STSCI 4780/5780
Parameter estimation with continuous data: The normal (Gaussian) distribution

Tom Loredo, CCAPS & SDS, Cornell University

© 2022-02-15

Context: Univariate parameter estimation settings

Important analytically tractable cases $\rightarrow$ build intuition, skills to treat some parameters in harder problems

- Lec05 — Binary sequence data (sequence or count):
  - Bernoulli, binomial, negative binomial dist'ns
  - Beta posterior and prior dist'ns
  - Beta-Bernoulli, beta-binomial conjugate families
- Lec06 — Point process data (count in an interval):
  - Poisson point process and count distribution
  - Gamma distribution posterior
  - Gamma-Poisson conjugate pair

Today

- Real-valued samples with additive noise:
  - Normal (Gaussian) sampling distribution for IID samples
  - Normal dist'n posterior for location (width known)
  - Normal-normal conjugate pair; stable estimation
  - Uncertain width $\Rightarrow$ Student-t distribution posterior

Inference With Normals/Gaussians

Gaussian PDF

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{over } [-\infty, \infty]$$

Common abbreviated notation:

$$x \sim \mathcal{N}(\mu, \sigma^2)$$
$$x \sim \mathrm{Norm}(\mu, \sigma^2)$$
$$p(x|\mu,\sigma) = \mathrm{Norm}(x; \mu, \sigma)$$

Parameters

$$\begin{aligned} \mu &= \langle x \rangle \equiv \int dx\, x\, p(x|\mu,\sigma) \\ \sigma^2 &= \langle (x-\mu)^2 \rangle \equiv \int dx\, (x-\mu)^2\, p(x|\mu,\sigma) \end{aligned}$$

Gauss's Observation: Sufficiency

Suppose our data consist of N measurements with additive noise:

$$d_i = \mu + \epsilon_i, \qquad i = 1 \text{ to } N$$

Suppose the noise contributions are IID normal, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

$$\begin{aligned}
p(D|\mu,\sigma,\mathcal{C}) &= \prod_i p(d_i|\mu,\sigma,\mathcal{C}) \\
&= \prod_i p(\epsilon_i = d_i - \mu|\mu,\sigma,\mathcal{C}) \\
&= \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(d_i-\mu)^2}{2\sigma^2}\right] \\
&= \frac{1}{\sigma^N (2\pi)^{N/2}} e^{-Q(\mu)/2\sigma^2} \\
&\quad \text{with } Q(\mu) \equiv \sum_i (d_i - \mu)^2
\end{aligned}$$

Find dependence of Q on μ by *completing the square*:

$$\begin{aligned}
Q &= \sum_i (d_i - \mu)^2 \qquad\qquad [\text{Note: } Q/\sigma^2 = \chi^2(\mu)] \\
&= \sum_i d_i^2 + \sum_i \mu^2 - 2\sum_i d_i\mu \\
&= \left(\sum_i d_i^2\right) + N\mu^2 - 2N\mu\overline{d} \qquad \text{where } \overline{d} \equiv \frac{1}{N}\sum_i d_i \\
&= N(\mu - \overline{d})^2 + \left(\sum_i d_i^2\right) - N\overline{d}^2 \\
&= N(\mu - \overline{d})^2 + Nr^2 \quad \text{where } r^2 \equiv \frac{1}{N}\sum_i (d_i - \overline{d})^2
\end{aligned}$$

Likelihood depends on $\{d_i\}$ only through $\overline{d}$ and r:

$$\mathcal{L}(\mu, \sigma) = \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right)$$

The sample mean and variance are *sufficient statistics*

This is a miraculous compression of information—the normal dist'n is highly *abnormal* in this respect!

Estimating a Normal Mean

Problem specification

Model: $d_i = \mu + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, σ is *known*

Denote contextual info with σ explicit (for later): $\sigma, \mathcal{C}$

Parameter space: μ

We seek $p(\mu|D, \sigma, \mathcal{C})$

Likelihood

$$
\begin{aligned}
p(D|\mu, \sigma, \mathcal{C}) &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right) \\
&\propto \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right)
\end{aligned}
$$

"Uninformative" prior

- *Translation invariance:* $\Rightarrow p(\mu) \propto C$, a constant
- *Reference prior:* Asymptotic information theory criterion $\Rightarrow p(\mu) \propto C$

This prior is *improper* (can't be normalized) unless bounded; formally we should bound it and take ∞ limit

Prior predictive/normalization

$$
\begin{aligned}
p(D|\sigma,\mathcal{C}) &= \int d\mu\; C \exp\left(-\frac{N(\mu-\overline{d})^2}{2\sigma^2}\right) \\
&= C(\sigma/\sqrt{N})\sqrt{2\pi}
\end{aligned}
$$

