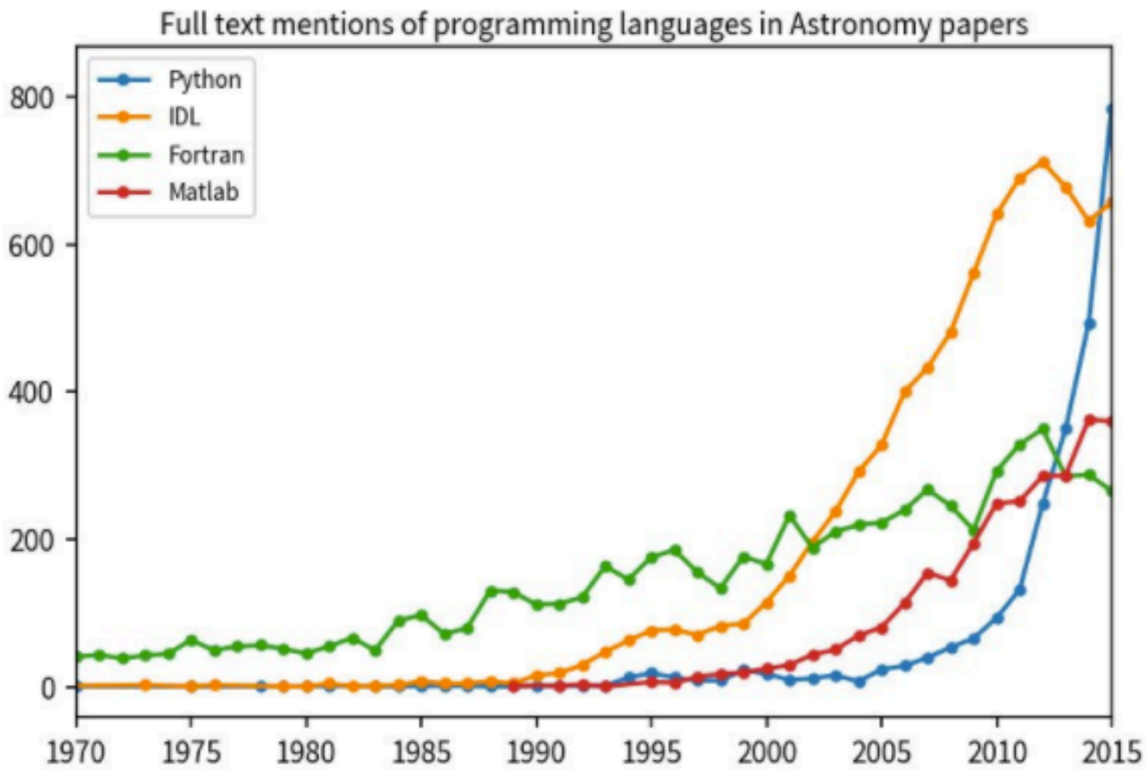
Trend of Programming languages in Astro literature



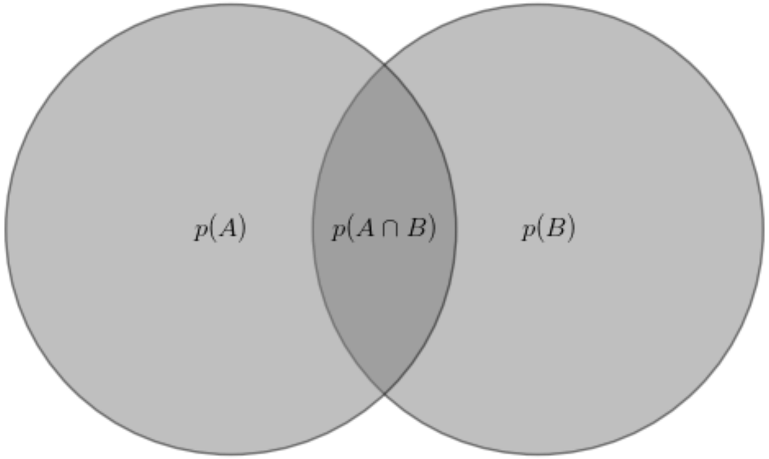
From https://twitter.com/astrofrog/status/787007261877166080

does not include R and C (as difficult to parse texts)

Statistical Data Analysis tasks in Astronomy

Photometric Redshifts (Regression)
Source Classification
Dimensionality Reduction/Visualization
Clustering
N-point statistics
Period Finding
Transient and Outlier Detection
Density Estimation
Matched Filtering
Source Extraction
Cross-Matching

Probability Axioms



$p(A \cup B) = p(A) + p(B) - p(A \cap B)$

P(A) is a probability if it satisfies three axioms:

P(A) > 0
Sum P(A) = 1 (for all possible outcomes)
For disjoint events A_1 A_2, etc

$$p(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} p(A_i)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Probability for both A and B to happen = $P(A \cap B)$

$$P(A \cap B) = P(A/B)P(B) = P(B/A)P(A)$$

Bayes Theorem

$P(A/B)$ is conditional probability of event A given that (on conditioned on) B has occurred

If events B_i i=1…..N are disjoint and union is set of all possible outcomes then

$$P(A) = \sum_{i=1}^{N} P(A \cap B_i) = \sum_{i=1}^{N} P(A|B_i)P(B_i)$$

This is called "Law of Total Probability"

Conditional probabilities also satisfy law of total probability. Assuming an event C_i is not mutually exclusive with A or any of the B_i then

$$P(A|B) = \sum_i P(A|B \cap C_i)P(C_i|B)$$

Probability rules were derived from two different sets of axioms by Cox and Kolmogorov (Jaynes)

If events C_i i=1…..N are disjoint and union is set of all possible outcomes then

$$P(A \cap B) = \sum_i P(A \cap B \cap C_i)$$

$$P(A \cap B) = \sum_i P(A|B \cap C_i) P(B \cap C_i)$$

$$P(A \cap B) = \sum_i P(A|B \cap C_i) P(C_i|B) P(B)$$

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

$$P(A|B) = \frac{\sum_i P(A|B \cap C_i) P(C_i|B) P(B)}{P(B)}$$

$$P(A|B) = \sum_i P(A|B \cap C_i) P(C_i|B)$$

Random Variables

- A random variables is a variable whose value results from the measurement of a quantity subject to stochastic variations.

- Independent identically distributed random variables are drawn from the same distribution and independent.

