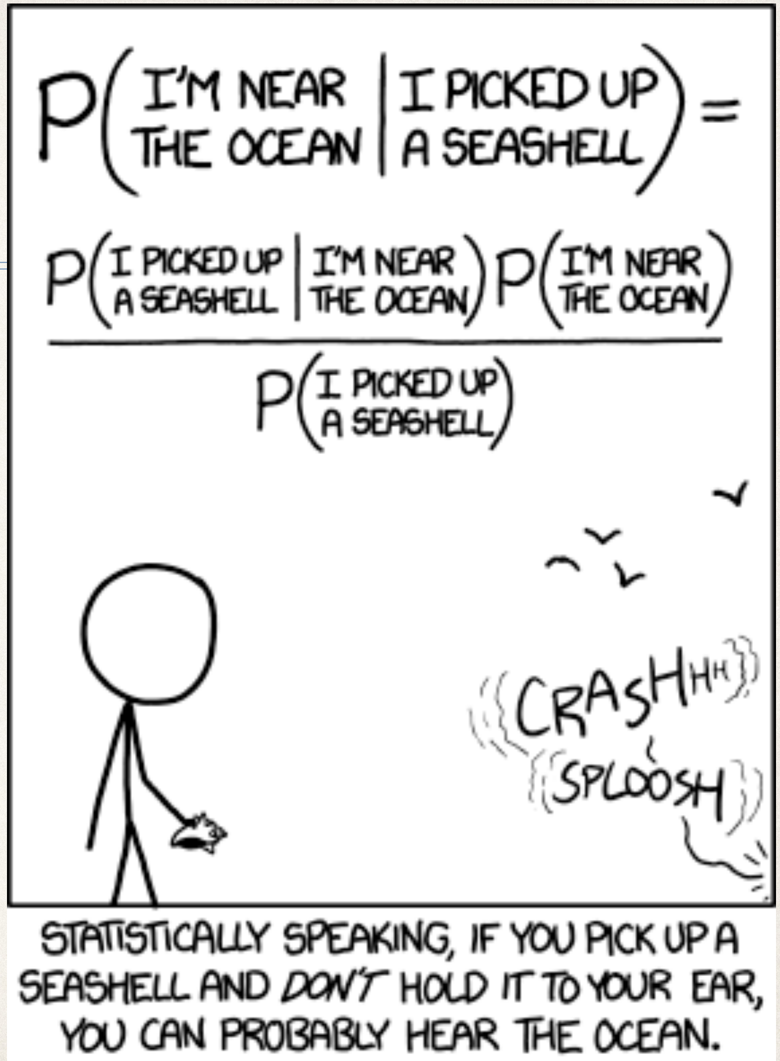
Probability Distributions

Statistics and Data Science

Spring 2025



http://xkcd.com/1236/

Goals for today: you should be able to…

- **Lecture 7 / 8 notebook:**
  - Use the Poisson distribution to describe real situations
  - Describe basic properties of Gaussian distributions
  - Choose appropriate priors for Gaussian parameters
- Identify log-normal, exponential, and Cauchy distributions (read slides offline if we do not get to them in class)

Review: the Binomial distribution

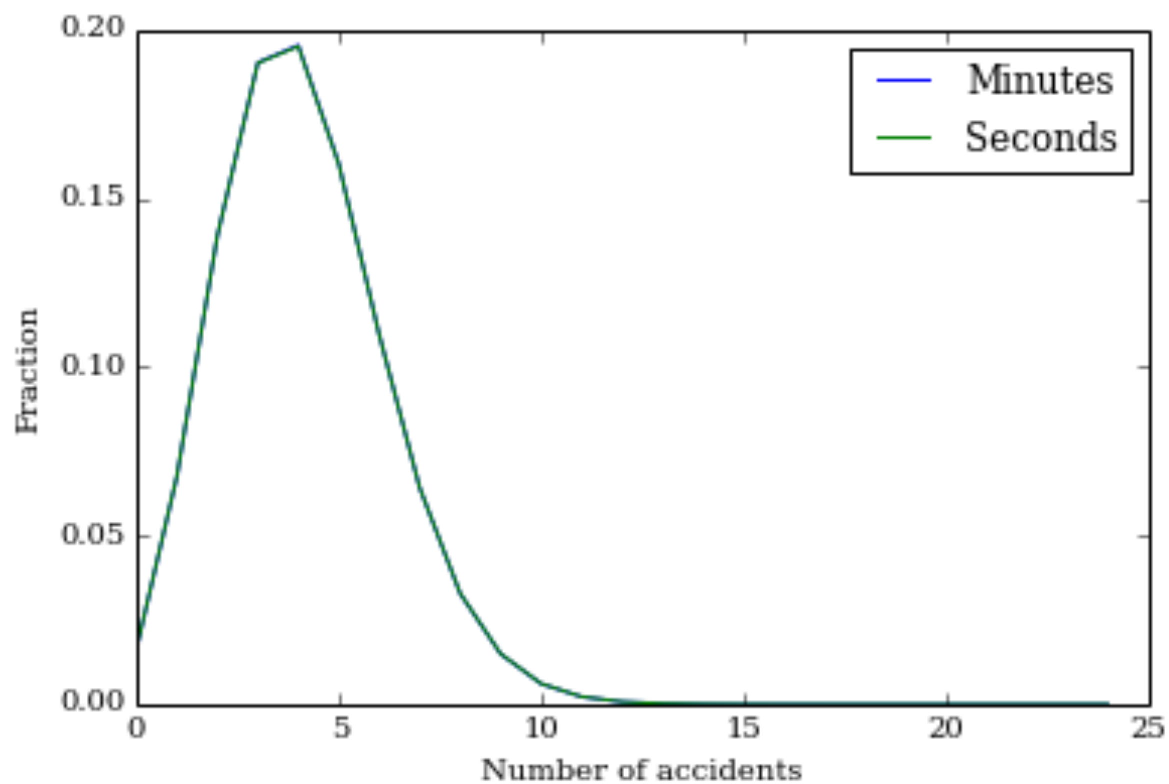
- In general, if there is a probability p of success, and we do N trials, then the probability of M successes is described by a binomial distribution:

$prob(M\ successes) = C(N,M)\ p^M\ (1-p)^{N-M}$

- The binomial distribution has mean $\mu = N\,p$ and variance $\sigma^2 = N\,p\,(1-p)$
- We are considering the case where $N\,p$ is finite but $p{\sim}0$, so $\mu = \sigma^2$
    - Many trials, but a small chance of an occurrence per trial...