... minus a tiny bit from tails, using a proper prior

Posterior

$$p(\mu|D,\sigma,\mathcal{C}) = \frac{1}{(\sigma/\sqrt{N})\sqrt{2\pi}} \exp\left(-\frac{N(\mu-\overline{d})^2}{2\sigma^2}\right)$$

Posterior is $N(\overline{d}, w^2)$, with standard deviation $w = \sigma/\sqrt{N}$

68.3% HPD credible region for μ is $\overline{d} \pm \sigma/\sqrt{N}$

Note that C drops out $\rightarrow$ limit of infinite prior range is well behaved

Informative Conjugate Prior

Use a normal prior, $\mu \sim N(\mu_0, w_0^2)$

Conjugate because the posterior turns out also to be normal

Posterior

Normal $N(\tilde{\mu}, \tilde{w}^2)$, but mean, std. deviation *"shrink"* towards prior

Define $B = \frac{w^2}{w^2 + w_0^2}$, so $B < 1$ and $B = 0$ when w_0 is large; then

$$\begin{aligned} \widetilde{\mu} &= \overline{d} + B \cdot (\mu_0 - \overline{d}) \\ \widetilde{w} &= w \cdot \sqrt{1 - B} \end{aligned}$$

Principle of stable estimation/precise measurement — "If observations are precise. . . relative to the prior, then the form and properties of the prior distribution have negligible influence on the posterior distribution." (ELS 1963)

Yes, prior probabilities often are quite vague and variable, but they are not necessarily useless on that account. . . . The impact of actual vagueness and variability of prior probabilities differs greatly from one problem to another. They frequently have but negligible effect on the conclusions obtained from Bayes' theorem, although utterly unlimited vagueness and variability would have utterly unlimited effect. If observations are precise, in a certain sense, relative to the prior distribution on which they bear, then the form and properties of the prior distribution have negligible influence on the posterior distribution. From a practical point of view, then, the untrammeled subjectivity of opinion about a parameter ceases to apply as soon as much data become available. More generally, two people with widely divergent prior opinions but reasonably open minds will be forced into arbitrarily close agreement about future observations by a sufficient amount of data.

Edwards, Lindman, and Savage (1963), 'Bayesian Statistical Inference for Psychological Research' (reprinted in *Breakthroughs in Statistics*

If plausible priors do not vary strongly over the region containing most of the volume of the integrated likelihood, the choice of prior negligibly affects inferences.

"When prior distributions can be regarded as essentially uniform"

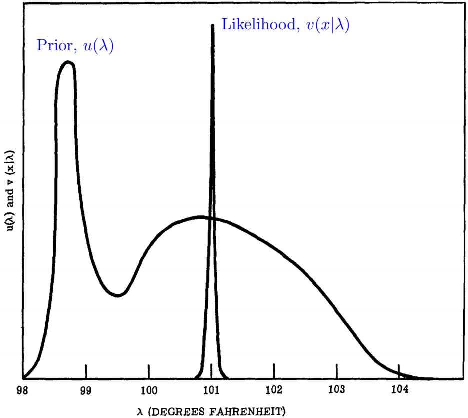
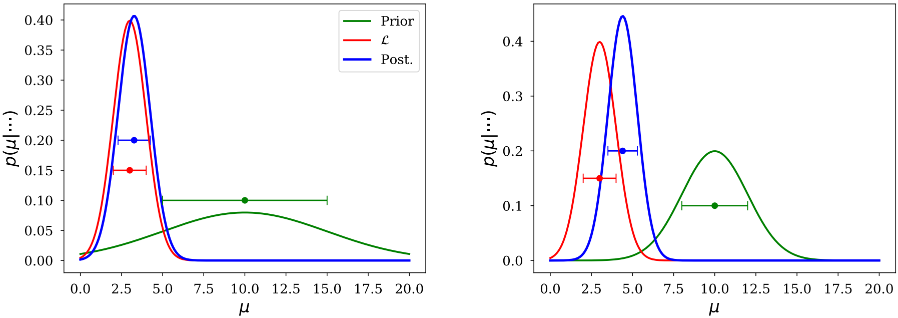


FIG. 1. $u(\lambda)$ and $v(x|\lambda)$ for the fever thermometer example. (Note that the units on the y axis are different for the two functions.)

"Headachy and hot, you are convinced that you have a fever but are not sure how much..."