  Two random variables x and y are *independent* if and only if

  $P(x,y) = P(x)\ P(y)$

Conditional Probability and Bayes Rule

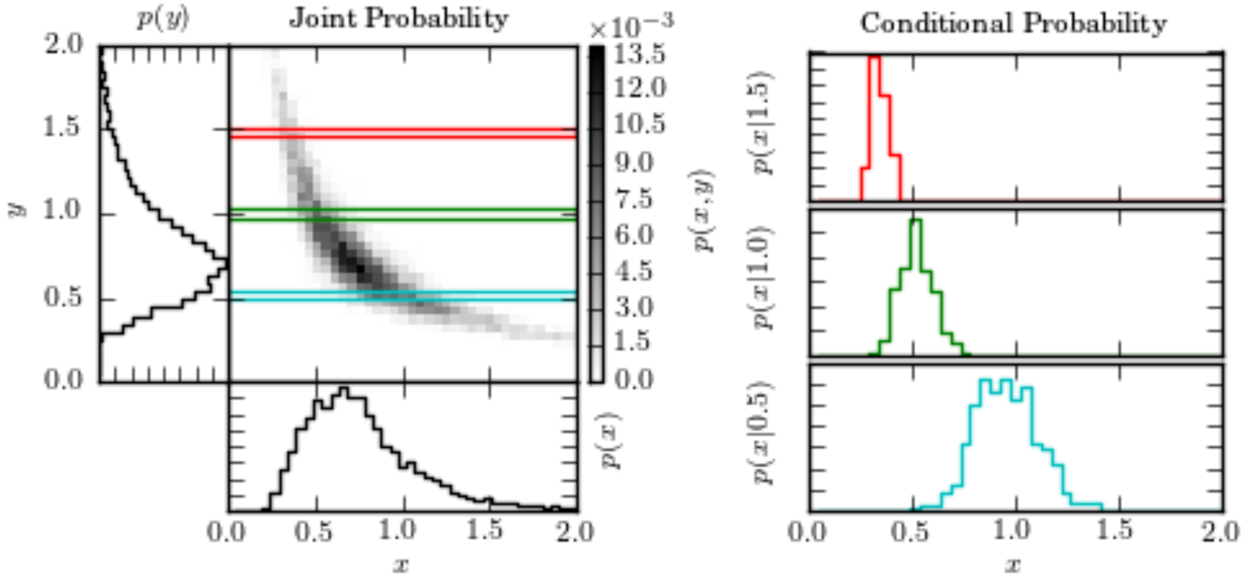
If two continuous random variables are not independent, it follows that

$$p(x,y) = p(x|y)p(y) = p(y|x)\,p(x)$$

Marginal Probability function defined as

$p(x) = \int p(x,y)dy$ By combining above two equations we get

$$p(x) = \int p(x|y)p(y)dy$$



$P(x|y=y_0)$ are one-dimensional "slices" through the two-dimensional Image $p(x,y)$ at given values of y_0 divided by the value of marginal distribution $p(y)$

$\int P(x)\,dx = 1$

Code to reproduce this available from astroml.org website

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

For a discrete random variable y_j with M possible values the above integral becomes a sum:

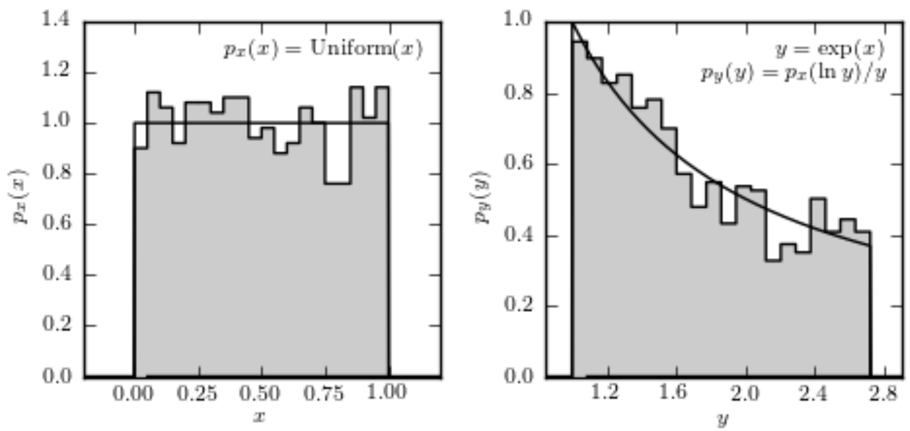
$$p(y_j|x) = \frac{p(x|y_j)p(y_j)}{\sum_{j=1}^{M} p(x|y_j)p(y_j)}$$

Homework : Read about Monty Hall problem

Transformations of Random Variables

Any function of a random variable x y = ϕ(x) is a random variable. We can calculate p(y) from p(x) as follows :

$$p(y) = p[\Phi^{-1}(y)] \left| \frac{d\Phi^{-1}(y)}{dy} \right|$$



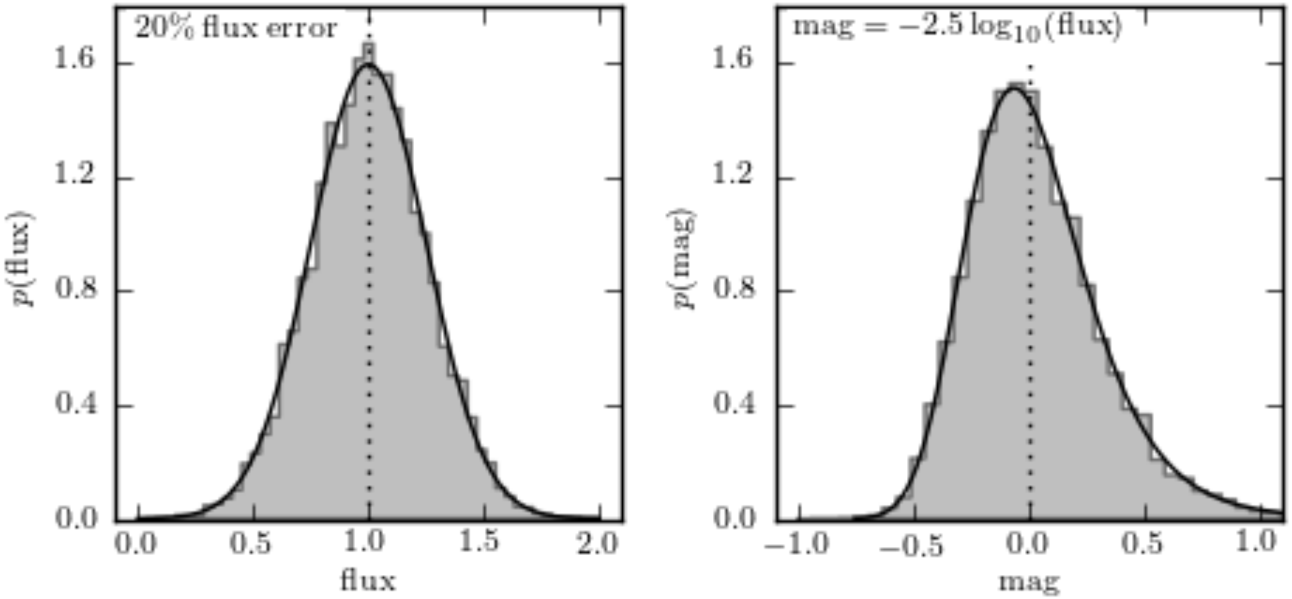
Cumulative statistics such as medians do not change their order under monotonic transformations

