Statistical Data Analysis
Discussion notes – week 3

- Lack of memory of the exponential distribution
- The log-normal distribution
- Generating random values from the Cauchy pdf
- Importance sampling

Example 1: "memorylessness" of exponential

Exponential pdf $f(x;\xi) = \frac{1}{\xi} e^{-x/\xi}, \quad x \geq 0$

First, find cumulative distribution

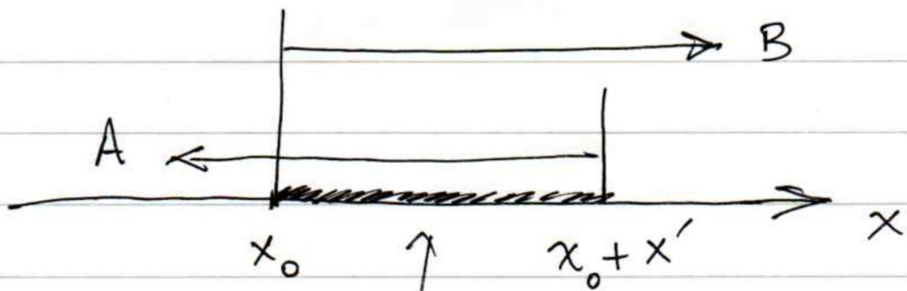
$$F(x) = \int_0^x \frac{1}{\xi} e^{-x'/\xi} dx' = -e^{-x'/\xi}\Big|_0^x = 1 - e^{-x/\xi}$$

Next, find $P(x < x_0 + x' \mid x > x_0)$

will show this is $P(x < x')$

Recall $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$

For $P(x < x_0 + x' \,|\, x > x_0)$



$A \cap B = x_0 < x < x_0 + x'$

$$\Rightarrow P(x < x_0 + x' \,|\, x > x_0) = \frac{P(x_0 < x < x_0 + x')}{P(x > x_0)}$$

$$= \frac{\int_{x_0}^{x_0+x'} \frac{1}{\xi} e^{-x/\xi} \, dx}{\int_{x_0}^{\infty} \frac{1}{\xi} e^{-x/\xi} \, dx} = \frac{F(x_0 + x') - F(x_0)}{1 - F(x_0)}$$