Conjugate normal examples:

- Data have $\overline{d} = 3$, $\sigma/\sqrt{N} = 1$
- Priors at $\mu_0 = 10$, with $w = \{5, 2\}$



Note we always have $\widetilde{w} < w$ (in the normal-normal setup)

Estimating a Normal Mean: Unknown σ

Problem specification

Model: $d_i = \mu + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, σ is *unknown*

Parameter space: (μ, σ), but *we're only interested in* μ; σ is a *nuisance parameter*

Seek $p(\mu|D, \mathcal{C})$ — a uinivariate *marginal posterior* in a bivariate setting

Likelihood

$$
\begin{aligned}
p(D|\mu, \sigma, \mathcal{C}) &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right) \\
&\propto \frac{1}{\sigma^N} e^{-Q(\mu)/2\sigma^2}
\end{aligned}
$$

$$
\text{where} \quad Q(\mu) = N\left[r^2 + (\mu - \overline{d})^2\right]
$$

Uninformative Priors

Assume priors for μ and σ are independent:

$$p(\mu, \sigma) = p(\mu)p(\sigma) \qquad || \, \mathcal{C}$$

Translation invariance $\Rightarrow p(\mu) \propto C$, a constant

Scale invariance $\Rightarrow p(\sigma) \propto 1/\sigma$ (flat in $\log \sigma$)

This is also the reference prior, and a "minimal sample size prior"

The posterior is improper in σ unless $N \geq 2$ (what we need to learn something about the width)

Joint Posterior for μ, σ

$$p(\mu, \sigma | D, \mathcal{C}) \propto \frac{1}{\sigma^{N+1}} e^{-Q(\mu)/2\sigma^2}$$

Marginal Posterior

The LTP instructs us to find the univariate posterior for μ by integrating (*marginalizing*) the joint PDF for (μ, σ) over σ:

$$p(\mu|D, \mathcal{C}) = \int \mathrm{d}\sigma \; p(\mu, \sigma|D, \mathcal{C})$$

Thus,

$$p(\mu|D, \mathcal{C}) \propto \int \mathrm{d}\sigma \; \frac{1}{\sigma^{N+1}} e^{-Q(\mu)/2\sigma^2}$$

Let $\tau = \frac{Q}{2\sigma^2}$ so $\sigma = \sqrt{\frac{Q}{2\tau}}$ and $|d\sigma| = \tau^{-3/2}\sqrt{\frac{Q}{2}}\, d\tau$

$$\begin{aligned} \Rightarrow p(\mu|D, \mathcal{C}) &\propto 2^{N/2} Q^{-N/2} \int \mathrm{d}\tau \; \tau^{\frac{N}{2}-1} e^{-\tau} \\ &\propto Q^{-N/2} \end{aligned}$$

$$\text{where} \quad Q(\mu) = N\left[r^2 + (\mu - \overline{d})^2\right]$$

Write $Q = Nr^2\left[1 + \left(\frac{\mu - \overline{d}}{r}\right)^2\right]$ and normalize:

$$p(\mu|D,\mathcal{C}) = \frac{\left(\frac{N}{2}-1\right)!}{\left(\frac{N}{2}-\frac{3}{2}\right)!\sqrt{\pi}}\frac{1}{r}\left[1 + \frac{1}{N}\left(\frac{\mu - \overline{d}}{r/\sqrt{N}}\right)^2\right]^{-N/2}$$

Student's t distribution, with $t = \frac{(\mu - \overline{d})}{r/\sqrt{N}}$

A symmetric "bell curve," but with power-law tails

Large N:

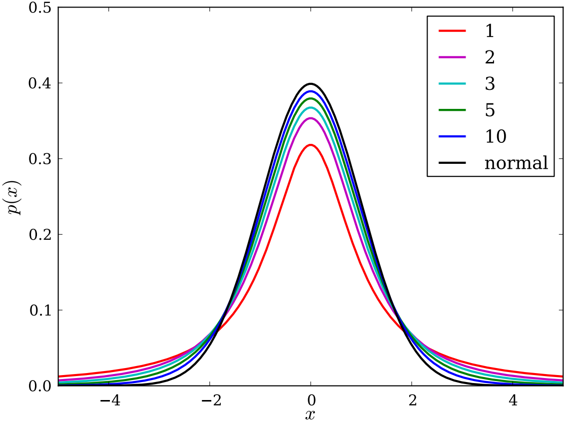
$$p(\mu|D,\mathcal{C}) \sim e^{-N(\mu-\overline{d})^2/2r^2}$$

A common *hack*: plug in an estimate of σ (e.g., by requiring $\chi^2/\text{dof} = 1$)

Marginalization doesn't just plug in a best σ; it slightly broadens the posterior to account for σ uncertainty

Student's t examples:

- $p(x) \propto \frac{1}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}$
- Location $= 0$, scale $= 1$
- Degrees of freedom $= \{1, 2, 3, 5, 10, \infty\}$



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

• • •

There are other experiments, however, which cannot easily be repeated very often; in such cases it is sometimes necessary to judge of the certainty of the results from a very small sample, which itself affords the only indication of the variability. Some chemical, many biological, and most agricultural and large scale experiments belong to this class, which has hitherto been almost outside the range of statistical enquiry.