If uncertainty in x at a given value of x_0 is given by σ_x then we can use Taylor series expansion to Estimate the uncertainty in y at $y_0 = \phi(x_0)$

$$\sigma_y = \left| \frac{d\Phi(x)}{dx} \right|_0 \sigma_x$$

Sometimes this can lead to misleading results for non-linear transformations
for example (in Astronomy) magnitude = -2.5log(Flux)

Example :



20% flux error
1.6
1.2
0.8
0.4
0.0
p(flux)
0.0
0.5
1.0
1.5
2.0
flux
mag = −2.5 log10(flux)
1.6
1.2
0.8
0.4
0.0
p(mag)
−1.0
−0.5
0.0
0.5
1.0
mag

Error Propagation (Without Covariances)

Consider G= $G(x_1,x_2,\ldots x_n)$ with uncertainties $\boldsymbol{\sigma}_1\ \boldsymbol{\sigma}_2\ldots\ldots\ \boldsymbol{\sigma}_n$

$$\sigma_G^2 = \sum_{i=1}^{N} \left(\frac{\partial G}{\partial x_i} \right)^2 \sigma_{x_i}^2$$

Iff the errors in x_1,x_2 … are uncorrelated

Eg. $\Delta m = m_1 - m_2$

$\boldsymbol{\sigma}^2{}_{\Delta m} = \boldsymbol{\sigma}_{m1}{}^2 + \boldsymbol{\sigma}_{m2}{}^2$

Error Propagation (with Covariances)

$$\sigma_G^2 = \sum_{i=1}^{N} \left(\frac{\partial G}{\partial x_i} \right)^2 \sigma_{x_i}^2 + 2\sigma_{x_1 x_2}^2 \frac{\partial G}{\partial x_1} \frac{\partial G}{\partial x_2} + ...$$

where

$$\sigma_{x_1 x_2}^2 = \frac{\sum_{i=1}^{N} [(x_{1i} - \bar{x_1})][(x_{2i} - \bar{x_2})]}{N}$$

Ref. Bevington's book

Descriptive Statistics

An arbitrary distribution $h(x)$ is characterized by its location parameters, scale or width parameters and "shape" parameters.
When they are based on the distribution $h(x)$ they are called *population* statistics. If they are based upon a finite-sized dataset, they are called *sample* statistics.

Arithmetic mean based upon expectation value

$$\mu = E(x) = \int_{-\infty}^{+\infty} x h(x)$$

Variance

$$V = \int_{-\infty}^{+\infty} (x-\mu)^2 h(x) dx$$

Standard Deviation

$$\sigma = \sqrt{V}$$

Skewness

$$\Sigma = \int_{-\infty}^{+\infty} \left(\frac{x-\mu}{\sigma}\right)^3 h(x)dx$$

Kurtosis

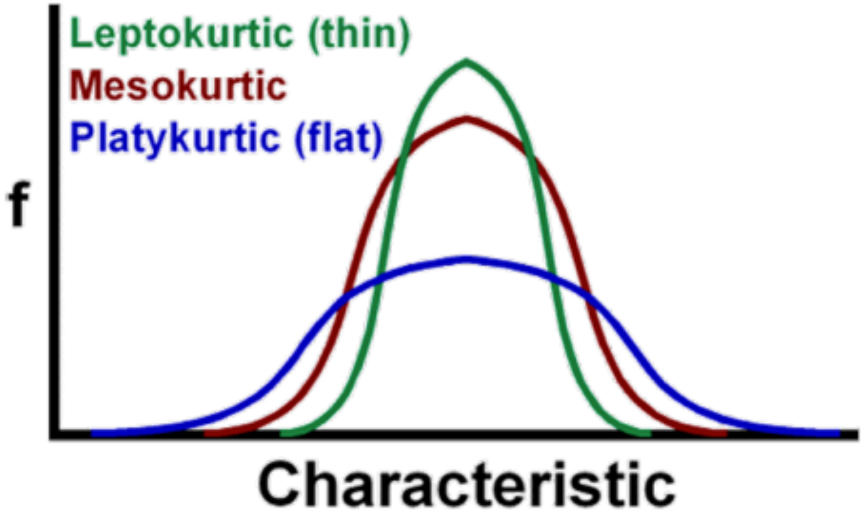
$$K = \int_{-\infty}^{+\infty} \left(\frac{x-\mu}{\sigma}\right)^4 h(x)dx - 3$$

Variance, skewness., kurtosis related to k^{th} central moment of a distribution (k=2,3,4)

Kurtosis

- Measure of the peakedness of the pdf. Describes the shape of the r,v.

$$Kurtosis = \frac{E(X-\mu)^4}{\sigma^4} = \frac{\mu_4}{\mu_2^2}$$



Kurtosis=3 $\rightarrow$ Normal

Kurtosis >3 $\rightarrow$ Leptokurtic
(peaked and fat tails)

Kurtosis<3 $\rightarrow$ Platykurtic
(less peaked and thinner tails)

Moments From Generating Function

Ref: arXiv:0712.3028

Generating Function allows you to calculate the moments of a distribution

$$Z(k) = \langle exp(ikx) \rangle = \int exp(ikx) P(x) dx$$

This can be written as an infinite series by expanding the exponential giving :

$$Z(k) = \sum_{n=0}^{\infty} \frac{(ik)^n}{n!} \hat{\mu}_n$$

$$\hat{\mu}_n = (-i^n) \frac{d^n}{dk^n} Z(k)|_{k=0}$$

Absolute Deviation about d

$$\delta = \int_{-\infty}^{+\infty} |x - d| h(x) dx$$

Absolute deviation about the mean (d=mean(x)) is called mean deviation

Mode (or most probable value in case of unimodal functions) x_m

$$\left(\frac{dh(x)}{dx} \right)_{x_m} = 0$$

P % quantiles (or p percentiles)

$$\frac{p}{100} = \int_{-\infty}^{q_p} h(x) dx$$

All the moments are built into NumPy and SciPy. Useful functions are

```
numpy.median, numpy.mean, numpy.var
numpy.percentile, numpy.std,scipy.stats.skew,
scipy.stats.kurtosis,
scipy.stats.mode
```

```
import numpy as np
x = np.random.random(100)
q25,q50,q75 = np.percentile(x,[25,50,75])
```

Difference between third and first quartile is called interquartile range

A useful relation between mode, median and mean valid for mildy non-gaussian distributions

$$\boxed{\text{Mode} = 3\,(\text{median}) - 2(\text{mean})}$$

Data-Based Estimates of Descriptive Statistics

If the above quantities are derived from data, they are called sample statistics (instead of population Statistics).

