Statistics for Astronomers: Lecture 19, 2019.05.08

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: Data Visualisation and Nonparametric Density Estimation

1. If you're only interested in the general trend in your data, use box plots. They'll also immediately identify outliers!
2. Histograms are fast but bad for various reasons – their shapes depend on bin size and bin location, and they degrade the information contained in the raw data.
3. There are ways to figure out the optimum bin size – both frequentist and Bayesian. The Bayesian versions are more sensitive to multimodal distributions, and allow for the computation of the optimum bin size without as few assumptions on the underlying distribution as possible.
4. The Bayesian Blocks method allows for variable bin size! It is especially applicable for small data sizes.
5. If you're really interested in generating a function that mimics the true population distribution, use KDEs.

Sampling techniques

Error propagation

Formulae for error propagation are based on linear approximations of Taylor Series. These fail for large fractional uncertainties.

Example:
IRAC 8 μm observation of a target results in $m = 7.27 \pm 0.54$ mag. What is the relative uncertainty in the flux?

$$F \propto \exp\left[-\frac{m}{\alpha}\right], \text{ with } \alpha = \frac{2.5}{\ln 10} \implies \frac{\Delta_F}{F} = \exp\left[-\frac{\Delta_m}{\alpha}\right] = 0.64.$$

Using just the first term in the Taylor series expansion instead,

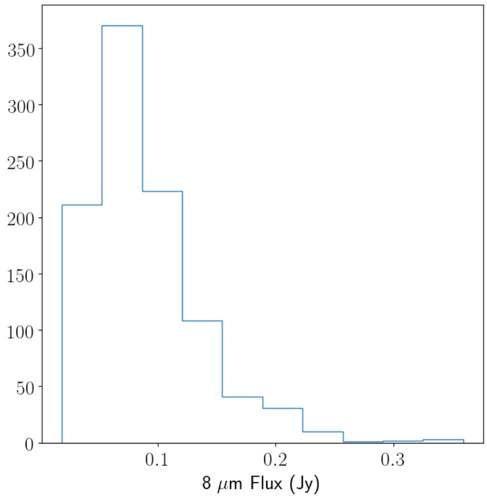
$$\frac{\Delta_F}{F} = \frac{\Delta_m}{\alpha} = 0.50.$$

Assumption: resulting distribution for F is symmetric so a standard deviation can be defined in the usual manner.

Error propagation (contd.)

Alternative method: Monte Carlo sampling.

Assume: magnitude $\sim \mathcal{N}(7.27, 0.54^2)$. Draw from this distribution $N = 1000$ times and compute the flux from each draw. Use the resulting distribution to estimate the location and scale parameters.



Distribution is severely skewed.
(Mean, median, mode) = (0.09, 0.08, 0.03) Jy.

Scale has to be evaluated in one of many ways (e.g., equal-tailed interval).

This method is extremely useful when (a) many errors have to be simultaneously propagated and/or (b) the relationship between the variables is nonlinear (*e.g.*, the blackbody flux in terms of its parameters).

Quadrature (Deterministic methods)

Based on weighted averages *e.g.*: Trapezoid Rule, Simpson's Rule, Gauss quadrature.

$$\int_a^b dx\ f(x) = \frac{b-a}{N}\sum_{i=1}^{N} w(x_i) f(x_i).$$

0^{th} order: $f(x)$ piecewise constant ($w(x_i) = 1$ above). Error $\sim N^{-1}$.
1^{st} order: $f(x)$ piecewise linear (Trapezoid Rule). Error $\sim N^{-2}$.
2^{nd} order: $f(x)$ piecewise quadratic (*e.g.*, Simpson's Rule). Error $\sim N^{-4}$.

d dimensions: error $\sim$ (one-dimensional error)$^{1/d}$. More efficient techniques required!

Quadrature (Stochastic/numerical methods)

Based on random draws *e.g.*: Monte Carlo, Markov Chain Monte Carlo.

Let $x \sim p(x)$. Suppose we want to evaluate $\mathbb{E}[g(x)] = \int dx\ p(x)\ g(x)$.
The simple Monte Carlo method is as follows:
Draw samples of x from $p(x)$, evaluate $g(x)$ for these x values, and approximate $\mathbb{E}[g(x)]$ with the average of these $g(x)$ values: $\mathbb{E}[g(x)] = \int dx\ p(x)\ g(x) \approx \frac{1}{N}\sum_{i=1}^{N} g(x_i)$.

Note that $\int_a^b dx\ g(x) = (b-a)\int_a^b dx\ p(x)\ g(x)$, where $p(x) = U[a,b] = \frac{1}{b-a}$.

Error $\sim \sqrt{\frac{Var(g(x))}{N}} \propto N^{-1/2}$.