$$F(x_0) = 1 - e^{-x_0/\xi}$$

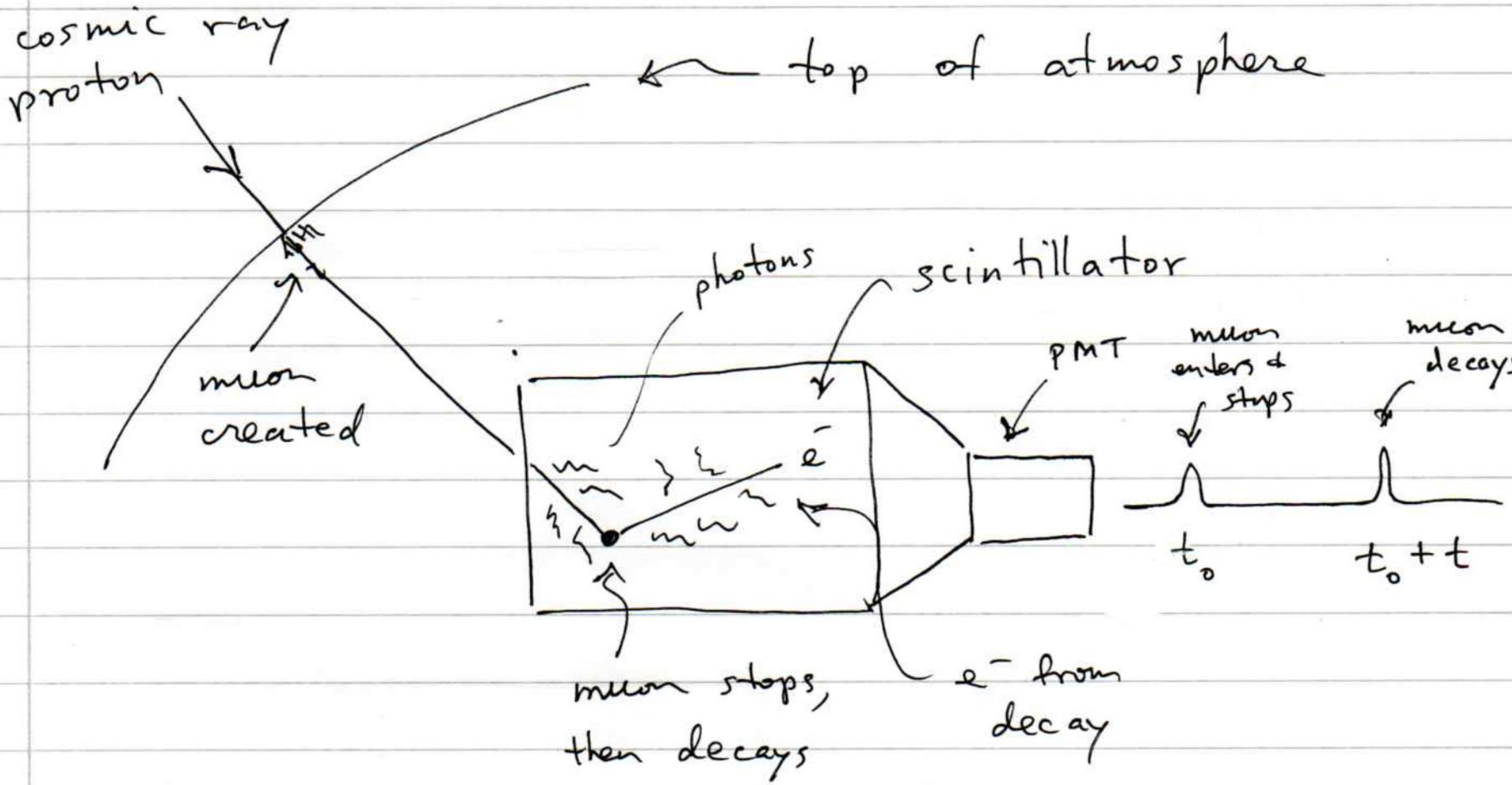
$$= \frac{-e^{-(x_0+x')/\xi} + e^{-x_0/\xi}}{e^{-x_0/\xi}}$$

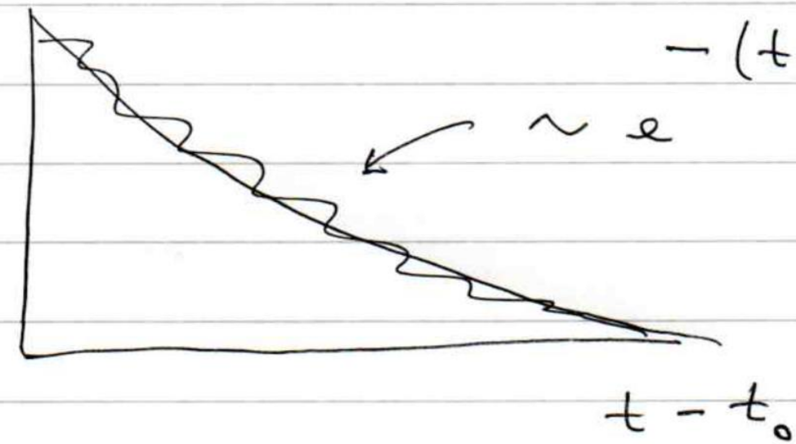
$$= 1 - e^{-x'/\xi} = F(x') = P(x \leq x')$$