Assume we have N given measurements x_i for i=1,........N abbreviated as $\{x_i\}$

For a sample of N measurements

$$\int_{-\infty}^{+\infty} g(x)h(x)dx \equiv (1/N)\sum_{i=1}^{N} g(x_i)$$

Sample arithmetic mean and standard deviation given by :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

If the samples have an error σ_i mean and error in mean are given by:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i/\sigma_i^2}{\sum_{i=1}^{N} 1/\sigma_i^2} \qquad \sigma_{\bar{x}}^2 = \frac{1}{\sum_{i=1}^{N}(1/\sigma_i^2)}$$

Sample Standard deviation is calculated as follows:

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

N-1 is used so that variance is unbiased

Uncertainty in the standard mean is given by :

(if errors in each data point are equal)

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{N}}$$

$$\sigma_s = \frac{s}{\sqrt{2(N-1)}}$$

For real data with outliers, calculation of s from data samples can lead to wrong estimates

Median Absolute Deviation (wikipedia)

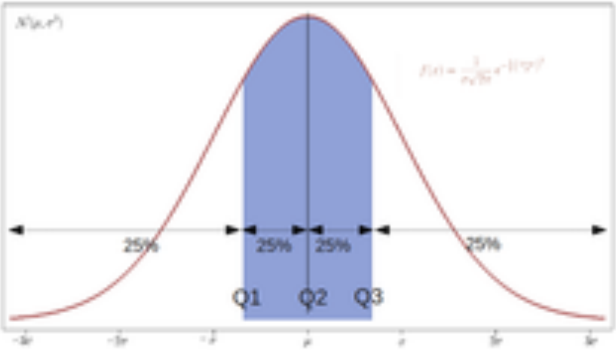
$$MAD = median(|x_i - median(x)|)$$

For a normal distribution, $\sigma = 1.482$ MAD where σ is the std deviation of Gaussian distribution

Alternately, you can use a rank based estimate of the standard deviation. Inter-quartile range $(q_{75} - q_{25})$ is a more robust estimator of the scale parameter than the standard deviation.

Even in absence of outliers for some distributions which do not have finite variance such as the Cauchy distribution, the median and interquartile range are the best choices for estimating the location and scale parameters.

For a Gaussian, $\sigma_G = 0.7413(q_{75} - q_{25})$ (σ_G can be computed with astroML library)



Quantiles

Source : wikipedia

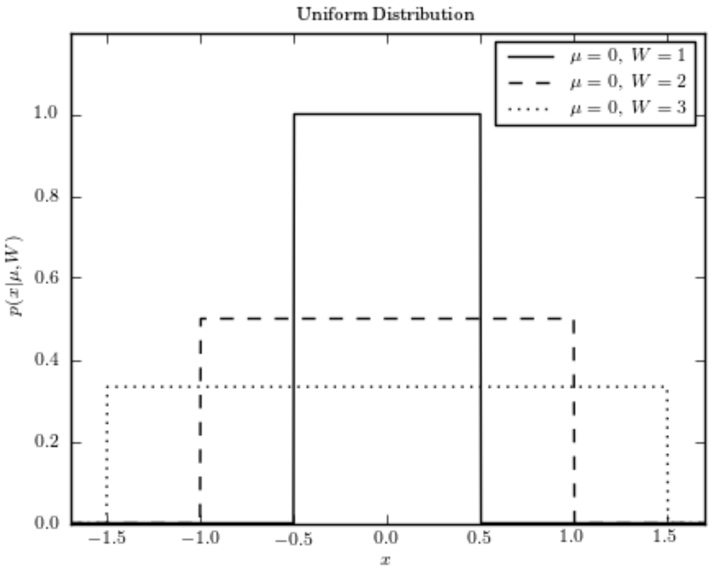
Specialized quantiles [edit]

Some q-quantiles have special names:[citation needed]

- The only 2-quantile is called the median
- The 3-quantiles are called tertiles or terciles → T
- The 4-quantiles are called quartiles → Q; the difference between upper and lower quartiles is also called the interquartile range, **midspread** or **middle fifty** → IQR = $Q_3 - Q_1$
- The 5-quantiles are called quintiles → QU
- The 6-quantiles are called sextiles → S
- The 7-quantiles are called septiles
- The 8-quantiles are called octiles → O

Examples of Distribution Functions

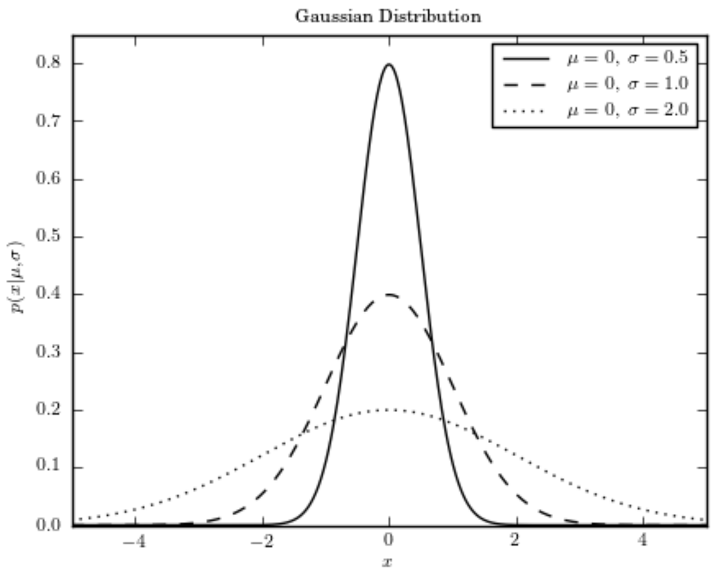
Uniform Distribution




```
from scipy import stats
dist = stats.uniform(0,2)
# left edge at 0 and width=2
r = dist.rvs(10)
# 10 random draws
P = dist.pdf(1)
#PDF evaluated at x=1
Look at scipy.stats page for
more information
```

$P(x|\mu,W) = 1/W$ for $|x - \mu| <= W/2$

Gaussian Distribution




```
dist = stats.norm(0,1)
#mean=0, stddev=1
r = dist.rvs(10)
p = dist.pdf(0)
```

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

Also called Normal distribution $\mathcal{N}(\mu,\sigma)$

Convolution of two Gaussians is also a Gaussian function

$$(f*g)(x) = \int_{-\infty}^{+\infty} f(x')g(x-x')dx' = \int_{-\infty}^{+\infty} f(x-x')g(x')dx'$$

Qt: Consider convolution of two normal distributions : $\mathcal{N}(\mu_0, \sigma_0)$ and $\mathcal{N}(b, \sigma_e)$

What is the mean and std. deviation of the resulting Gaussian?

