Probability and Random Numbers

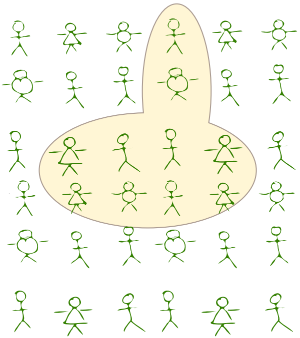
- Some basics of probability
  - population and sample
  - random variables
  - probability distribution function (PDF)
  - cumulative distribution functions (CDF)
  - expectation values
- Pseudorandom Numbers
  - Linear Congruential Generators
  - properties of good random number generators
- Read section 3 of T2.2 Lecture Notes on standard 1D distributions

Population and Sample

Population
set of all entities that are the object of a statistician's interest

- all American males,
- all graduate students at FSU
- all US voters

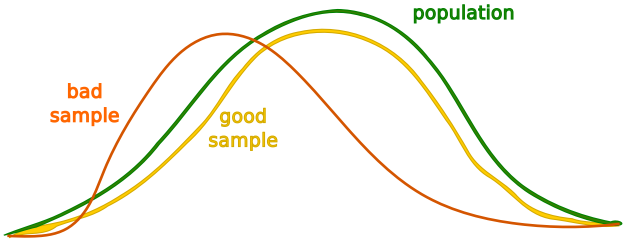
Sample
a *subset* of the population which is observed



The aim is to make inferences about the population from the sample.

Population and Sample

Constructing a "representative" sample is key (but difficult).[1]



Bad: using NBA players to estimate the height of an American male

Bad: using New York voters to predict the outcome of a presidential race.

[1] "double-blind randomized control trials" are the gold-standard in many fields to study effects of intervention

Random Variable

A random variable is a well-defined attribute of entities in a population.

Examples of population : random variable

- all American males : height
- FSU graduate students : GPA

The random variable may be discrete or continuous

Continuous

- person's height or weight
- finish times at a marathon
- fraction of women in a population

Discrete

- roll of a die
- result of coin toss
- number of children per household

Probability Distribution Function

The distribution of random variables can be described by the *probability distribution function*, or the *probability density function*.

Luckily the abbreviation for both of them is the same - PDF.

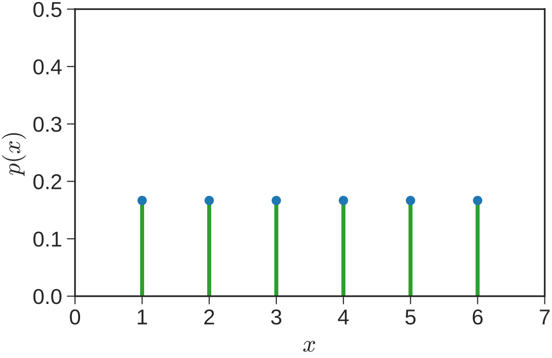
For example, the roll of a fair die may be described by the following PDF.

$$p(x) = \frac{1}{6}, \quad x \in \{1, 2, .., 6\}$$

x is the random variable. This PDF says that the probability of any outcome x is equal.

Thus, the probability $p(x = 3) = p(x = 1) = 1/6$.

Probability Distribution Function



Note that the sum

$$\sum_{x=1}^{x=6} p(x) = 1.$$

This is a *normalized* discrete PDF.

Probability Distribution Function

As an example of a *continuous* PDF, consider the heights of American males or females.

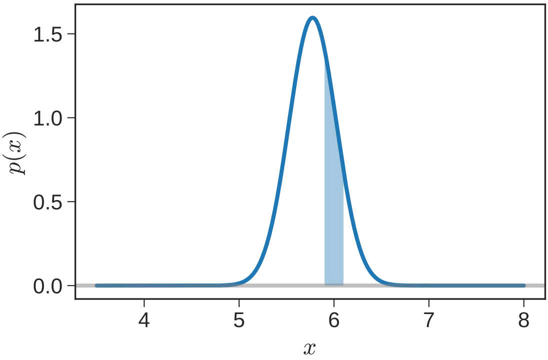
Perhaps it can be described by a *normal* distribution:

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

Note: actual data for men fits $\mu = 69.3$, and $\sigma = 3.0$ inches, while for women, $\mu = 64.0$ and $\sigma = 3.0$ inches.