Simple Monte Carlo not very efficient! Better than Simpson only for $d > 8$!! For comparable errors, more function evaluations required in the simple MC case.

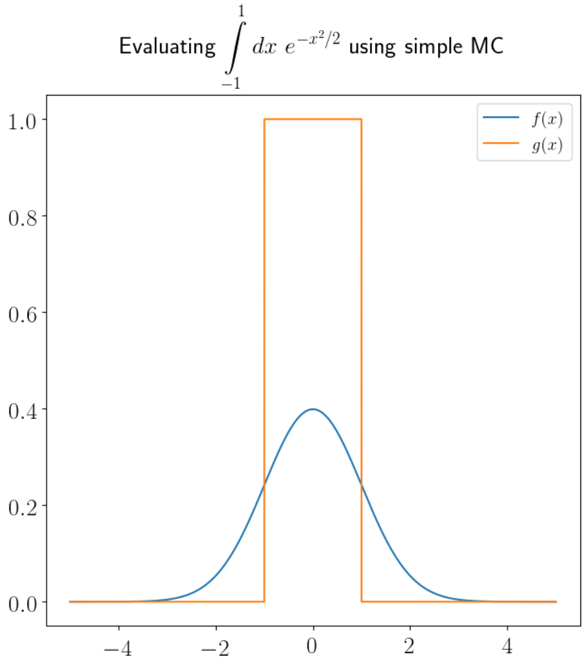
Example using Simple MC

For $X \sim \mathcal{N}(0,1)$,
compute $p(-1 < x < 1)$.

$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $g(x) = \mathbb{I}_{x\in[-1,1]}(x)$.

Exact value: $\Phi(+1) - \Phi(-1)$.

Report/plot absolute relative error between computed value and true value for
$N = 10, 100, 1000, 10000, 1000000$.



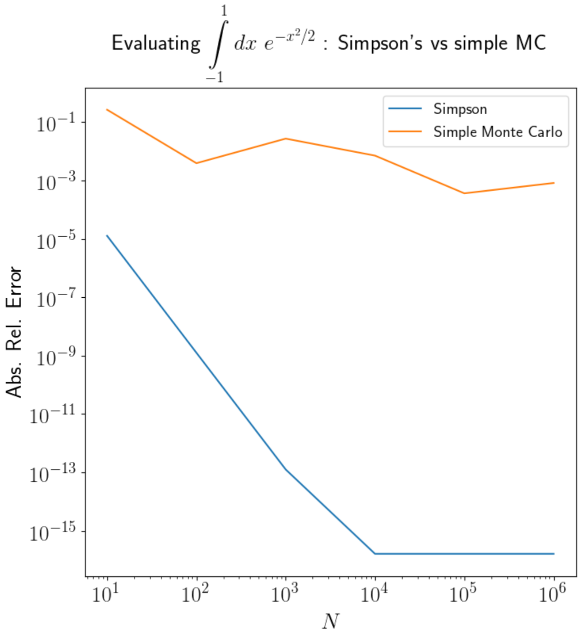
Example using Simple MC

For $X \sim \mathcal{N}(0,1)$,
compute $p(-1 < x < 1)$.

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \; g(x) = \mathbb{I}_{x \in [-1,1]}(x).$$

Exact value: $\Phi(+1) - \Phi(-1)$.

Report/plot absolute relative error between computed value and true value for
$N = 10, 100, 1000, 10000, 1000000$.



Example using Simple MC: another way

Note that the "mask function" $g(x)$ on the previous slide is nothing but $U[-1,1]$.

Let's try the following steps:
Generate a pair of variables X, Y such that $X \sim U[-1,1]$ (the range of x values for which $p(x)$ is desired) and $Y \sim U[0,1]$ (the range of heights of $g(x)$).
Reject pairs with $y > f(x)$, accept pairs with $y \leq f(x)$.

What is the ratio of accepted pairs to total pairs?

This is another way to evaluate the integral, called rejection sampling.

Rejection sampling

Rejection sampling samples from a proposal distribution $g(x)$ instead of the target distribution $p(x)$.

$g(x)$ is such that for some $M > 1, f(x) \leq M\ g(x)$.

The general procedure for rejection sampling is as follows

1. Sample an x value from the proposal distribution.
2. For this x value, sample a y value from $U[0, g(x)]$ (that is, find a height that is between zero and the value of the proposal distribution at this x value).
3. If the sampled $y \leq f(x)$ for the corresponding x value, accept this x value. If not, reject it and go back to step 1.

The fraction $\nu = \dfrac{N_{\text{accepted}}}{N_{\text{total}}}$ of accepted values is such that $\displaystyle\int dx\ f(x) = \nu \int dx\ g(x)$.

Another example of rejection sampling: computing the value of π using a circle inscribed in a square.

Future