Cumulative distribution function of a Gaussian distribution is given by :

$$P(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \quad = \frac{1}{2}(1 \pm \mathrm{erf}\left(\frac{|\mathrm{x}-\mu|}{\sqrt{2}\sigma}\right))$$

where erf = Gauss error function

Gauss error function is available in scipy.special

```
from scipy.special import erf
erf(1) = 0.842
```

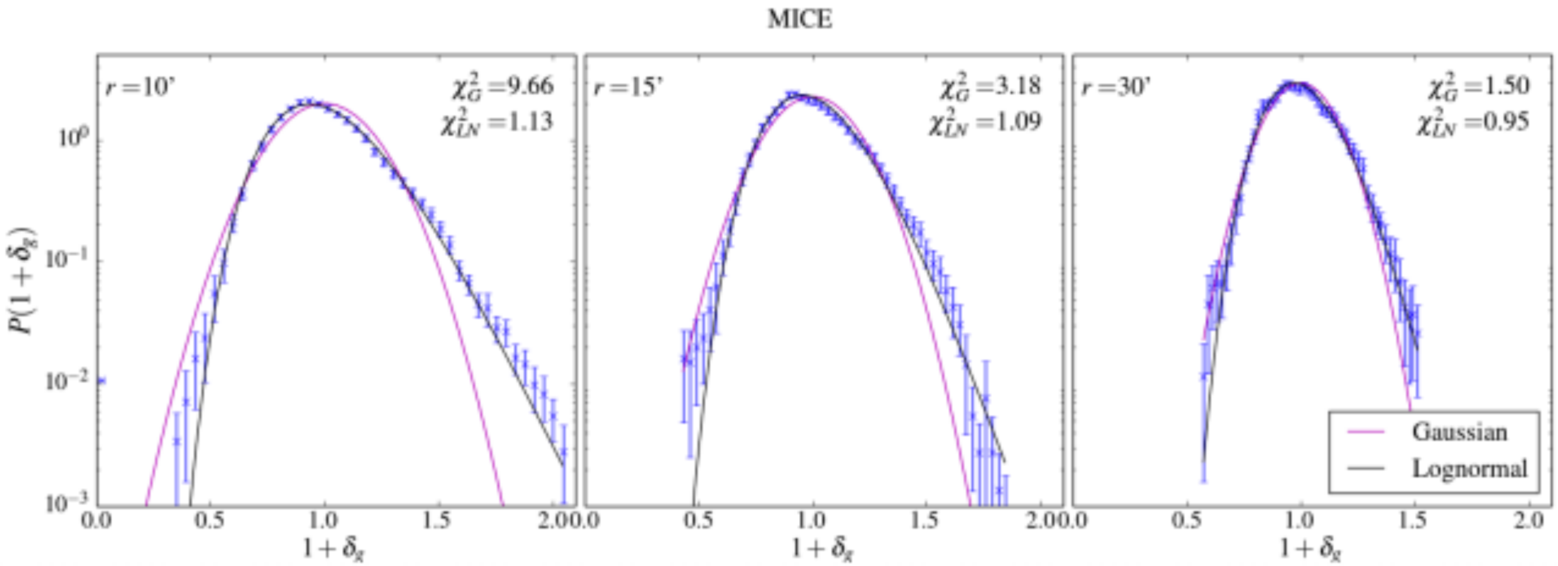
$$\int_a^b p(x|\mu,\sigma)dx = P(b|\mu,\sigma) - P(a|\mu,\sigma)$$

For $a = \mu - M\sigma$ and $b = \mu + M\sigma$, the above integral = erf($M/\sqrt{2}$)

M=1,2,3, give values of 0.68, 0.954, 0.997 respectively for erf($M/\sqrt{2}$)

If x follows a Gaussian distribution, exp(x) follows a log-normal distribution.

LogNormal Distribution



Number distribution of galaxies as a function of density contrast

arxiv:1605.02036

Lognormal Distribution Examples

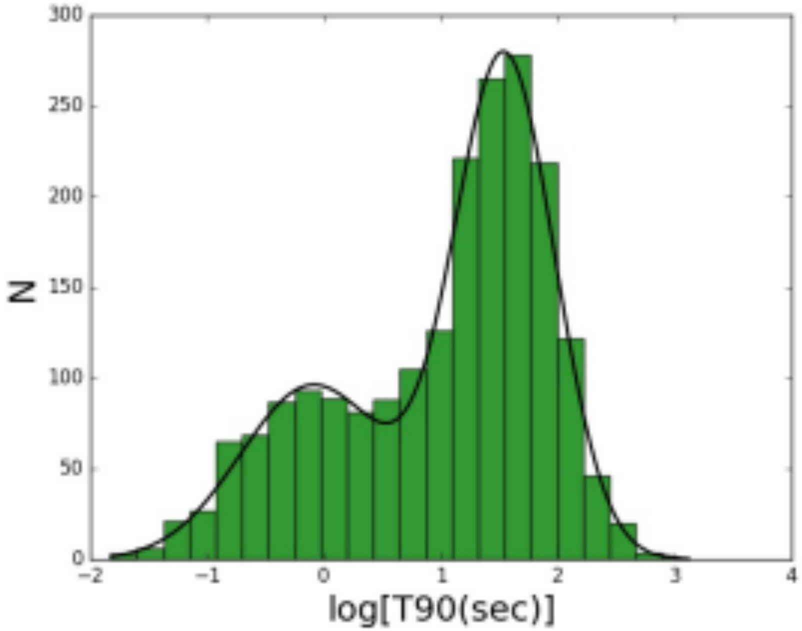


Fig. 1 A fit for the 2-component model for BATSE GRBs.

Gamma-Ray Burst Duration (Soham Kulkarni arXiv:1612.08235)

How to Gaussianize a distribution

Perform a Box-Cox (1964) transformation on the data (arXiv:1508.00931)

$$y_\lambda(a) = \begin{cases} \frac{a^\lambda - 1}{\lambda} & \text{if} \quad \lambda \neq 0 \\ \log a & \text{if} \quad \lambda = 0 \end{cases}$$

$$\bar{y_\lambda} = \sum_{i=1}^{N} \frac{y_\lambda(a_i)}{n}$$

Maximum likelihood estimate of the variance of the transformed data is given by

$$s_\lambda^2 = \sum_{i=1}^{N} \frac{(y_\lambda(a_i) - \bar{y_\lambda})^2}{n}$$

We choose λ such that we maximize the log likelihood function

$$I(\lambda) = -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log s_\lambda^2 + (\lambda - 1)\sum_{i=1}^{N}\log(a_i)$$