Binomial distributions with p~0

- Expected distribution of number of traffic accidents in Pittsburgh on a given day, treating every minute OR every second as a separate binomial (AKA Bernoulli) trial



The Poisson distribution

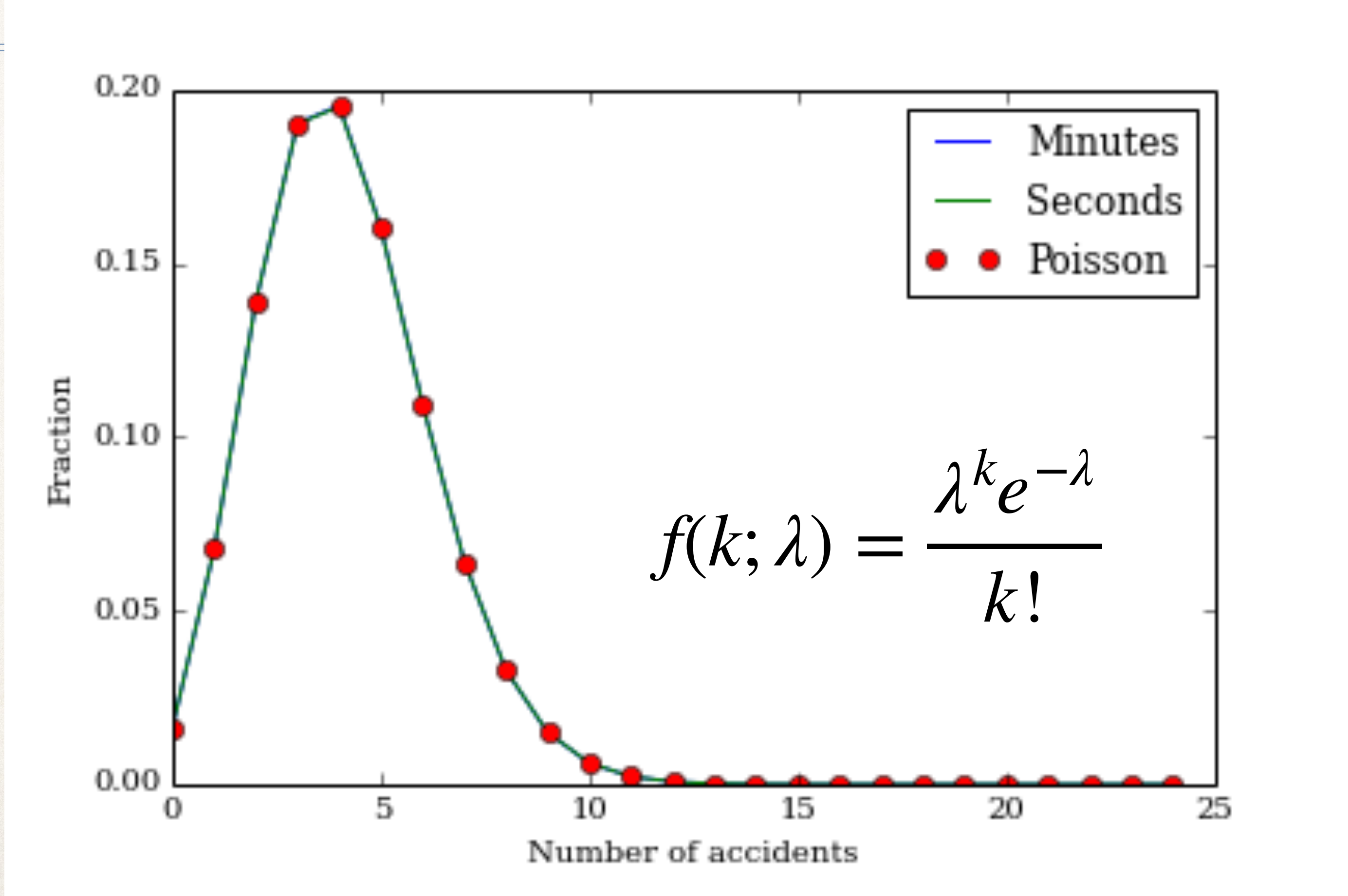
- If we have a scenario where we have a process that occurs at a constant rate per unit time (so that it would be described by a binomial distribution with N large and p small, if we choose a small enough time interval), the number of times that process occurs, k, will follow a Poisson distribution:

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where λ is the mean total number of times that event will occur, integrating over the whole time interval (e.g., 4.1 for traffic accidents in Pittsburgh)

- The Poisson distribution has mean $\mu=\lambda$ and variance $\sigma^2=\lambda$

The Poisson distribution matches the binomial result



Why do physicists and astronomers care about Poisson distributions?

- In the UV, optical, near-IR, X-ray, and γ-ray parts of the spectrum, our detectors respond to individual photons .
- If observing conditions are constant and the source is not variable, we'd expect a constant number of photons per hour (or minute, or second) to be detected.
- As a result, the total number of photons we detect (e.g., in some pixel of a detector) after some amount of integration time will follow a Poisson distribution (presuming our detector returns a number of counts proportional to the number of photons entering).
- Similarly, the number of events of a certain type we observe in a particle collider will be Poisson-distributed, since we have many events total but only a small fraction that are of a given type

Poisson distributions for λ=1,4,10

- Let's plot a few Poisson distributions.
- We can only get integer results, so we can do:

```
k=np.arange(25)
```

- to set up an array from 0 to 25.

- Then we can plot the distribution for any choice of lambda:

```
plt.plot(k,lambda**k * ??? (-lambda)/factorial(k) )
```

- Like most Python mathematical functions, factorial() works fine when given an array as input.
- **Plot Poisson PMFs for lambda=1, lambda=4, and lambda=10.**

Creating legends

- Putting legends in plots is easy in Python.

