Continuous distributions

Continuous distributions are defined on a range: \(0 \leq x \leq 1, 0 \leq x, \ldots\)

They can be univariate or multivariate

**Definitions**

\[
\text{Prob}[a < X < b] = \int_a^b f(x) \, dx \\
\text{Prob}[X < b] = \int_{-\infty}^b f(x) \, dx = F(b)
\]

\(f(x)\) is the probability density

\(F(b)\) is the cumulative probability distribution

Hopefully, \(\int_{-\infty}^{\infty} f(x) \, dx = F(\infty) = 1\)

**A continuous probability density may have discrete components:**

\[x = x \Rightarrow \int \frac{dx}{dy} = \int \frac{\sqrt{y}}{dy} = \frac{1}{2\sqrt{y}}\]

**Transforming the scale**

\[
\text{Prob}[x < X < x + dx] \sim f(x) \, dx \\
\text{Prob}[y < Y < y + dy] \sim g(y) \, dy = f(x) \, dx
\]

So, \(g(y) = f(x) \frac{dx}{dy}\)

\(f(x) = 1\) for \(0 < x < 1\), and \(y = x^2 \Rightarrow g(y) = 1 \times \frac{dx}{dy} = \frac{d\sqrt{y}}{dy} = \frac{1}{2\sqrt{y}}\)
A distribution is described by its moments:

- Mean \( \bar{x} = \int_{-\infty}^{\infty} x f(x) \, dx \) ... \( k^{\text{th}} \) moment \( M_k = \int_{-\infty}^{\infty} x^k f(x) \, dx = \mathbb{E}[x^k] \)

- Variance \( \sigma^2 = \int_{-\infty}^{\infty} (x - \bar{x})^2 f(x) \, dx \) ... \( k^{\text{th}} \) central moment \( \int_{-\infty}^{\infty} (x - \bar{x})^k f(x) \, dx = \mathbb{E}[(x - \bar{x})^k] \)

Moments can be infinite: e.g. a Cauchy distribution \( f(x) = \frac{1}{\pi(1+x^2)} \)

\[ \text{Sort}[\text{RandomReal}[\text{CauchyDistribution}[0, 1], 100]] \]
Some examples

- **Uniform distribution**
  
  range \([a, b]\) density \(f(x) = \frac{1}{b-a}\).

  \[
  \text{Sort}[\text{RandomReal}[\text{UniformDistribution}[[0, 1]], 100]]
  \]

- **Exponential distribution**

  range \([0, \infty]\) density \(f(x) = \lambda e^{-\lambda x}\) where \(\lambda\) is the rate

  Mean: \(\bar{x} = \mathbb{E}[x] = \frac{1}{\lambda}\) Variance \(\mathbb{E}[(x - \frac{1}{\lambda})^2] = \frac{1}{\lambda^2}\)

  An exponential is the distribution of the smallest of very many uniformly distributed values

  \[
  \text{tab} = \text{Sort}[\text{Table}[\text{Min}[\text{RandomReal}[\text{UniformDistribution}[[0, 100]], 100]], \{10000\}] ;
  \]

  \[
  \{\text{Mean}[\text{tab}], \text{Variance}[\text{tab}]\}
  \]

  \[
  \text{BarChart}[\text{BinCounts}[\text{tab}, \{0, 5, 0.1\}]]
  \]

  2.5% chance that \(x < 0.025 \bar{x}\); 2.5% chance that \(x > 3.7 \bar{x}\)

  If events occur at exponentially distributed intervals at rate \(\lambda\), the number of events in time \(T\) is Poisson, with expectation \(\lambda T\)

- **Gamma distribution**

  What is the distribution of the sum of two exponentials?

  \[
  f_2(x) = \int_0^x f_1(y) f_1(x-y) \, dy = \int_0^x e^{-y} e^{-(x-y)} \, dy = xe^{-x}
  \]

  This is a convolution: \(f_2 = f_1 * f_1\)

  The sum of \(k\) exponentials has a distribution:

  \[
  f_k(x) = f_1 * f_1 \ldots = \frac{x^{k-1}}{\Gamma[k]} e^{-x} \quad \text{or} \quad \frac{x^{k-1}}{\Gamma[k]} e^{-x} \quad \text{where} \ \Gamma[k] \ \text{is the gamma function}
  \]
More generally, \( f_k(x) = \frac{\lambda^k x^{k-1}}{\Gamma(k)} e^{-\lambda x} \) The mean is \( \frac{k \lambda}{\lambda^2} \) and the variance is \( \frac{k \lambda^2}{\lambda^2} \) \( \frac{\text{variance}}{\text{mean}^2} = \frac{1}{k} \)

- **Normal distribution**

The sum of many independent random variables follows a normal distribution:

\[
f(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-(x-\mu)^2 / (2\sigma^2)}
\]