y_λ (a) will be an exact normal distribution if λ=0 or 1/λ is an even integer

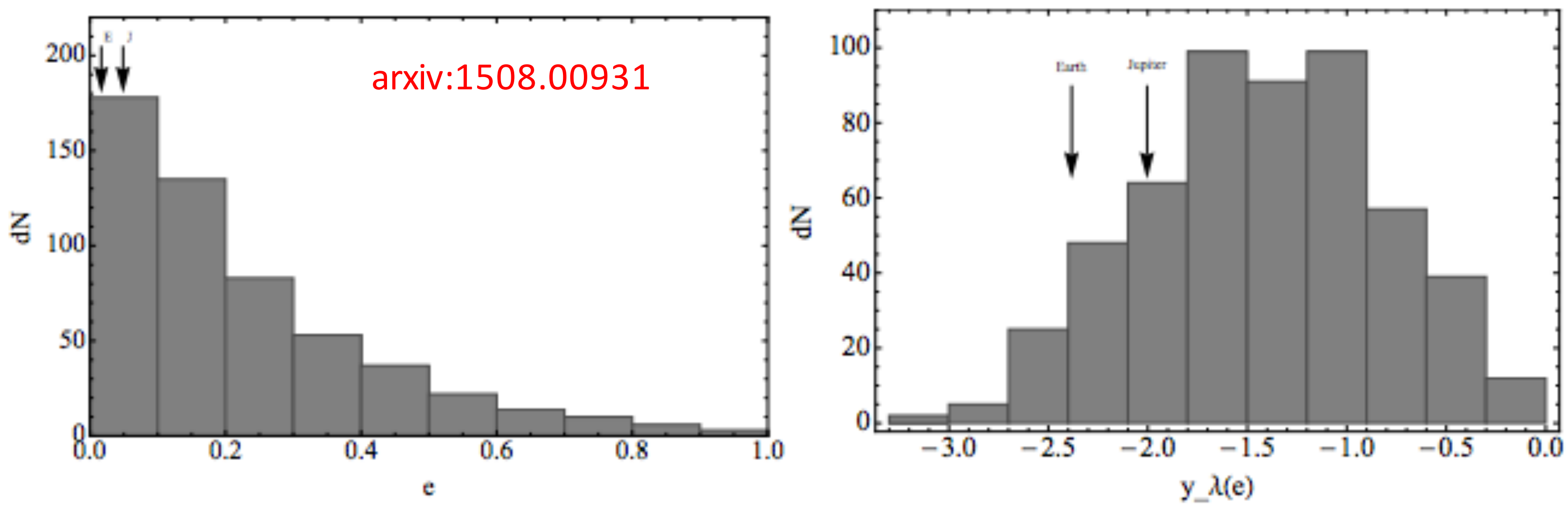


FIG. 1.— Left: Eccentricity distribution for all observed exoplanets with a measured orbital eccentricity. Right: Box-Cox transformed distribution of exoplanet eccentricities. The total number of exoplanets is 539.

Eccentricity distribution of exoplanets before and after Box-Cox transformation

Box-Cox Transformation in Python (see also stackexchange for examples)

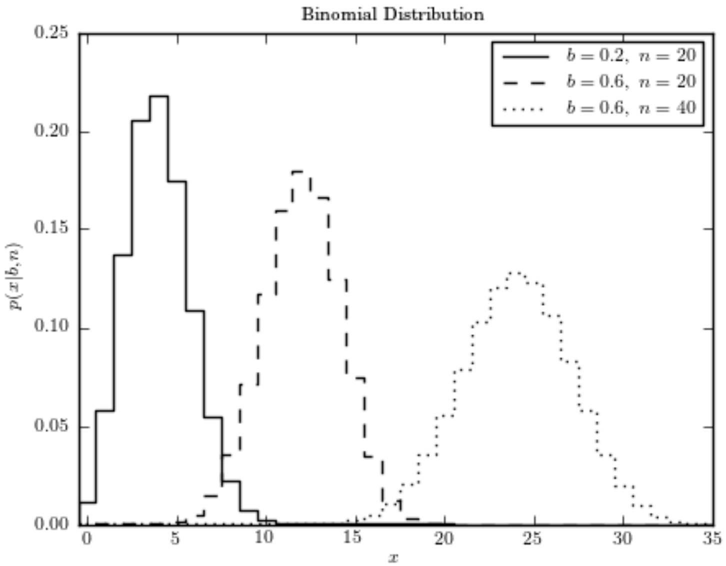
scipy.stats.boxcox

scipy.stats.**boxcox**(*x*, *lmbda=None*, *alpha=None*) [source]

Return a positive dataset transformed by a Box-Cox power transformation.

Parameters:	**x** : *ndarray* Input array. Should be 1-dimensional. **lmbda** : *{None, scalar}, optional* If *lmbda* is not None, do the transformation for that value. If *lmbda* is None, find the lambda that maximizes the log-likelihood function and return it as the second output argument. **alpha** : *{None, float}, optional* If `alpha` is not None, return the `100 * (1-alpha)%` confidence interval for *lmbda* as the third output argument. Must be between 0.0 and 1.0.
Returns:	**boxcox** : *ndarray* Box-Cox power transformed array. **maxlog** : *float, optional* If the *lmbda* parameter is None, the second returned argument is the lambda that maximizes the log-likelihood function. **(min_ci, max_ci)** : *tuple of float, optional* If *lmbda* parameter is None and `alpha` is not None, this returned tuple of floats represents the minimum and maximum confidence limits given `alpha`.

Binomial Distribution




```
from scipy import stats
dist=stats.binom(20,0.7)
r= dist.rvs(10)
P = dist.pmf(8) # prob.
evaluated at k=8
```

Binomial distribution describes the distribution of a variable that can only take discrete values
If probability of success is *b*, distribution of a discrete variable *k* that measures how many times Success occurs in N trials is given by Probability Mass Function

$$p(k|b, N) = \frac{N!}{k!(N-k)!} b^k (1-b)^{N-k}$$

N=1 is called Bernoulli distribution

Mean of a binomial distribution is given by

$\bar{k}$ = Nb

Standard deviation is given by :

$\sigma_k = [N\,b\,(1-b)]^{1/2}$

Binomial distribution can be generalized to a Multinomial distribution in case a variable has more than two discrete values.