Interpretation: The probability of the height lying between x and $x + dx$ is given by $p(x)dx$.

Probability Distribution Function



area of the shaded area = probability that the height of a randomly chosen individual from this population is between 5.9 and 6.1 feet.

Probability Distribution Function

The *normalization* condition on continuous PDFs is

$$\int_{-\infty}^{\infty} p(x)dx = 1.$$

In other words, the area under the curve is one.

The corresponding *normalization* condition for discrete PDFs is

$$\sum_{x=-\infty}^{+\infty} p(x) = 1.$$

Question: What does the normalization condition *mean*?

Cumulative Distribution Function

The *cumulative distribution function* or CDF $F(x)$ represents the probability that the random variable $X \leq x$.

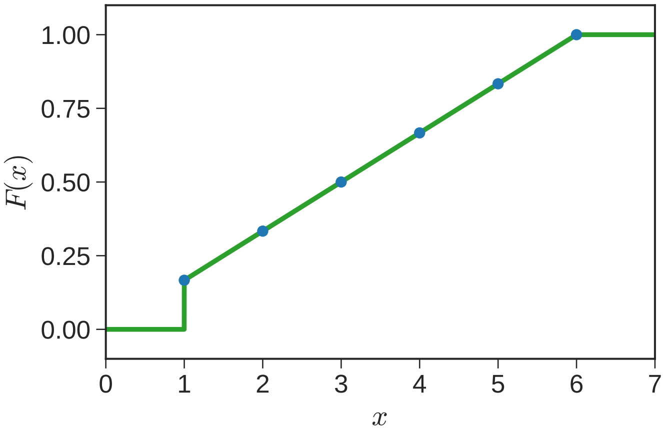
For a discrete PDF, $p(x)$,

$$F(x) = \sum_{X=-\infty}^{x} p(X)$$

For the roll of a die:

$$F(x) = \begin{cases} 0 & x < 1 \\ x/6 & 1 \leq x \leq 6 \\ 1 & x > 6 \end{cases}$$

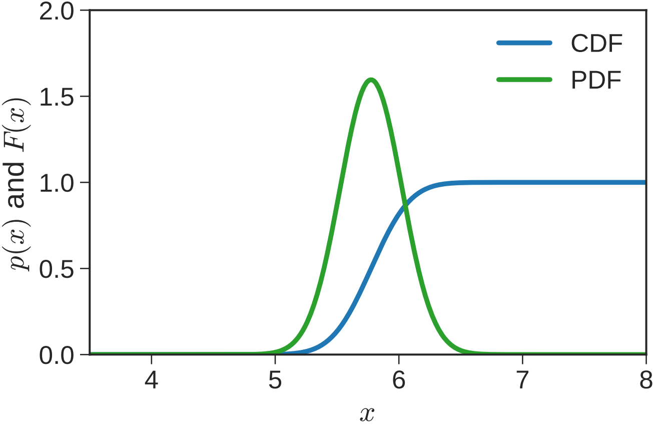
Cumulative Distribution Function



Cumulative Distribution Function

For a continuous PDF,

$$F(x) = \int_{-\infty}^{x} p(X) dX$$



CDF and PDF

We can get the PDF from the CDF:

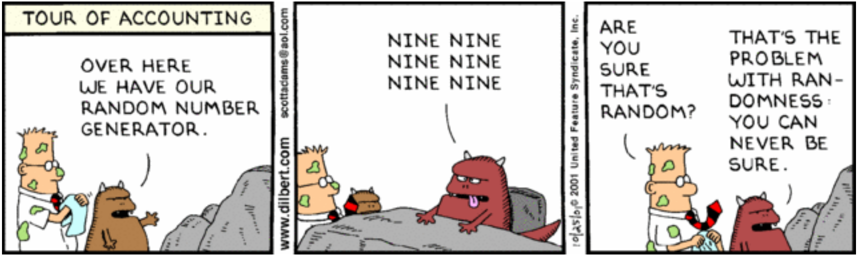
Discrete distribution

$$p(x) = F(x) - F(x-1)$$

Continuous distribution

$$p(x) = \frac{dF(x)}{dx}$$

Pseudorandom Numbers



dilbert.com

We will first discuss the generation of pseudo-random numbers $x \in [0, 1]$ from the continuous uniform distribution.

Once we have these, we can sample other distributions.