1) In all your plot commands, use the `label=` keyword to provide a string label for each set of points/curve in the plot; e.g.:

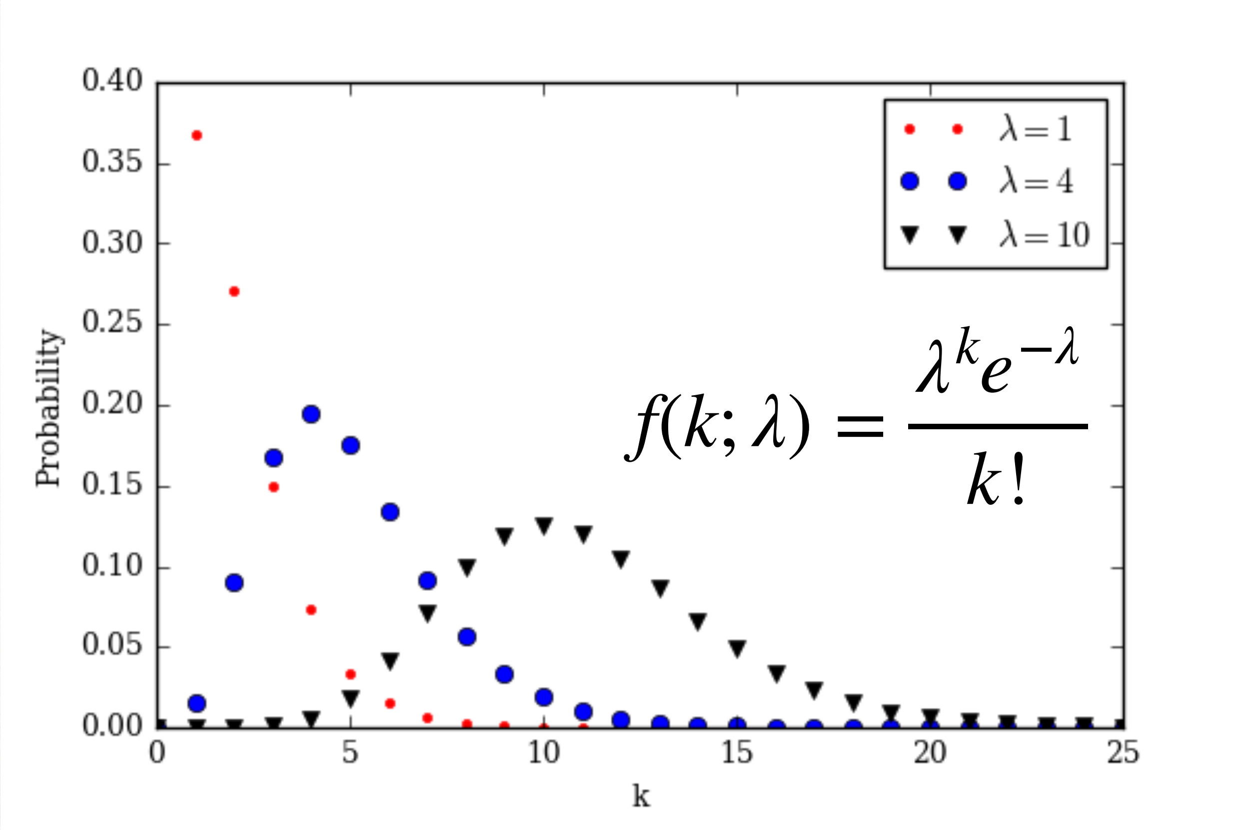
```
plt.plot(x,events_from_minutes, label='Minutes')
```

2) Use the pyplot command `plt.legend()` to draw the legend on your plot. Take a look at the help information for it; you can set legend positions, etc. with keywords.

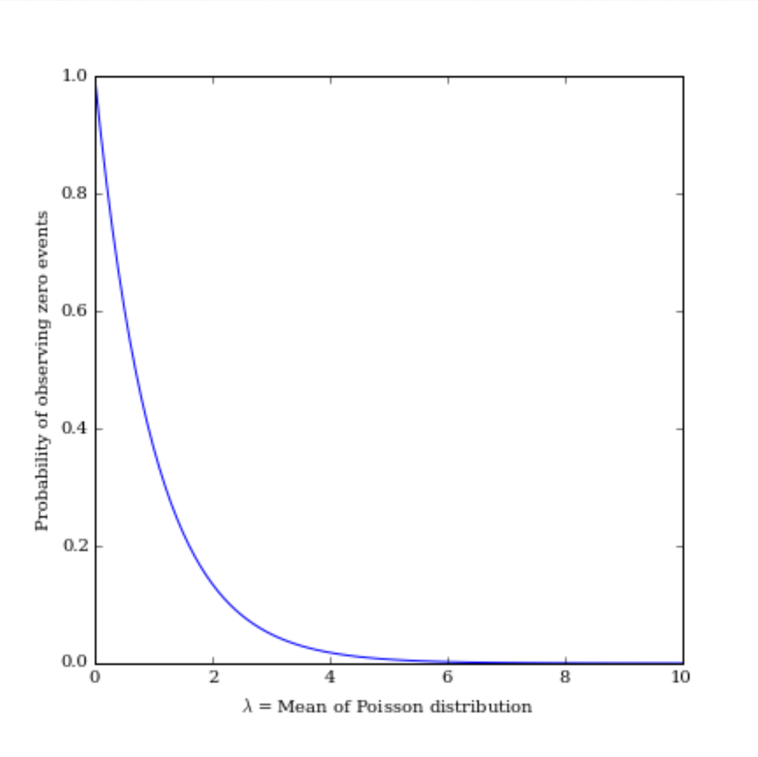
```
plt.legend()
```

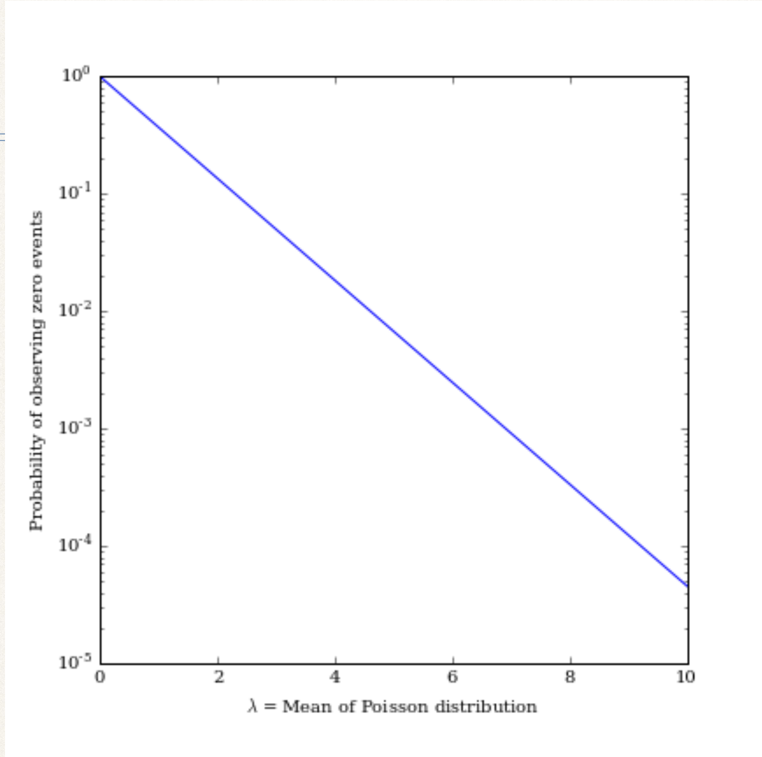
Poisson distributions for λ=1,4,10

Even for λ=10, the Poisson distribution is a bit asymmetric - as <0 events is impossible.