"Student" = William Sealy Gosset, at Guinness & Son, Dublin!

Illustration III. In 1899 and in 1903 "head corn" and "tail corn" were taken from the same bulks of barley and sown in pots. The yields in grammes were as follows:

	1899	1903
Large seed	13·9	7·3
Small seed	14·4	8·7
	+ ·5	+ ·6

• • •

To test whether it is of advantage to kiln-dry barley seed before sowing, seven varieties of barley were sown (both kiln-dried and not kiln-dried) in 1899 and four in 1900; the results are given in the table.

Helmert & Luroth presented a Bayesian derivation in 1876

TRANSACTIONS
of THE NEW YORK
ACADEMY OF SCIENCES

STIGLER'S LAW OF EPONYMY*

Stephen M. Stigler

Department of Statistics
University of Chicago
Chicago, Illinois 60637

No reader of Robert K. Merton's work on the reward system of science could fail to be struck by his insightful and engaging discussions of the role of eponymy in the social structure of science. The uninitiated should read (and reread) his 1957 address, "Priorities in Scientific Discovery,"[1] but for present purposes I must at least repeat his definition of eponymy, as "the practice of affixing the name of the scientist to all or part of what he has found, as with the Copernican system, Hooke's law, Planck's constant, or Halley's comet."[2] Merton went on to discuss three levels of a hierarchic

• • •

I have chosen as a title for this paper, and for the thesis I wish to present and discuss, "Stigler's law of eponymy." At first glance this may appear to be a flagrant violation of the "Institutional Norm of Humility,"[4] and since statisticians are even more aware of the importance of norms than are members of other disciplines, I hasten to add a humble disclaimer. If there is an idea in this paper that is not at least implicit in Merton's *The Sociology of Science,* it is either a happy accident or a likely error. Rather I have, in the Mertonian tradition of the self-confirming hypothesis, attempted to frame the self-proving theorem. For "Stigler's Law of Eponymy" in its simplest form is this: "No scientific discovery is named after its original discoverer."

Supplementary material

Normal mean confidence & credible regions

Problem

Estimate the location (mean) of a Gaussian distribution from a set of samples $D = \{x_i\}$, $i = 1$ to N

Report a *point estimate*, and a *region* summarizing the uncertainty

Model

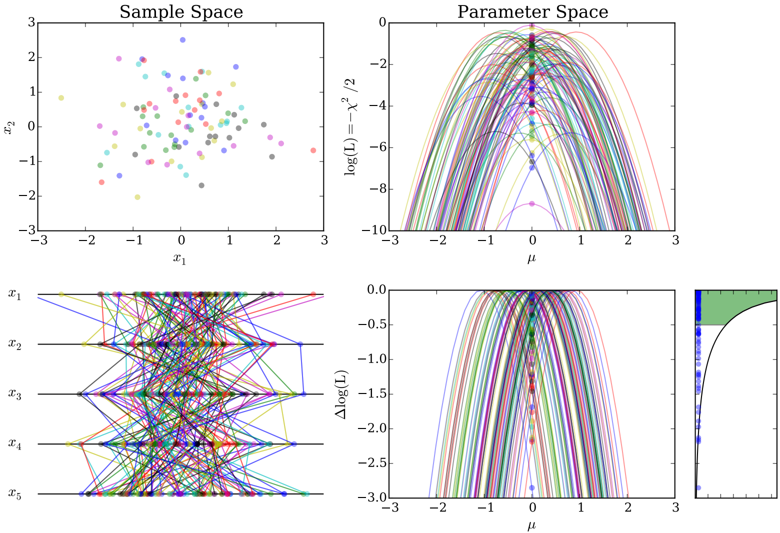
$$p(x_i|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$\text{Equivalently,} \quad x_i \sim \mathcal{N}(\mu,\sigma^2)$$

Here assume σ is *known*; we are uncertain about μ

Confidence interval construction (frequentist)

Likelihoods for 100 simulated data sets, $\mu = 0$



Credible interval construction

Normalize the likelihood for the observed sample; report the region that includes 68.3% of the normalized likelihood