Mean and Variance

The **mean** or **expected value** of a random variable

Discrete distribution

$$E[x] = \sum_{x=-\infty}^{\infty} xp(x)$$

Continuous distribution

$$E[x] = \int_{-\infty}^{\infty} xp(x)dx$$

The **variance** of the random variable

$$V[x] = E[x^2] - E[x]^2.$$

Expected Value of $g(x)$

Expected value of a function $g(x)$ of the random variable:

Discrete distribution

$$E[g] = \sum_{x=-\infty}^{\infty} g(x)p(x)$$

Continuous distribution

$$E[g] = \int_{-\infty}^{\infty} g(x)p(x)dx$$

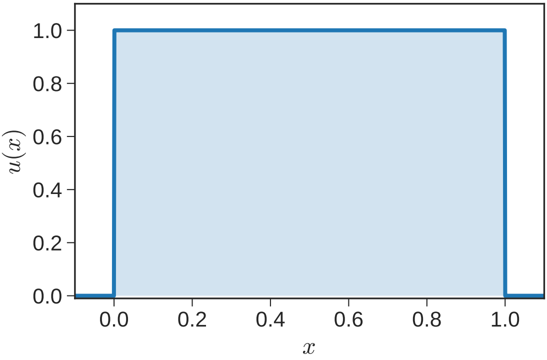
Sometimes a subscript denoting the particular probability distribution is added,

$$E[g] \to E_p[g]$$

Uniform Random Number Generators

PDF is given by:

$$u(x) = \begin{cases} 1, & 0 \leq x < 1 \\ 0, & \text{elsewhere} \end{cases}$$



Uniform Random Number Generators

The basic idea is to "toss a coin" for a (say) 32-bit binary number, so that each of the 2^{32} possibilities

$$\begin{gathered} 0000\ldots000 \\ 0000\ldots001 \\ 0000\ldots010 \\ 0000\ldots011 \\ \vdots \\ 1111\ldots111 \end{gathered}$$

is visited in an "apparently" random fashion.

Usually the random numbers are based on a deterministic algorithm and hence called "pseudo"-random.

Linear Congruential Generators

The simplest RNGs are linear congruential sequence RNGs.

They involve multiplication and truncation of leading bits of an integer.

$$n_{i+1} = (a n_i) \mod m,$$

where n_i is an integer, a is the multiplier, and m is the modulus.

$x \mod y$ is the modulo or remainder operator
e.g., $8 \mod 3 = 2$, $5 \mod 5 = 0$.

n_0, the initial seed, has to be supplied. Thus, a particular choice of a and m specify a particular method.

To get a real number between 0 and 1, compute n_i/m, which is guaranteed to be less than 1.

LCG Example

A Python program

```
def LinCongGen(a, m, n0, num):
    # returns "num" integers "n"
    n = np.zeros(num, dtype=int)
    n[0] = n0

    for i in range(1,num):
        n[i] = a * n[i-1] % m # % = modulo operator

    return n
```

Not all choices of a and m result in a good RNG.

For an example of a bad choice, consider $a = 3, m = 7$.

LCG Bad Choice Example

```
print(LinCongGen(3, 7, 12, 50))
```

```
[12  1   3   2   6   4   5   1   3   2   6   4   5
     1   3   2   6   4   5   1   3   2   6   4   5
     1   3   2   6   4   5   1   3   2   6   4   5
     1   3   2   6   4   5   1   3   2   6   4   5
     1]
```

Notice that the sequence 1, 3, 2, 6, 4, 5 is periodic.

Let us try a different seed $n_0 = 8$.

```
print(LinCongGen(3, 7, 8, 50))
```

```
[8   3   2   6   4   5   1   3   2   6   4   5
 1   3   2   6   4   5   1   3   2   6   4   5
 1   3   2   6   4   5   1   3   2   6   4   5
 1   3   2   6   4   5   1   3   2   6   4   5
 1]
```

In general, we choose an m that is very large.

Consider a better choice: $a = 16807$ and $m = 2147483647$.