As a rule of thumb, it's OK to ignore this asymmetry if λ>5





Here we plot the probability of getting 0 as a function of lambda:

```
plt.plot(x,x**0*np.exp(-x)/factorial(0))
plt.yscale('log')
```

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Priors for the Poisson distribution

- It turns out the Jeffreys prior for a Poisson distribution is *prob*(λ)$\propto \lambda^{-1/2}$. So if we measure *N* counts from a Poisson process, the resulting posterior for λ is:

  $prob(\lambda \mid N) \propto prob(N \mid \lambda)\, prob(\lambda) \propto \lambda^N e^{-\lambda}\, \lambda^{-1/2} / N! \propto \lambda^{N-1/2} e^{-\lambda}$

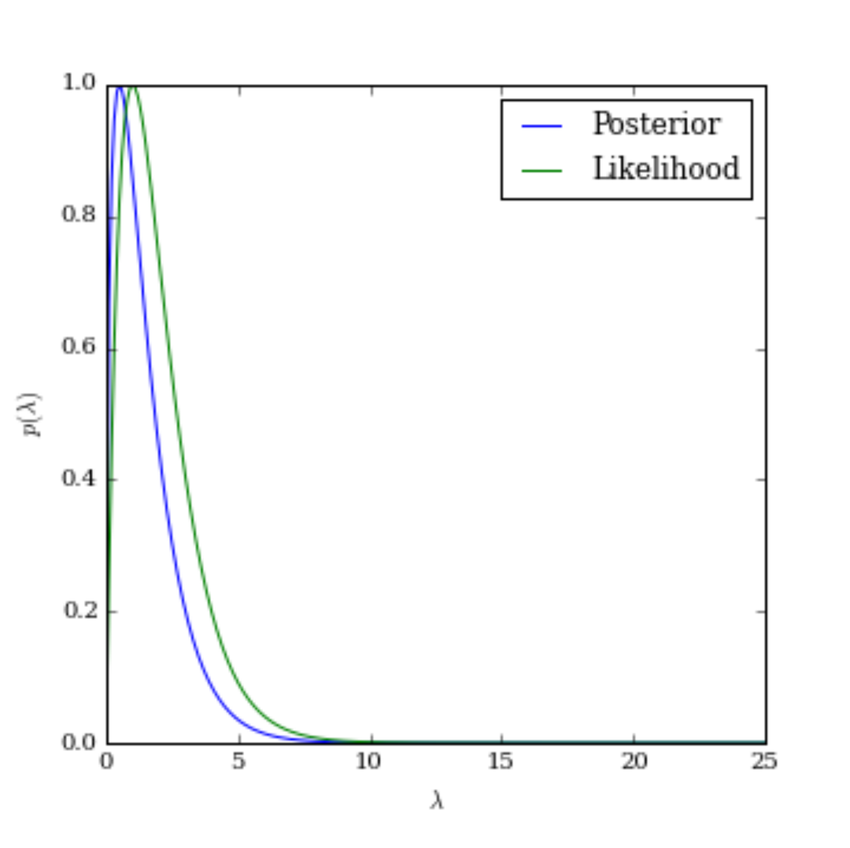
- Let's plot up the distribution, for N=1,3,5,10, with and without the prior:

```
lam_arr=np.linspace(0,25,251)

N=1

likelihood = lam_arr**(N)*np.exp(-lam_arr)

posterior = lam_arr**(N-0.5)*np.exp(-lam_arr)
```

- now try what happens with N=3,5,10...

