Statistics for Astronomers: Lecture 14, 2019.04.11

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: quadratic alternatives to the KS test

There are two important alternatives to the KS test. They are both quadratic ECDF tests.

In quadratic ECDF tests, the distance is computed as $N \int_{-\infty}^{\infty} dF(x) \left[\hat{F}_n(x) - F(x)\right]^2 w(x)$,

where $w(x)$ is a weight function to emphasize different regions of the distribution.

The Cramér-von Mises test uses $w(x) = 1 \ \forall \ x$.

The Anderson-Darling test uses $w(x) = \dfrac{1}{F(x)(1 - F(x))} = \dfrac{1}{Var(\hat{F}_n(x))}$, placing more weight on observations near the tails of the distribution.

The AD test is more sensitive than the KS test to deviations in the tails of the distribution.

Recall: Spearman's rank correlation test

Given a sequence X_i of size N, the rank $\widetilde{X}_i$ of the data point X_i is its location in the ordered sequence. For example, if $X = [5.2, 2.1, 1.0, -1.1, 4.3]$, then $\widetilde{X} = [5, 3, 2, 1, 4]$.

The relationship between two variables X and Y can be quantified using Pearson's correlation coefficient: $\rho_{XY} = \dfrac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$.

Use Pearson's correlation coefficient, except with $\widetilde{X}, \widetilde{Y}$ instead of X, Y.

This is Spearman's rank correlation coefficient, $r_S = \dfrac{Cov(\widetilde{X},\widetilde{Y})}{\sqrt{Var(\widetilde{X})Var(\widetilde{Y})}}$.

ρ_{XY} measures the extent of linear relationship between X and Y. r_S, on the other hand, measures the extent of any monotonic relationship (think of the definition of d_i).

r_S is much more resistant to outliers!!

Recall: Bayes' Theorem, reframed

Given some prior information I, we can select a model M that includes parameters $\vec{\theta}$. Bayes' Theorem is then

$$\overbrace{\underbrace{p(M,\vec{\theta}|D,I)}_{\text{for model and model parameters}}}^{\text{posterior prob. distrib.}} = \frac{\overbrace{p(D|M,\vec{\theta},I)}^{\text{likelihood}}\ \overbrace{p(M,\vec{\theta}|I)}^{\text{prior}}}{\underbrace{p(D|I)}_{\text{prior predictive prob.}}}$$

This form is appropriate for parameter estimation.
Frequentist: parameters are fixed!
Bayesian: this is our degree of belief in a given value of the parameter.

For model selection, we expand the prior: $p(M,\vec{\theta}|I) = p(\vec{\theta}|M,I)\ p(M|I)$
(second term on RHS $\neq 1$ for model selection).

Priors

(from Ivezić et al.)

In terms of information, priors can be informative or "non-informative".

Informative prior
Specific information about parameter(s). Progressively increasing amounts of data $\Longrightarrow$ posterior is evidence-dominated.

Example: "Data from the past ten years suggests that there is a 2% change of rain in Morelia today between 2 and 3 PM."

Non-informative prior
Vague information about parameters, typically based on general principles/objective information (also called objective prior). "Light" modification to observations $\Longrightarrow$ posterior is likelihood-dominated.

Example: "The flux from this star is non-negative" ($0 \leq F < \infty$).
This is also an example of an improper prior, as it does not integrate to unity.
However, we are still OK if the resulting posterior is well-defined
(bus example from last week – $p(\tau|I) \propto 1/\tau, t \leq \tau < \infty$).

The Principle of Indifference is a classic example of an uninformative prior.

Priors (contd.)

Let $p(\theta|I) = C\theta^k$ for constants C, k ($k = 0$ gives the uniform distribution).

Define $y = a\,\theta$ (scaled version of θ, similar to changing units).
Activity: What must k be if we want the form of the prior in terms of y to remain unchanged?

$p(y) = Cy^k/a^{k+1}$.
For $k = -1, p(y) = Cy^k$, same form as $p(\theta)$. $\Longrightarrow$ scale-invariant prior for θ is $p(\theta|I) \propto 1/\theta$.

This was why, in the bus example, we chose $p(\tau|I) \propto \tau^{-1}$.

This is an example of a non-informative prior: "The prior for the scale parameter is independent of the choice of units." A similar prior for a location parameter demands independence from translations.

Two more probability distributions

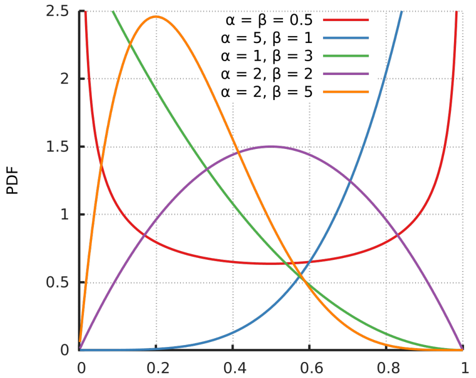
We'll use these in the current lecture as well as in the future for Bayesian inference.

$\text{Beta}(\alpha, \beta) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$ with $x \in (0,1)$ and $\alpha, \beta > 0$.