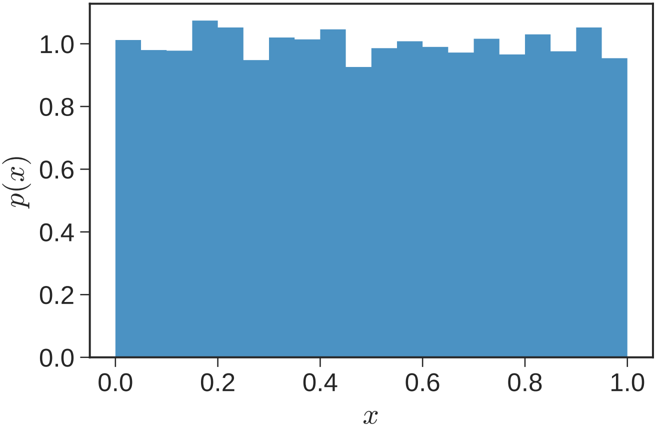
```
a = 16807
m = 2147483647
x = LinCongGen(a, m, 12, 10000).astype(float)/m

n, bins, patches = plt.hist(x, 20, density=True)
```

Note I am converting the integers to real numbers by dividing by m.

LCG: Better choice

Histogram of 10,000 numbers



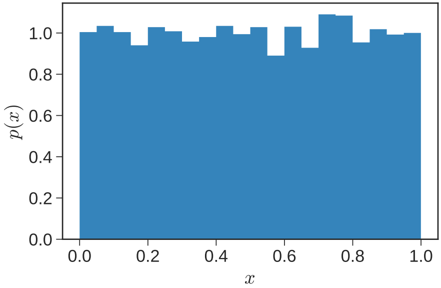
Flat histograms, while a useful diagnostic measure, is not sufficient.

RANDU

A famous disastrous bad choice is the RANDU RNG, which used $a = 65539$ and $m = 2^{31}$.

The histogram looks flat enough ...

```
a = 65539
m = 2**31
x = LinCongGen(a, m, 12, 9999).astype(float)/m
```



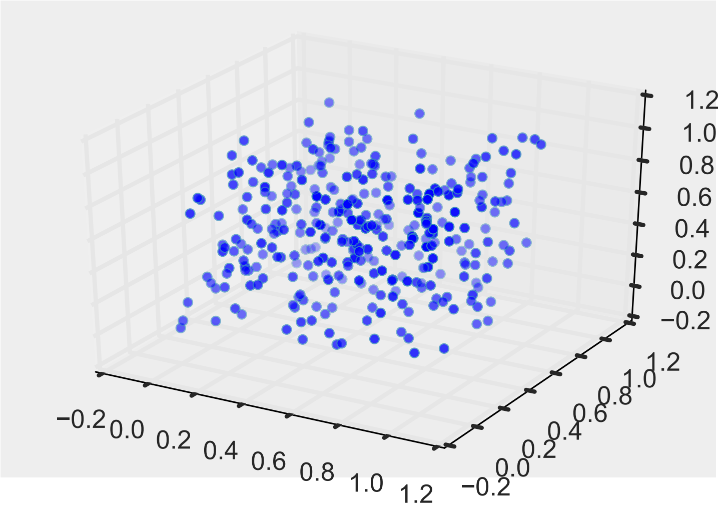
RANDU

However, if you plot three successive deviates in 3D, ...

```
a = 65539
m = 2**31
x = LinCongGen(a, m, 12, 999).astype(float)/m

from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
ax.scatter(x[0::3], x[1::3], x[2::3],'.')
```

RANDU



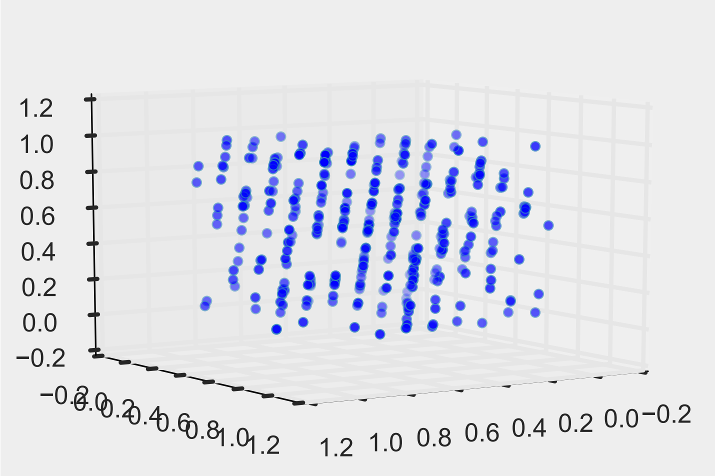
RANDU

... you see a disturbing pattern, if you look at it from the right angle.

```
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
ax.view_init(9, 57) # this adjusts the viewing angle
ax.scatter(x[0::3], x[1::3], x[2::3],'.')
```

George Marsaglia, PNAS 1968,
Random Numbers Fall Mainly in the Planes
SCRI (predecessor of DSC), Statistics, FSU
1924-2011

RANDU



What makes for a good RNG?