Poisson Distribution

Poisson distribution special case of the binomial distribution describing the distribution of a discrete variable, when the number of trials (N) goes to infinity and probability of success (p=k/N) stays fixed.
Distribution of number of success k is controlled by μ = k N and is given by

$$P(k|\mu) = \frac{\mu^k \exp(-\mu)}{k!}$$

Poisson distribution is ubiquitous describes independent point process:
photon noise, radioactive decay, galaxy distribution for very few galaxies, point sources

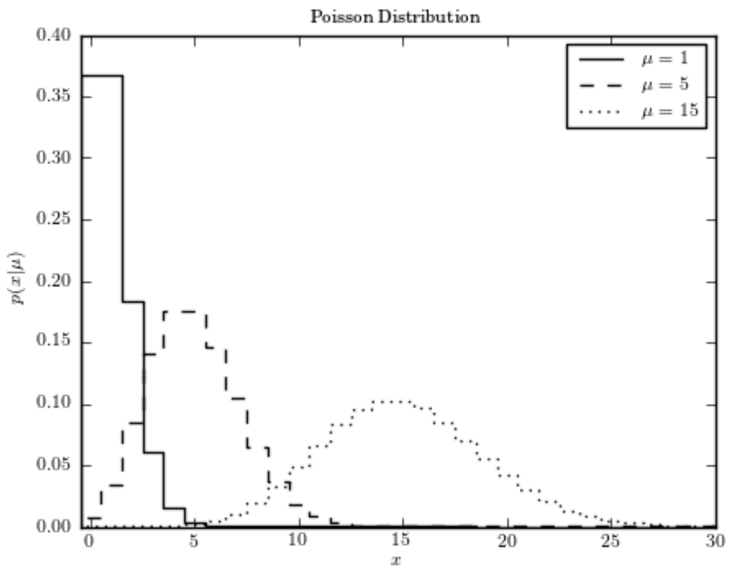
Mean (or expectation value) = μ
Standard deviation = √μ

As μ increases the Poisson distribution becomes more and more similar to Gaussian distribution

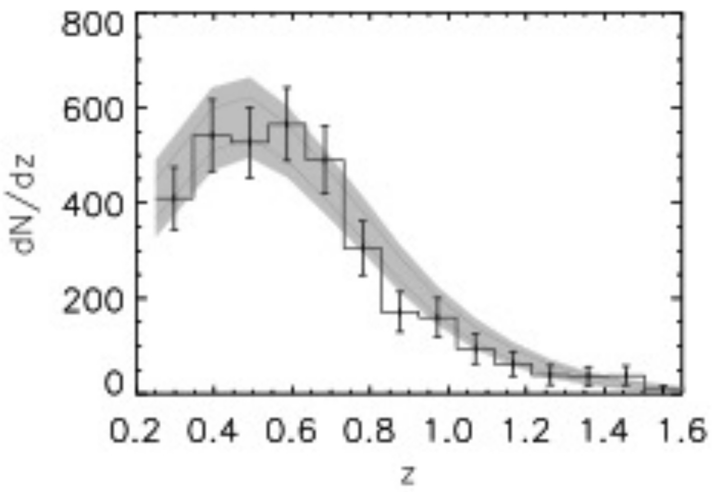
As μ increases the Poisson distribution becomes more and more similar to a Gaussian distribution given by $\mathcal{N}(\mu, \sqrt{\mu})$

Difference between Mean and Median does not become 0 but becomes 1/6

```
from scipy import stats
dist = stats.poisson(5)
r = dist.rvs(10)
p = dist.pmf(3)
```

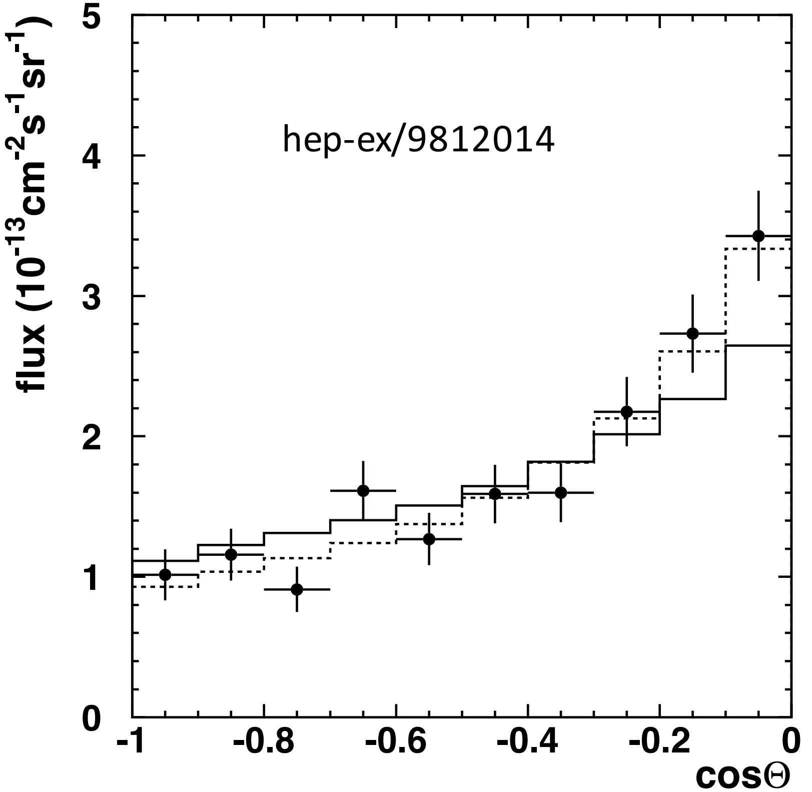


Examples of Poisson Distribution



Galaxy clusters discovered with SPT
as a function of redshift

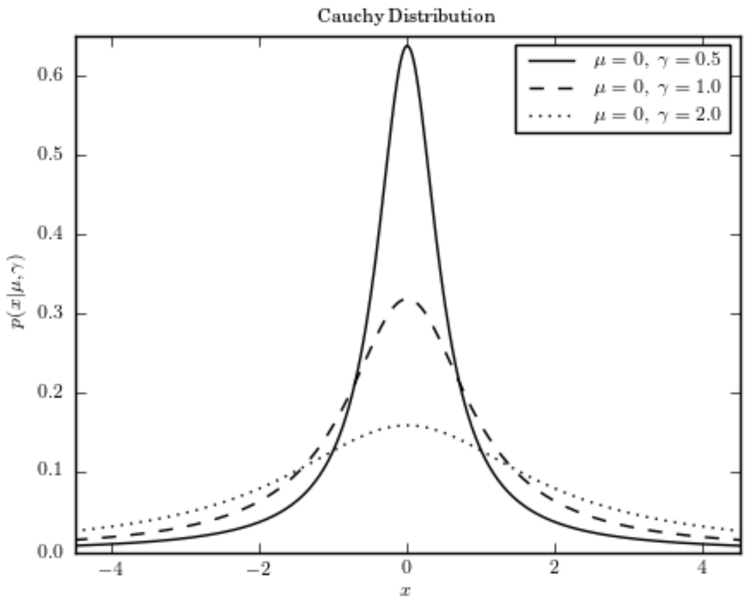
arXiv:1603.06522



Flux of upward muons in Super-K as a function
of zenith angle (1998)