And from this using $f(x) = \frac{\partial F}{\partial x}$

$$f(x - x_0 \mid x > x_0) = f(x)$$

Example "memoryless" exponential





$\sim e^{-(t-t_0)/\tau}$

$\tau = 2.2\ \mu s$ = mean lifetime of muon

Time that muon lived before t_0 is irrelevant. Muon is just as "young" at t_0 as when it was first born:

$$f(t - t_0 \mid t > t_0) = f(t)$$

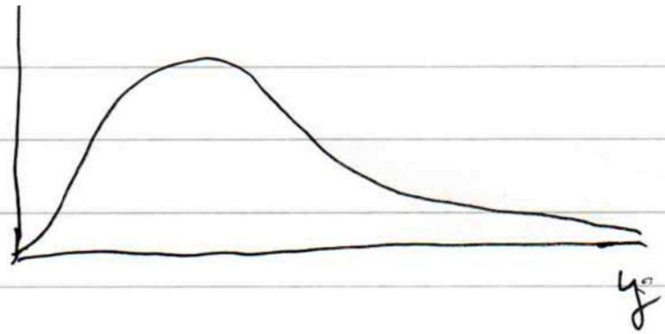
Example 3 Log-normal dist. & variable trans.

Gaussian $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2}$

Let $y = e^x$ & find pdf of y

$$x = \ln y \,, \qquad \frac{dx}{dy} = \frac{1}{y}$$

$$g(y) = f(x(y)) \left| \frac{dx}{dy} \right| = \frac{1}{\sqrt{2\pi}\,\sigma\, y} \exp\left[- \frac{(\ln y - \mu)^2}{2\sigma^2} \right]$$



$(y > 0)$

μ, σ^2 are mean, variance of Gaussian x, not of the log-normal y. Can find

$$E[y] = \exp\left[\mu + \frac{\sigma^2}{2}\right], \quad V[y] = \left[e^{\sigma^2} - 1\right] \exp(2\mu + \sigma^2)$$

$$x = \sum_{i=1}^{\text{many}} u_i \xrightarrow{\text{CLT}} x \sim \text{Gauss}$$

$$y = e^x = \exp\left[\sum_i u_i\right] = \prod e^{u_i} \xrightarrow{\text{CLT}} \text{log-normal}$$