Meets a set of statistical tests.

- What is the period of the algorithm?
- Is histogram uniform?
- Systematic correlation of deviates?
- Florida State: Marsaglia, Diehard Battery of Tests for Randomness

Good RNGs include:

- Mersenne-Twister
- SIMD-oriented Fast Mersenne-Twister
- Well Equidistributed Long-period Linear (WELL)
- Xorshift

Standard Distributions

Many libraries are available to generate random numbers from standard distributions.

If you can't find these for your system, do not fret.

As long as you have a decent uniform random number generator, you can generate random numbers from any other distribution.

Standard Distributions

For concreteness, let us catalog some standard distributions.

Discrete

- Uniform
- Binomial
- Poisson

Continuous

- Uniform
- Gaussian
- Exponential

Standard Distributions

For concreteness, let us catalog some standard distributions.

Discrete

- Uniform
- Binomial
- Poisson

Continuous

- Uniform
- Gaussian
- Exponential

Discrete: Binomial

Consider N trials of an experiment with possible outcomes "success" or "failure" ("heads" or "tails", 0 or 1 etc.)

Suppose the probability of "success" is p

The discrete random number n is the number of successes

Probability of a particular outcome say 'ssffs' is $pp(1-p)(1-p)p = p^3(1-p)^2$

Suppose, the order of successes and failures is unimportant (ssffs $\leftrightarrow$ sssff $\leftrightarrow$ sfsfs $\leftrightarrow$ etc.).

Then the number of ways to generate n successes and $N-n$ failures is

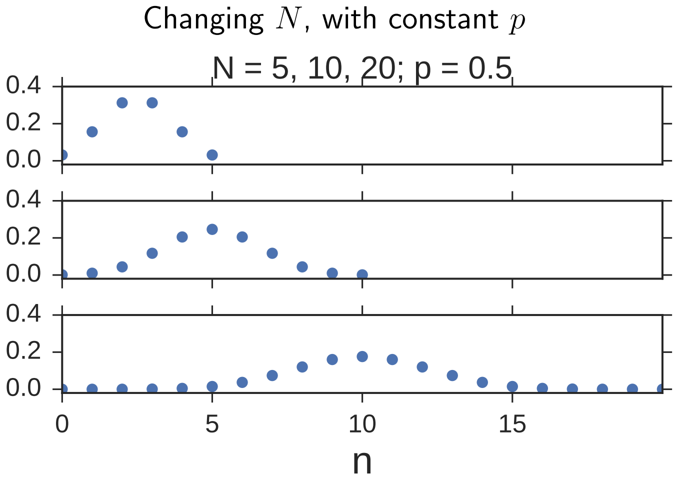
$$^{N}C_n = \frac{N!}{(N-n)!n!}$$

Binomial

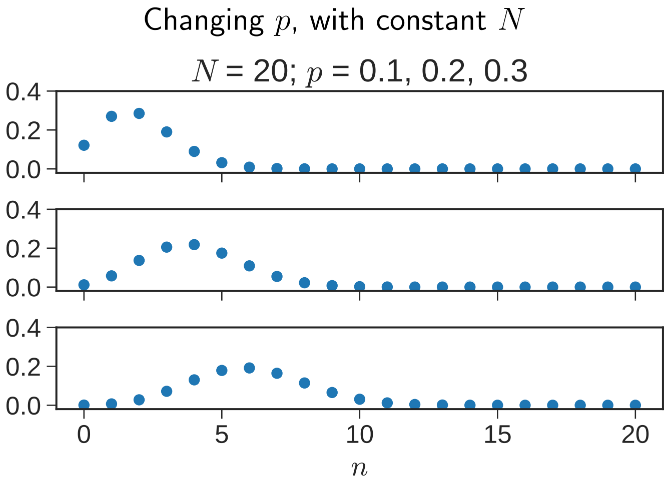
The binomial distribution of n successes in N trials, when p is the probability of success is given by

$$f(n; N, p) = \frac{N!}{(N-n)!n!} p^n (1-p)^{N-n}$$

Binomial



Binomial



$E[x] = Np$ and $V[x] = Np(1-p)$.