A linear combination of normal variables also follows a normal distribution
Moment generating functions

The generating function of a discrete distribution is $\mathbb{E}[z^j]$

The moment-generating function of a continuous distribution is $\hat{f}(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x) \, dx$

- Differentiating $\hat{f}$ around zero gives the moments:
  
  \[
  \hat{f}(0) = 1 \\
  \hat{f}'(0) = \mathbb{E}[X] \\
  \hat{f}''(0) = \mathbb{E}[X^2] \\
  \vdots \\
  \hat{f}^{(k)}(0) = \mathbb{E}[X^k]
  \]

- The MGF for the convolution of two distributions is the product of their MGFs:
  
  \[
  f * g = \int_{-\infty}^{\infty} f(y) g(x-y) \, dy \\
  \hat{f} \ast \hat{g} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{tx} f(y) g(x-y) \, dy \, dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{ty} e^{tx} f(y) g(x-y) \, dy \, dx \\
  = (\int_{-\infty}^{\infty} e^{ty} f(y) \, dy) (\int_{-\infty}^{\infty} e^{t \cdot -y} g(z) \, dz) = \hat{f} \hat{g}
  \]

- Examples

  Exponential: $\lambda e^{-\lambda t} \rightarrow \frac{1}{\lambda-t}$

  Gamma: $\frac{\lambda^k t^{k-1}}{\Gamma(k)} e^{-\lambda t} \rightarrow \left(\frac{\lambda}{\lambda-t}\right)^k$

  Normal: $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \rightarrow e^{tx} \sigma^2 t^{1/2}$
Central limit theorem

“I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the “Law of Frequency of Error”. The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.”  

Galton, 1889

The sum of a large number of independent random variables, with finite mean and variance, tends to a normal distribution.

MGF of the distribution of the sum of many independent variables is the product of their MGF:

\[ \tilde{f}_1(t) \tilde{f}_2(t) \tilde{f}_3(t) \ldots \]

Near \( t = 0 \), \( \tilde{f}_i \sim \exp\left( t \mu_i + \left( \frac{t^2}{2} \right) \right) \) so we have overall \( \exp\left( t \sum \mu_i + \sum \left( \frac{t^2}{2} \right) \right) \) which is just the MGF of a normal distribution with mean \( \sum \mu_i \) and variance \( \sum \sigma_i^2 \).
Problems

Problems 1 & 5 are for homework

1) For the following distributions, find the mean, the variance, and the probability of exceeding two standard deviations: a) uniform, b) exponential, c) normal, d) Cauchy.

2) This code gives a table of the lifetimes of elementary particles:

```mathematica
pd = ParticleData[];
lT = Select[Table[ParticleData[pd[[i]], "Lifetime"], {i, Length[pd]}], NumberQ]
```

You can do the same for other properties or other kinds of data - try ?ParticleData or ?GenomeData. (You need to be online for this to work)

This is the number of times 1, 2, … appears in the first digit:

```mathematica
BinCounts[Table[First[RealDigits[lT[[i]]][1]], {i, Length[lT]}], {1, 10}]
```

This is the distribution for the last digit:

```mathematica
BinCounts[Table[Last[RealDigits[lT[[i]]][1]], {i, Length[lT]}], {1, 10}]
```

{78, 76, 130, 101, 59, 98, 41, 107, 196}

Explain why these distributions are different.

3) During the day, there are 6 buses per hour. What is the mean time you have to wait? The variance in time? The probability that you have to wait more than twice the mean? Assume that you arrive at a random time, and that you are i) in Vienna, where the buses run to a regular timetable, ii) in Edinburgh, where buses are uniformly distributed throughout the day or iii) in London, where buses come two at a time, at uniformly distributed times.

4) Molecular evolution can be modelled as the accumulation of mutational changes at a constant rate along each lineage. Different kinds of mammals diverged from each other around ~ 50 million years ago; a typical protein has been accumulating changes in amino-acid sequence at ~ 10^{-6} per year. i) What would you expect the mean and standard deviation of the number of differences between two species to be? ii) In a sample of 100 species pairs, the actual number of differences in amino-acid sequence between a pair of species is observed to have mean 100, and standard deviation 20. Is this significantly higher than expected? What might account for the discrepancy?

5) i) Draw a histogram of the distribution of numbers of heads in 2, 5, 10, 50 coin tosses. What are the mean and variance? ii) Compare these distributions with the normal distribution with the same variance. Choose the scale carefully. iii) Make the same comparison for the sums of 2, … 50 exponential distribution, and iv) for a Cauchy distribution. For iv), it is harder to calculate the distribution mathematically, but you can use Mathematica to explore the problem by drawing random variables.