Again, as N gets larger, the prior has less impact

- Here are the results for N=1, normalized to have the same maximum

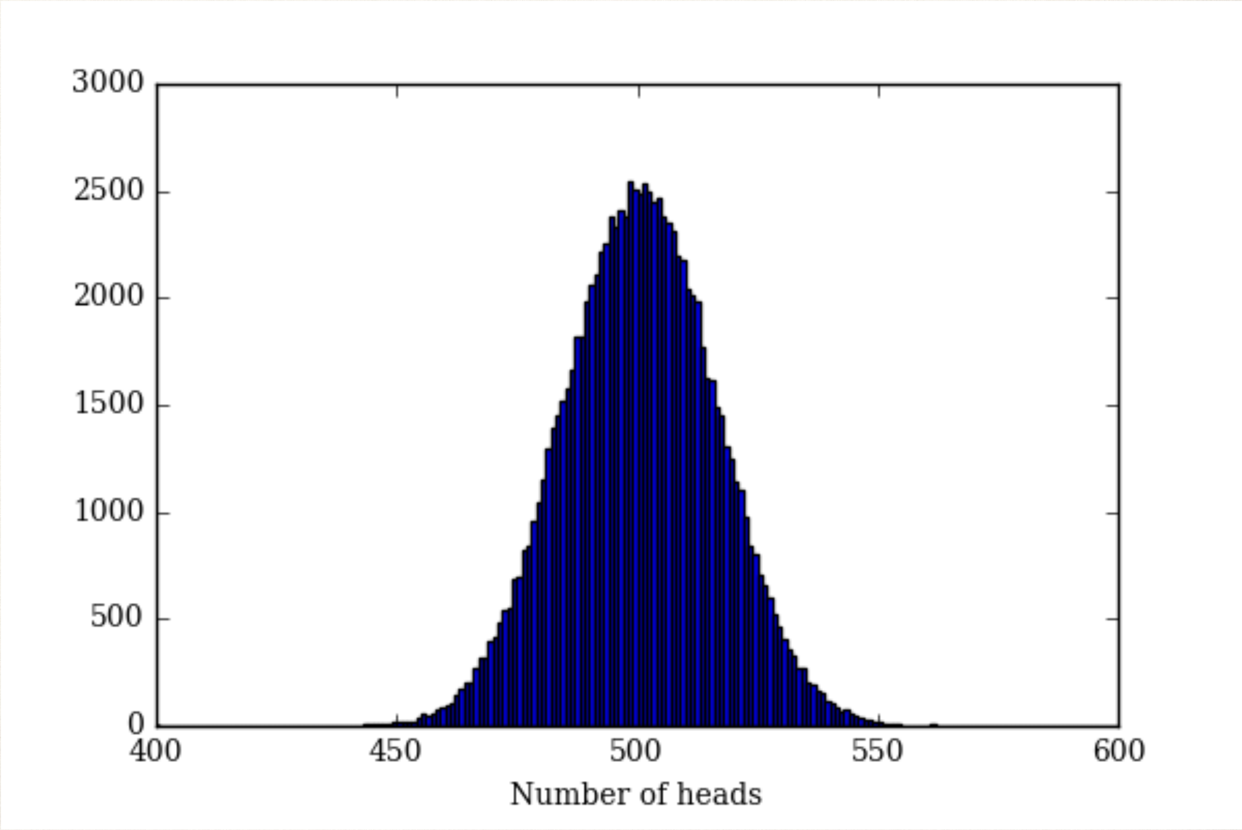


A few more Poisson facts

- If we have two measurements, x_1 and x_2, drawn from Poisson distributions with means μ_1 and μ_2, then their sum, $x_1 + x_2$, will follow a Poisson distribution with mean $\mu_1 + \mu_2$
- E.g., if we have two images with just Poisson noise, we can just add the two images and get a result equivalent to a single image with a longer exposure time (given by the sum of the 2 exposure times).
- One thing to beware of: the Poisson distribution controls the values of actual counts. If we converted to counts per hour, say (or otherwise multiply x by some constant c), the new variable cx will have mean $c\lambda$, but variance $c^2\lambda$, so **the mean and variance are no longer equal**. Similar caution is needed when averaging, instead of summing, multiple Poisson-distributed measurements.

The Gaussian Distribution

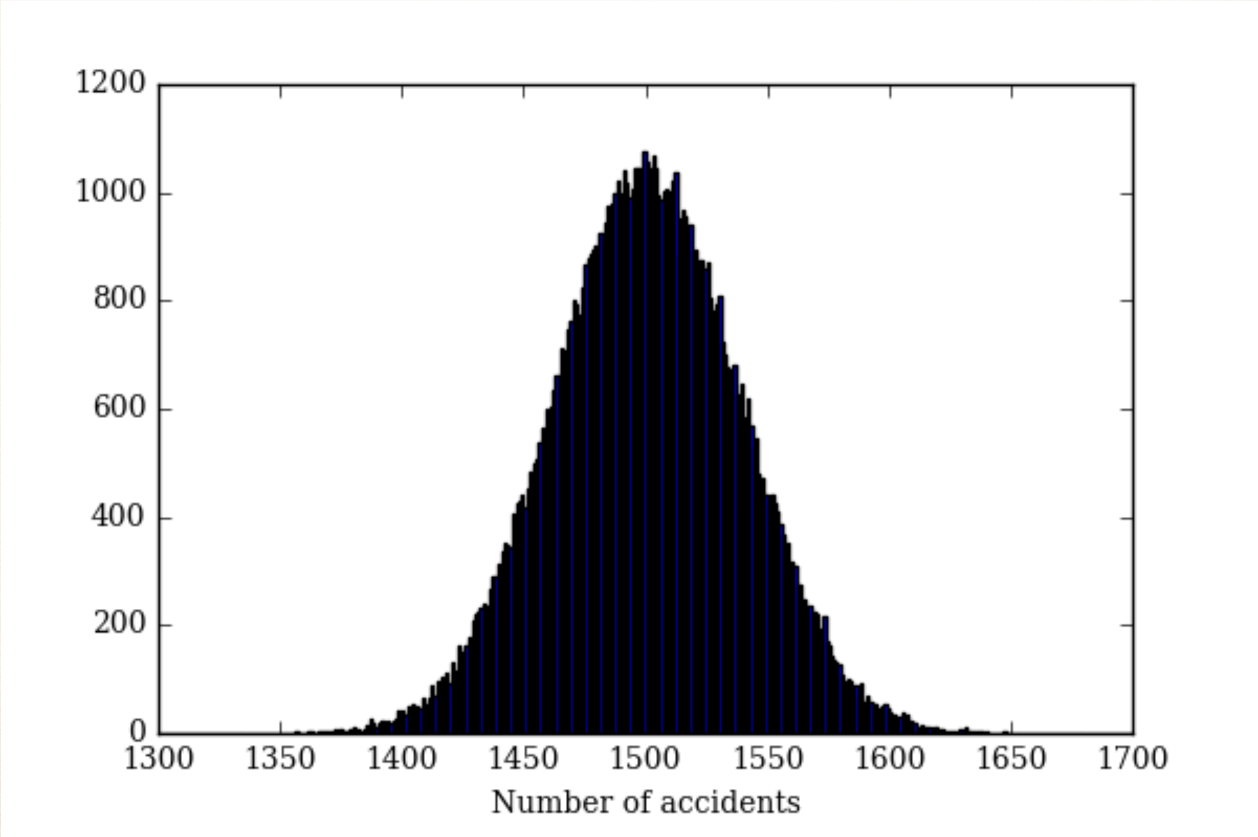
- Data drawn from a binomial distribution with large N follows a familiar shape:



Results from flipping 1000 coins & counting the total # of heads 10^5 times

The Gaussian Distribution

- So does data from a Poisson distribution with large N



Results from simulating # of accidents in Pittsburgh per year

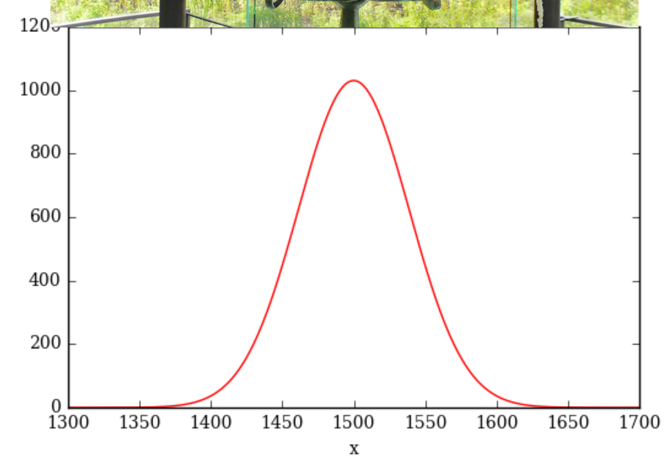
The Gaussian Distribution

- Both of these distributions can be matched closely by a Gaussian (or "Normal") distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

which has mean $\mu=\mu$ and variance $\sigma^2 = \sigma^2$

- Sometimes this is called a "bell curve"

