$ and $\alpha \to 0$, but with $\mu = \alpha N$ fixed and not necessarily small (but $\mu \ll N$); focus on $n \sim \mu$ so $n \ll N$ as well:

$$\frac{p(n)}{p(n-1)} \approx \frac{N\alpha}{n} = \frac{\mu}{n}$$

In that same limit, writing α in terms of μ and N,

$$p(0) = (1-\alpha)^N = \left(1 - \frac{\mu}{N}\right)^N \approx e^{-\mu}$$

Now evaluate $p(n)$ using the recurrence relation:

$$\begin{aligned} p(1) &= \frac{\mu}{1} \times p(0) = \mu e^{-\mu} \\ p(2) &= \frac{\mu}{2} \times p(1) = \frac{\mu^2}{2} e^{-\mu} \\ p(n) &= \frac{\mu}{n} \times p(n-1) = \frac{\mu^n}{n!} e^{-\mu} = \textit{Poisson} \end{aligned}$$

The *Poisson limit theorem* or *law of rare events* (events rare in *proportion*, though possibly numerous)