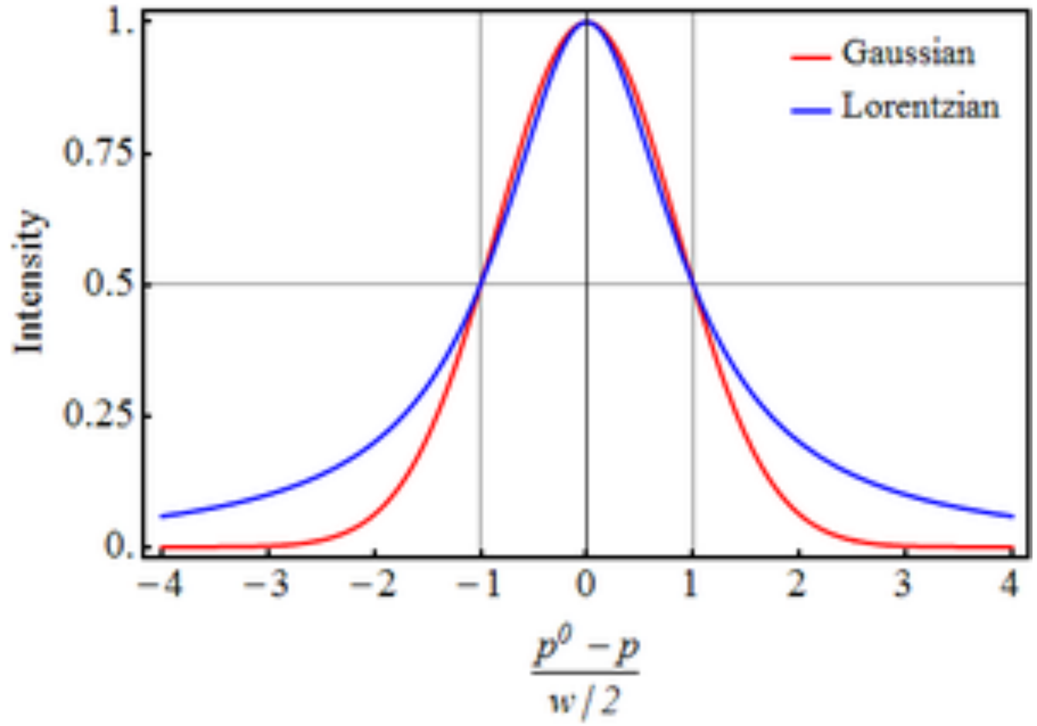
Cauchy (Lorentzian) Distribution

$$p(x|\mu,\gamma) = \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{\gamma^2+(x-\mu)^2}\right)$$



Cauchy distribution described by location Parameters μ and scale parameter γ

Exercise : Redo the above plots with $\mu=1$ and $\gamma=4$



Lorentzian line shape function

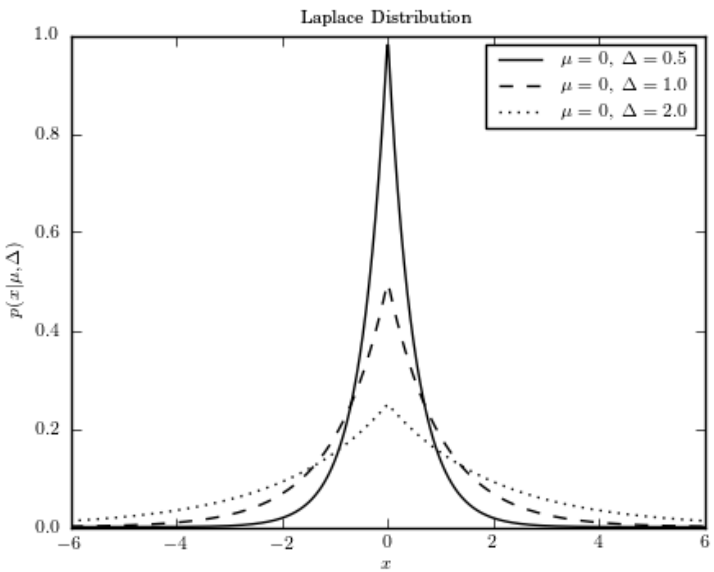
```
from scipy import stats
dist = stats.cauchy(0,1)
r = dist.rvs(10)
P = dist.pdf(3) # pdf evaluated at x=3
```

Ratio of two independent standard normal variables $z = (x-\mu)/\sigma$ with z drawn from a Normal Distribution with μ=0 and $\boldsymbol{\sigma}$=1 follows a Cauchy distribution with μ=0 and γ=1

However, ratio of two random variables drawn from two different Gaussian distributions is more complicated *(follows the Hinkley distribution)*

Exponential (Laplace) Distribution

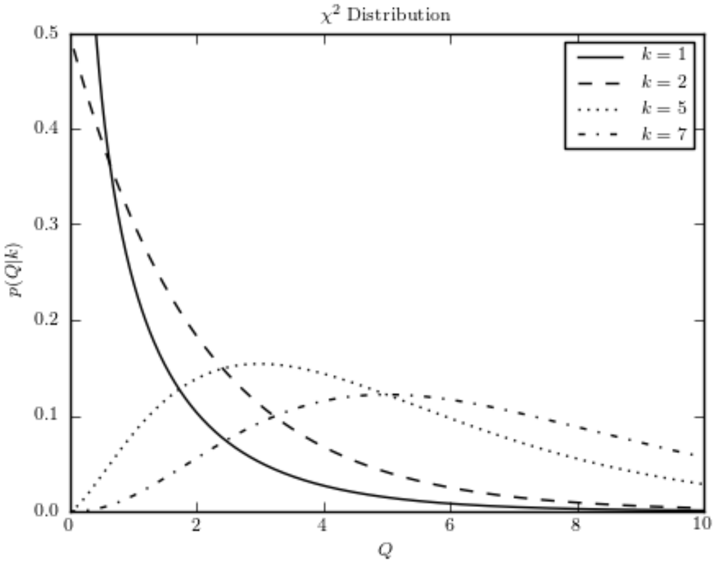
$$p(x|\mu, \Delta) = \frac{1}{2\Delta} \exp\left(-\frac{|x-\mu|}{\Delta}\right)$$




```
from scipy import stats
dist = stats.laplace(0,0.5)
r = dist.rvs(10)
P = dist.pdf(3)
```

Describes time between two successive events which occur continuously and independently at constant Rate

Chi-Square Distribution




```
from scipy import stats
dist = stats.chi2(5) #k=5
r = dist.rvs(10) #10 random draws
P = dist.pdf(3) #evaluated at x=1
```

If $\{x_i\}$ are drawn from a Gaussian distribution and if we define

$$z_i = (x_i - \mu)/\sigma$$

$$Q = \sum_{i=1}^{N} z_i^2$$

follows a $\chi 2$ distribution with k=N degrees of freedom

$$p(Q|k) \equiv \chi^2(Q|k) = \frac{1}{2^{k/2}\Gamma(k/2)} Q^{k/2-1} \exp(-Q/2)$$

Γ is incomplete Gamma function