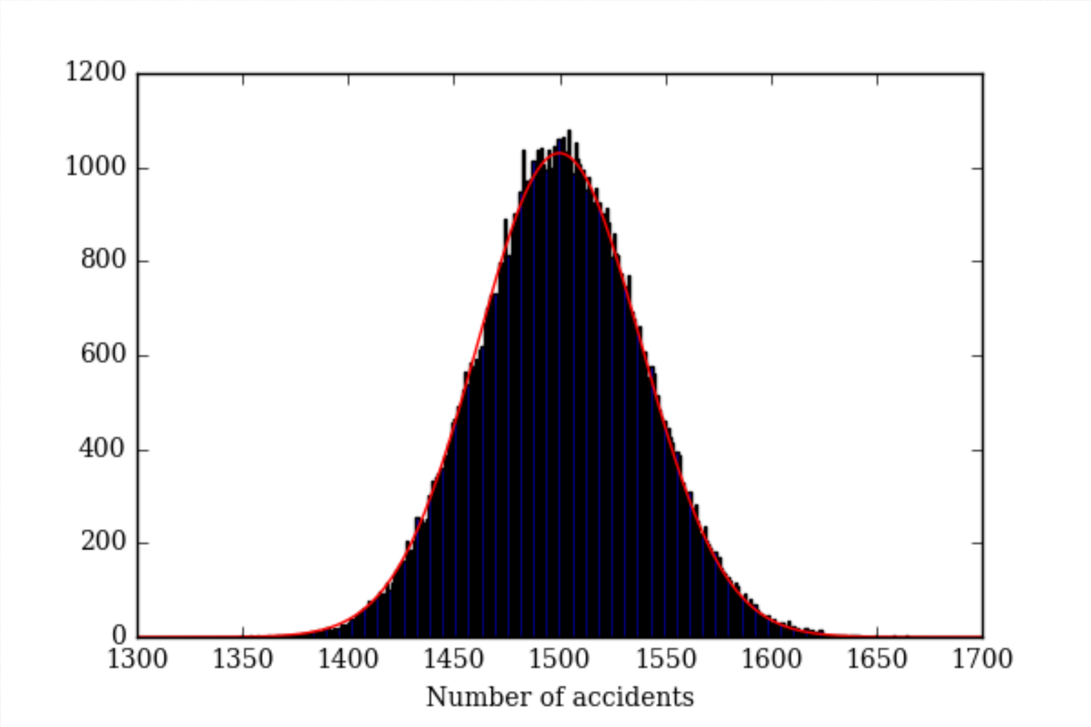


The Gaussian Distribution

- A Poisson distribution has mean $\mu=\lambda$ and variance $\sigma^2 = \lambda$; if we just plug those values into

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

we get an excellent match to the Poisson-distributed data (red curve).



Parameters of the Gaussian

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

- μ determines where the center of the Gaussian occurs. Note that the expression is symmetric about x=μ.
- Let's try it:

```
x=np.linspace(-10,10,20001)
mu=0
sigma=1
plt.plot(x,1/np.sqrt(2*np.pi*sigma**2)*np.exp(-(x-mu)**2/2/sigma**2))
mu=1
plt.plot(x,1/np.sqrt(2*np.pi*sigma**2)*np.exp(-(x-mu)**2/2/ sigma**2),'g--')
```

Parameters of the Gaussian

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

- σ determines how broad the curve is. The factor in front of the exponential just normalizes the curve to have integral 1.

```
mu=0
sigma=1
y1=1/np.sqrt(2*np.pi*sigma**2)*np.exp(-(x-mu)**2/2/sigma**2)
```

- **Now define y2, using sigma=2, and y0_2, using sigma=0.2.**
- **Plot up y1, y2, and y0_2 vs. x to see the difference sigma makes. Also try making the y axis logarithmic (how do we do that?) and explain what you see.**

Integrals of the Gaussian

- For any Gaussian distribution, the height at a point where $|x-\mu| = A^*\sigma$ will always be the same fraction of the peak height (i.e., $\exp(-A^2/2)$).
- As a consequence, for any Gaussian distribution, the integral between $-A^*\sigma$ and $+A^*\sigma$ will always be the same fraction of the overall integral (1).

```
mu=0
sigma=10
y10=1/np.sqrt(2*np.pi*sigma**2)*np.exp(-(x-mu)**2/2/sigma**2)
interp_10=interpol.interp1d(x,y10,kind='cubic')
print(f'integrals for sigma=10 Gaussian: \
 {integrate.quad(interp_10,-10,10,epsrel=1.e-4)[0]:.8f}, \
 {integrate.quad(interp_10,-20,20,epsrel=1.e-4)[0]:.8f}')
```

Integrals of the Gaussian

- For a Gaussian distribution, here are tables of the fraction of times a random number will be within $N\sigma$ of μ :

N	fraction $<N\sigma$
1	0.682689492137
2	0.954499736104
3	0.997300203937
4	0.999936657516
5	0.999999426697
6	0.999999998027

fraction $<N\sigma$	N
0.80	1.28155
0.90	1.64485
0.95	1.95996
0.98	2.32635
0.99	2.57583
0.995	2.80703
0.998	3.09023
0.999	3.29052
0.9999	3.8906
0.99999	4.4172

Source:
wikipedia.org

Transforming Gaussians

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

- We can generate Gaussian random variables with mean 0 and sigma 1 using the function `numpy.random.randn()`, which works just like `numpy.random.rand()`:

```
import numpy.random as random
nrandom=int(1E4)
bin_array=np.linspace(-5,5,1001)
n,bins,patches=plt.hist(random.randn(nrandom),bins=bin_array,histtype='step')
```

- In statistics parlance, these numbers are drawn from the distribution $N(0,1)$: a normal distribution with mean 0 and variance 1.
- Since the center shifts with μ and the width is proportional to σ, we can get numbers distributed as $N(\mu,\sigma^2)$ by transforming variables:

  $x' = \sigma x + \mu$

where x is distributed as $N(0,1)$ (Note: sometimes the notation used is $N(\mu,\sigma)$. you need to be careful!)