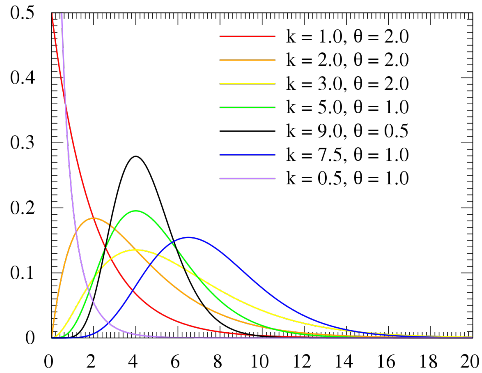
Mean: $\dfrac{\alpha}{\alpha+\beta}$; mode: $\dfrac{\alpha-1}{\alpha+\beta-2}$.

$\text{Gamma}(k, \theta) = \dfrac{1}{\theta^k \Gamma(k)}\, x^{k-1} \exp\left[-\dfrac{x}{\theta}\right]$ with $x \in (0,\infty)$ and $k, \theta > 0$.

Mean: $k\theta$; mode: $(k-1)\theta$.



Credit: de:User:Horas/en:User:Krishnavedala, Public domain.



Credit: w:User:Marksweep/en:User:Cburnett, CC BY-SA 3.0, via Wikimedia Commons.

Bayesian point/location and interval estimates

Once $p(\theta|\text{data})$ is computed, we can compute the location estimates (mean, median, mode).

For example, the Bayesian estimator of the parameter mean is $\bar{\theta} = \displaystyle\int d\theta\ \theta\ p(\theta|\text{data})$.

We can also compute Bayesian interval estimates, also called posterior intervals or credible intervals (abbreviated in these lectures as CrI).

One example of a $100(1-\alpha)\%$ CrI is $[a, b]$ such that

$$\int_{-\infty}^{a} d\theta\ p(\theta|\text{data}) = \int_{b}^{\infty} d\theta\ p(\theta|\text{data}) = \alpha/2.$$

Another type of CrI is the highest posterior density (HPD) interval, defined as the narrowest interval that contains $100(1-\alpha)\%$ of the posterior probability.

Numerical computation of HPD interval

1. Obtain N random deviates $x[i]$ drawn from the posterior density distribution.
2. Sort them in ascending order.
3. For each $x[i]$, find the point that is $w = (1-\alpha)N$ points away.
4. Compute the widths $w[i] = x[w+i] - x[i]$.
5. Find the location $i = i_0$ corresponding to the smallest width. The HPD interval is then $(x[i_0], x[i_0+w])$.

Write your own script! You'll need it for your research if you're using Bayesian inference.

Example from Wasserman's "All of Statistics"

A coin has an unknown probability θ of coming down heads. Flipping the coin N times, we observe s heads. Find the posterior distribution of θ.

Let us pick a prior $p(\theta) = U(0,1)$ so that the prior mean is 1/2 (expected for a fair coin).

The likelihood of obtaining s heads is $\mathscr{L}(\theta) \propto \theta^s(1-\theta)^{N-s}$.

The posterior is then $p(\theta|\mathrm{data}) = \mathscr{L}(\theta)p(\theta) \propto \theta^s(1-\theta)^{N-s} = \mathrm{Beta}(\alpha,\beta)$,

What are α and β? $\alpha = s+1,\ \beta = N-s+1$.

Posterior mean $\bar{\theta} = \dfrac{\alpha}{\alpha+\beta} = \dfrac{s+1}{N+2}$.

We can rearrange the above:

$$\bar{\theta} = \frac{s+1}{N+2} = \frac{s}{N+2} + \frac{1}{N+2} = \underbrace{\frac{s}{N}}_{\text{data mean}} \times \frac{N}{N+2} + \underbrace{\frac{1}{2}}_{\text{prior mean}} \times \frac{2}{N+2}$$

The posterior mean is thus the weighted average of the data mean and the prior mean. The effective sample size is then $N+2$.

Prior-dominated posterior

(from Andreon & Weaver, "Bayesian Methods for the Physical Sciences")

The prior can drive the posterior away from the data (likelihood) if it is steep and/or has very little overlap with the region where the likelihood dominates.

One example: inferring the true (photon) count rate from a faint source.
I observe a faint source once and get a photon count rate of $S_{\mathrm{obs}} = 4\ \mathrm{s}^{-1}$. Based on this observation, what is the constraint on the true photon count rate S from the source?

If the distribution of photon counts from sources in the Universe were uniform (uniform prior), the photon-counting uncertainty would symmetrically scatter values on either side of the population mean $\Longrightarrow$ 95% CI from data nicely constrains true count rate.
However, there are way more faint sources in the Universe.
e.g., in Euclidean space, $\dfrac{dN}{dS} \equiv p(S) \propto S^{-5/2}$ (steep prior, small intersection with likelihood).

$\Longrightarrow$ more likely that a lower photon count gets observed as a higher value due to Poisson uncertainty. This is a form of Eddington Bias.

Prior-dominated posterior (contd.)

Prior: $p(S) \propto S^{-5/2}$.
Likelihood of obtaining data $S_{\mathrm{obs}} = 4$ from Poissonian uncertainties acting on S:
$\mathscr{L}(S) \propto S^{S_{\mathrm{obs}}} \exp\Big[-S\Big] = S^4 \exp\Big[-S\Big]$.
Posterior $p(S|S_{\mathrm{obs}}) \propto S^{3/2} \exp\Big[-S\Big]$

$$= \mathrm{Gamma}\left(\frac{5}{2}, 1\right).$$

$\Longrightarrow$ Mean: 5/2; Mode: 3/2. Can also compute HPD (homework).