Discrete: Poisson

Wikipedia has a succinct description:

"the Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event."

Examples: the number of

- phone calls received by a call center per hour
- taxis passing a particular street corner per hour
- decay events per second from a radioactive source
- homicides in a city in a year

Poisson

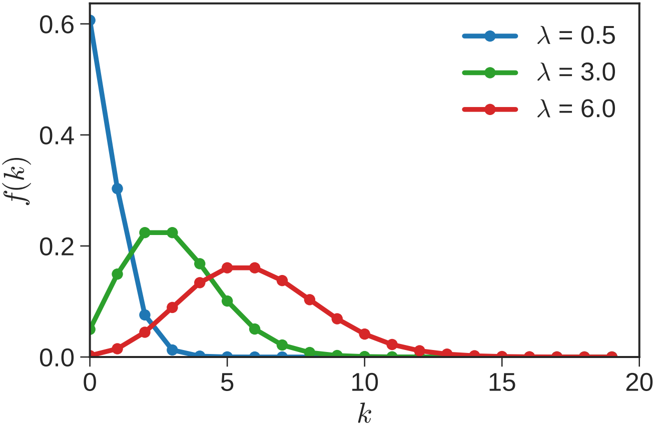
The probability of observing k events in an interval where the average rate is λ is given by:

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, ..., \infty$$

Poisson is related to the binomial distribution, under the following special conditions

$$N \to \infty$$
$$p \to 0$$
$$Np = \lambda$$

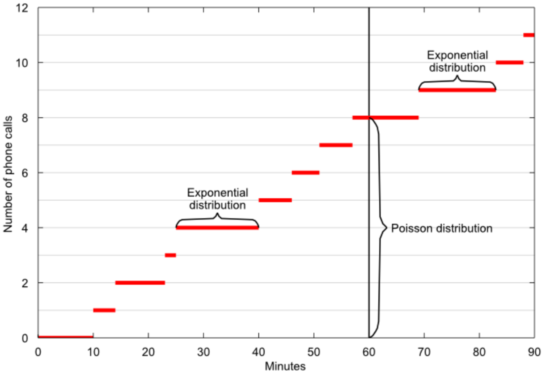
Poisson



The mean and variance of the Poisson distribution are both equal to λ.

Continuous: Exponential

The exponential distribution is closely related to the Poisson distribution.



http://www.statlect.com/uddpoi1.htm

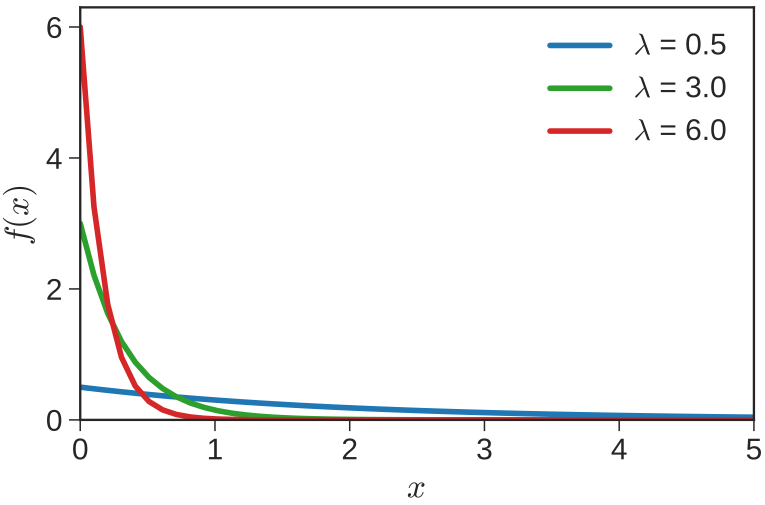
Continuous: Exponential

It describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate.

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

Exercise: Show that the mean and the variance of the exponential distribution are $1/\lambda$ and $1/\lambda^2$, respectively.

Exponential



Lists of Probability Distributions

- NIST has a list of commonly encountered distributions
- Wikipedia has a fairly comprehensive list
- SciPy not only has a comprehensive list, but also has a convenient common interface for plotting, sampling, getting statistics etc.