- **Make a histogram for Gaussian random data drawn from the distribution $N(5,5^2)$**

So why is the Gaussian distribution important?

- It is not a coincidence that, for large N, both binomial and Poisson distributions match a Gaussian well (as we saw in HW 1). This is a consequence of the Central Limit Theorem:

    If you sum N different numbers, each independently randomly drawn from the same probability distribution which has finite variance, then as N increases, that sum will be closer and closer to being distributed as a Gaussian distribution as N gets larger.

- It was first postulated in the same work (in 1733) that introduced the Gaussian (as a limit of the binomial distribution).

Why does the Central Limit Theorem apply to the binomial & Poisson cases?

- In the binomial case, we are recording the probability of M successes in N trials. This is obviously the same as summing the number of successes from each individual trial; since the only possibilities for each one are 0 or 1, the variance for a single trial must be finite (it's $p\,(1-p)$). So we are summing N numbers drawn from identical PDFs with finite variance.
- Recall that, if X_1 and X_2 are Poisson distributed with parameter λ_1 & λ_2, then X_1+X_2 is Poisson distributed with parameter $\lambda = \lambda_1 + \lambda_2$
  - So a Poisson-distributed variable with parameter λ is equivalent to the sum of N Poisson-distributed random variables with parameter λ/N; each of them then has variance λ/N, which is finite.

Applying the Central Limit Theorem

- The book adopts a more practical definition of the CLT:

  If you form averages M_n of samples of size n from a population with finite mean μ and variance σ^2, then the distribution of $(M_n - \mu)/(\sigma/\sqrt{n})$ approaches a Gaussian with mean 0 and variance 1 as n goes to infinity.

- Of course, averages are just sums divided by n.

- We won't bother proving the CLT here; you can look at a stats book or on the internet. The typical proof is based on what in statistics are called "characteristic functions" but we typically call Fourier transforms.

Priors for the Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

prob(params | data) = prob(data | params) prob(params) / prob(data)

- In the case of a binomial or Poisson distribution, we had one parameter which could be fit for (p or λ).
- For a Gaussian, we have two, μ and σ, so *prob(params)* is a function of 2 variables
- Typically, we assume those variables are independent, so $prob(\mu, \sigma) = prob(\mu)\, prob(\sigma)$
- Jeffreys priors are:
  - $prob(\mu) = 1$
  - $prob(\sigma) = 1/\sigma$

(compare to the prior for a Poisson distribution, $prob(\lambda) \propto \lambda^{-1/2}$)

Priors for the Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

prob(params | data) = prob(data | params) prob(params) / prob(data)

- Note that, if $prob(\sigma) = 1/\sigma$, then we have a prior that is uniform not in σ , but in $\ln(\sigma)$:

  let $u = \ln(\sigma)$, $f(\sigma)$=the PDF for σ, and $g(u)$=the PDF for u

  then: $g(u)\, du = f(\sigma)\, d\sigma$

  $g(u)\, d\sigma \,/\, \sigma = 1/\sigma\, d\sigma$

  so $g(u) = 1$

- The Jeffreys prior for the variance, $V = \sigma^2$, is $1/V = 1/\sigma^2$ ...

Some rules of thumb on priors

- Typically, in the no-information case, you won't be too far wrong if you go with one of the two cases we see for the Gaussian:

- Numbers that could take any real value, positive or negative, or that have some known, characteristic scale:
    - ➡ use a uniform prior (sometimes over a fixed range), like $prob(\mu) = 1$

- Numbers that could take any nonzero value, or that could have any unknown scale/order of magnitude (like a physical constant):
    - ➡ use a prior uniform in ln x, like $prob(\sigma) = 1/\sigma$

Bayesian use of Gaussians

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

- Suppose we have some measured value for x, where we expect x to come from a Gaussian of some unknown μ, but with known $\sigma=2$, e.g., $x \sim N(\mu, 2^2)$.
- What is the posterior for μ given a measurement (say, x=5), using the Jeffreys prior $prob(\mu) = 1$?

```
x=5
sigma=2
mu=np.linspace(-10,10,201)
likelihood = ???
prior= ???
plt.plot(mu,likelihood,label='Likelihood')
plt.plot(mu,likelihood*prior,label='Posterior')
plt.legend()
```

You should find this is just a Gaussian, centered at our measured value, with standard deviation sigma!

Bayesian interpretation of the measurement

- So if we measure x=5, with a known uncertainty (σ) of 2, we'd expect μ to be within 2 units of 5 (i.e., <1 σ away) 68% of the time, within 4 (=2 σ) 95% of the time, etc.

Bayesian estimate for σ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

- Now suppose we have some measured value for x, where we expect x to come from a Gaussian of known mean $\mu=0$, but with unknown σ.
- What is the likelihood for σ and posterior for σ given a single measurement (say, $x=5$), with prior $prob(\sigma) = 1/\sigma$? **This time, be sure to normalize the posterior distribution to have integral 1.**

```
x=5
mu=0
sigma=np.linspace(0.,50.,501)+1.E-3 # want to avoid dividing 1/0
likelihood=???
prior = ???
norm = ???

plt.plot(sigma,likelihood,label='Likelihood')
plt.plot(sigma,likelihood*prior/norm,label='Posterior')
```