<u>Sum</u> of many terms $\xrightarrow{\text{CLT}}$ Gauss

Product " " factors $\xrightarrow{\text{CLT}}$ log-normal

Random numbers from the Cauchy pdf

Cauchy pdf $\quad f(x) = \frac{1}{\pi} \frac{1}{1+x^2}$

Cumulative dist. $\quad F(x) = \int_{-\infty}^{x} \frac{dx'}{\pi(1+x'^2)}$

$\Rightarrow F(x) = \frac{1}{\pi} \tan^{-1} x' \Big|_{-\infty}^{x}$

$= \frac{1}{\pi} \left(\tan^{-1} x + \frac{\pi}{2} \right)$

$\overset{\text{set}}{=} \; r \qquad$ and solve ~~for~~ x

$r \sim U[0,1]$

$\Rightarrow \quad x(r) = \tan\left[\pi\left(r - \frac{1}{2}\right)\right]$

i.e. if $r_1, r_2, \ldots$ indep. & $\sim U[0,1]$

then $x_i = x(r_i)$ indep & $\sim \dfrac{1}{\pi(1+x^2)}$

Code: cauchyMC.py

cauchyMC.ipynb

cauchyMC.py, .ipynb

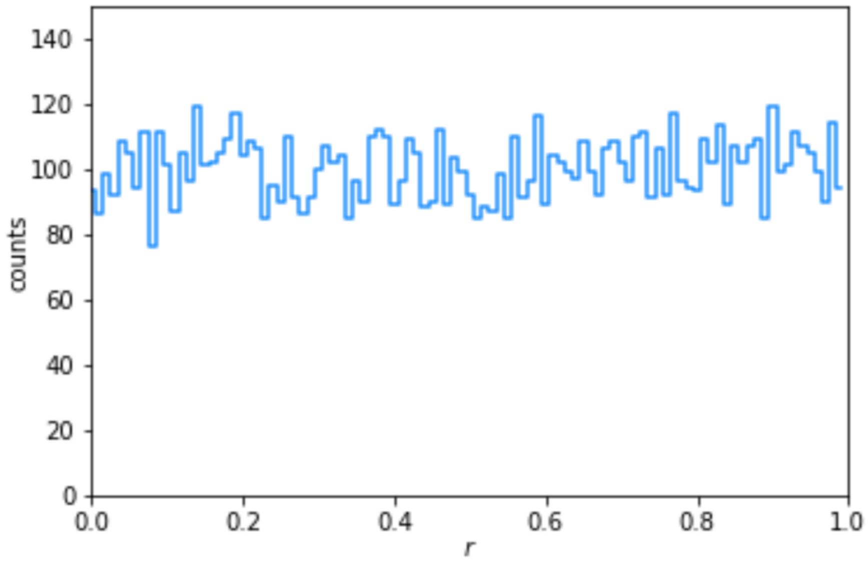
```
# cauchMC.py
# simple Monte Carlo program to make histogram of uniform and Cauchy
# distributed random values and plot
# G. Cowan, RHUL Physics, updated October 2024

import matplotlib.pyplot as plt
import numpy as np

# Set random seed and other parameters
np.random.seed(12345)
numVal = 10000
nBins = 100

# Generate uniformly distributed numbers
rMin, rMax = 0., 1.
rData = np.random.uniform(rMin, rMax, numVal)

# Using transformation method, generate Cauchy distributed numbers
xMin, xMax = -10., 10.
xData = np.tan(np.pi * (rData - 0.5))
```

cauchyMC.py, .ipynb (continued)

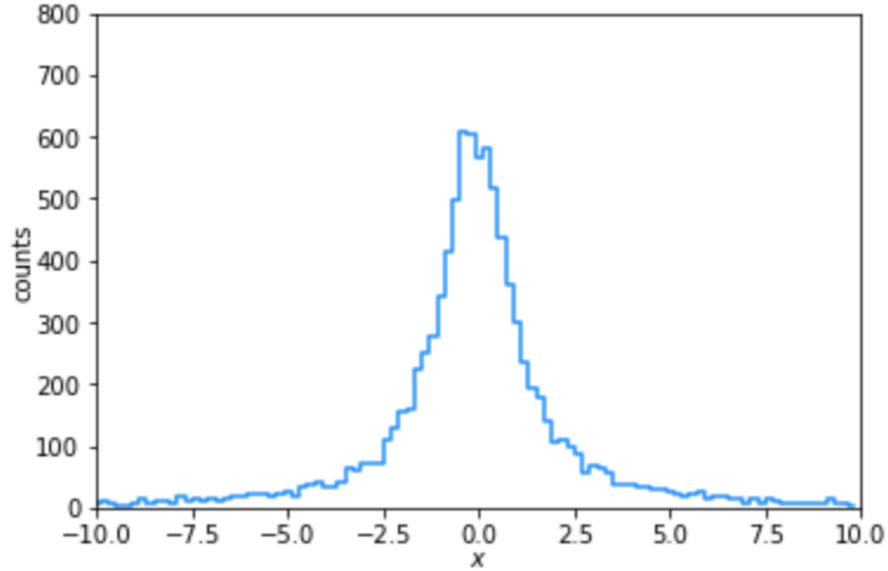
```python
# Define a function for plotting histograms
def plot_histogram(data, nBins, data_range, xlabel, ylabel, ylim, filename):
    fig, ax = plt.subplots()
    hist, bin_edges = np.histogram(data, bins=nBins, range=data_range)
    ax.step(bin_edges[:-1], hist, where='mid', linewidth=1.5,
            color='dodgerblue')
    ax.set_xlim(data_range)
    ax.set_ylim(ylim)
    ax.set_xlabel(xlabel, labelpad=0)
    ax.set_ylabel(ylabel, labelpad=0)
    fig.subplots_adjust(bottom=0.15, left=0.15)    # Adjust layout
    plt.savefig(filename, format='pdf')            # Save and show the plot
    plt.show()

# Plot and save histograms
plot_histogram(rData, nBins, (rMin, rMax), r'$r$', 'counts',
               (0, 150), "uniform_histogram.pdf")
plot_histogram(xData, nBins, (xMin, xMax), r'$x$', 'counts',
               (0, 800), "cauchy_histogram.pdf")
```

cauchyMC.py, .ipynb (output)