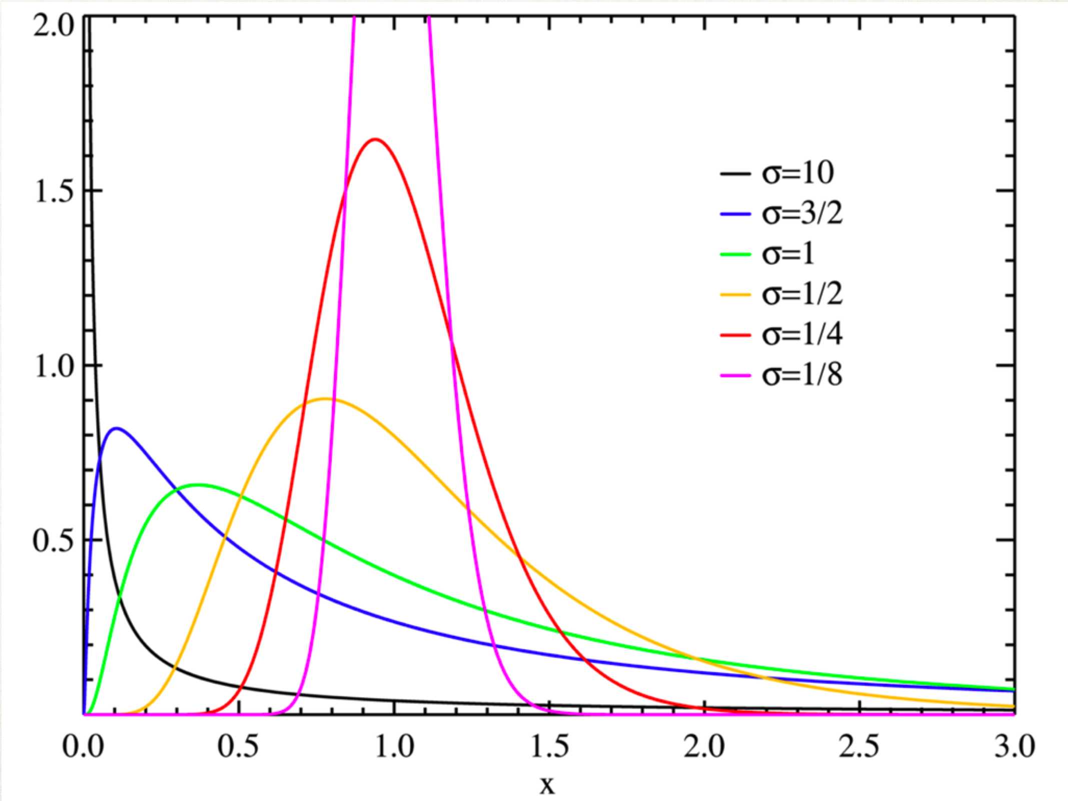
Other interesting distributions

- There are a few other classical distributions you are likely to encounter in physics and astronomy:

  1) **the Log-Normal distribution**: if x is described by a log-normal distribution, then $\ln x$ is distributed as some Normal distribution; i.e., $x = e^y$, where y is drawn from some Normal distribution, $y \sim N(\mu, \sigma^2)$

- Just like Gaussians result when you look at the sum of many different random variables, log-normal distributions tend to occur when you look at the result of the product of many different random variables (of nonzero mean). E.g.: the possible distribution of stock prices a year from now is the combination of many different daily ups & downs.
- See `scipy.stats.lognorm`

Lognormal distributions



Plot: wikipedia.org

Other interesting distributions

2) the Exponential distribution:

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & , x \geq 0, \\ 0 & , x < 0. \end{cases}$$

- This has mean $\mu=\beta$ and variance $\sigma^2 = \beta^2$
- The exponential distribution occurs, for instance, for the distribution of heights of molecules in an atmosphere (or stars in a galaxy above or below its disk).
- Cf. `scipy.stats.expon` (which uses a parameter lambda which is equivalent to 1/ β in the equation above)

Other interesting distributions

3) **The power-law distribution:**

$$f(x) = (x/x_0)^\alpha$$

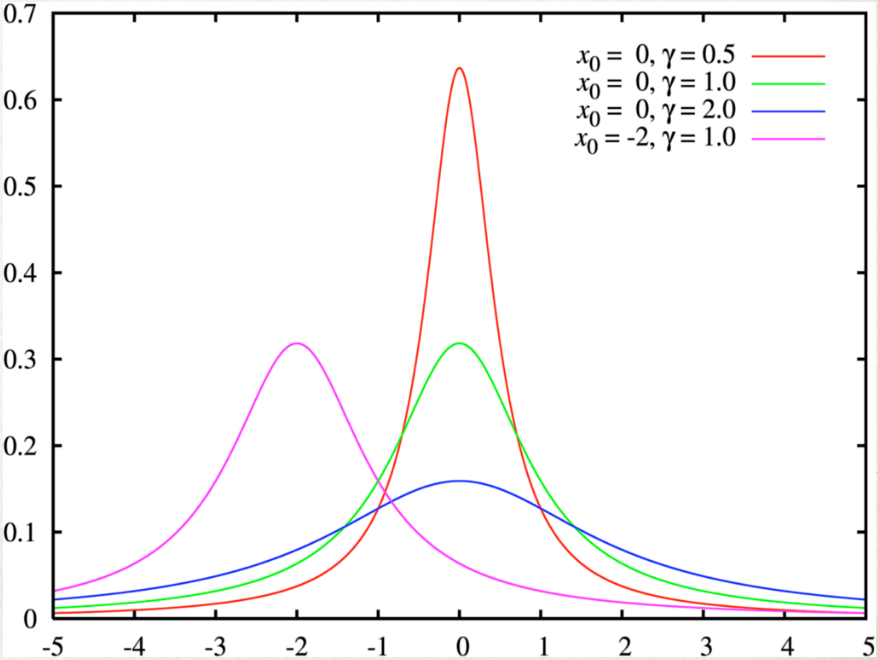
- This is an improper distribution: although we'd like to define it over $0 < x < \infty$, it will have infinite integral over that range.
- Normally, we choose to truncate it somehow.
- The flux distribution of some class of objects often follows a power-law distribution (or energy distribution of particles from a collision, or distributions in some CM systems).
- See: `scipy.stats.powerlaw`

Other interesting distributions

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]} = \frac{1}{\pi}\left[\frac{\gamma}{(x - x_0)^2 + \gamma^2}\right]$$

4) The Cauchy (or Lorentz or Lorentzian) distribution:

- In this case, x_0 determines the center/peak location of the distribution and γ determines its width. It looks innocuous...
- Implemented as `scipy.stats.cauchy`



Plot: wikipedia.org

The Cauchy Distribution

$$f(x;x_0,\gamma) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]} = \frac{1}{\pi}\left[\frac{\gamma}{(x-x_0)^2+\gamma^2}\right]$$

- Notice that the distribution falls off as x^{-2} far from the peak.
- If we calculate the mean = **E***x* we get an undefined number; the same thing happens if we calculate the variance.
- The Cauchy distribution is a classic counterexample in statistics for many theorems. In astronomy, spectral lines and point spread functions often have Cauchy wings; the spectral profile of a laser does too due to collisional broadening.
- The ratio of two Gaussian variables with mean 0 will have a Cauchy distribution; the sum of Cauchy-distributed variables is also Cauchy-distributed, **not** Gaussian - the Central Limit Theorem doesn't apply to it!