Uniform



Cauchy



Importance Sampling

Often the goal of an MC calculation is to estimate the mean value of a function $h(x)$, where $x \sim f(x)$.

$$E_f[h(x)] = \int h(x) f(x) \, dx \equiv \mu$$

An estimator $\hat{\mu}_{\rm MC}$ for μ is the average of N values of $h(x)$ with x generated from $f(x)$:

$$\hat{\mu}_{\rm MC} = \frac{1}{N} \sum_{i=1}^{N} h(x_i)$$

This has a variance of

$$V[\hat{\mu}_{\rm MC}] = \frac{1}{N} V_f[h(x)] = \frac{1}{N} \left(E_f[h^2(x)] - \mu^2 \right)$$

Importance Sampling (2)

With importance sampling, one can reduce the variance for a given number of generated random values N.

Rewrite the desired mean value as

$$\mu = \int h(x) f(x) \, dx = \int \frac{h(x) f(x)}{g(x)} \, g(x) \, dx = E_g \left[\frac{h(x) f(x)}{g(x)} \right]$$

where $g(x)$ is any other pdf nonzero on the same interval as $f(x)$ from which we can generate random values.

Thus μ is the expectation with respect to g of $h(x)f(x)/g(x)$, and can be estimated by generating N values of $x \sim g(x)$ and using

$$\hat{\mu}_{\mathrm{IS}} = \frac{1}{N} \sum_{i=1}^{N} \frac{h(x_i) f(x_i)}{g(x_i)}$$

Importance Sampling (3)

The variance of $\hat{\mu}_{\rm IS}$ is

$$V[\hat{\mu}_{\rm IS}] = \frac{1}{N} V_g \left[\frac{h(x) f(x)}{g(x)} \right] = \frac{1}{N} \left(E_g \left[\frac{h^2(x) f^2(x)}{g^2(x)} \right] - \mu^2 \right)$$

By choosing $g(x)$ such that $h(x)f(x)/g(x)$ is as constant as possible, one can minimize the variance of $\hat{\mu}_{\rm IS}$.

Minimum achieved when: $g(x) \propto |h(x)| f(x)$

Alternative estimator (smaller variance at cost of small bias):

$$\hat{\mu}_{\rm IS} = \frac{\sum_{i=1}^{N} \frac{h(x_i) f(x_i)}{g(x_i)}}{\sum_{i=1}^{N} \frac{f(x_i)}{g(x_i)}}$$

References:

C.P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed., (Springer, New York, 2004).

J.S. Liu, *Monte Carlo Strategies in Scientific Computing*, (Springer, New York, 2001).

Importance Sampling Example

We can generate uniform random numbers with the pdf:

$$f(x) = 1\,, \qquad 0 \le x \le 1$$

Suppose we want the mean value on [0,1] of: $h(x) = x^3 e^x$

$$E_f[h(x)] = \int h(x) f(x)\, dx = \int_0^1 x^3 e^x \cdot 1\, dx$$

In practice we don't know the exact result.

$$= e^x (x^3 - 3x^2 + 6x - 6)\Big|_0^1 = 6 - 2e = 0.563436$$

We need a pdf $g(x)$ from which we can sample such that

$$\frac{h(x) f(x)}{g(x)}$$

is approximately constant on the relevant interval $0 < x < 1$.

Importance Sampling Example (2)

Guess: $g(x) = (\theta + 1)x^\theta$ for some θ that we can adjust.

The cumulative distribution is $G(x) = \int_0^x g(x')\,dx' = x^{\theta+1}$

To sample from $g(x)$, set $g(x) = r$ where $r \sim \mathrm{U}[0,1]$ and solve for x: $x(r) = r^{1/(\theta+1)}$

For e.g. $\theta = 3.4$ and $N = 1000$,

Exact value = 0.563436

Monte Carlo estimate:
0.525473 +/- 0.022596

Importance sampling estimate:
0.564376 +/- 0.001159

20 times smaller err. than MC