Probabilities & Data

Week 7: Continuous Distributions

DR NICK FEWSTER-YOUNG

Topics ☺

- Motivate why we want to study continuous distributions.
- Recap quantitative data types and the connection to continuous distributions
- Definitions of expectation, variance and probability density for the continuous setting.
- Calculating probabilities.
- Expectation and Variance

Motivation

Physical quantities are often best described as continuous: temperature, duration, speed, weight, etc. In order to model such quantities probabilistically we could discretise the range to a set of values and represent them as discrete random variables. However, we may not want our conclusions to depend on how we choose the discretisation grid. Constructing a continuous model allows us to obtain insights that are valid for sufficiently fine grids without worrying about discretisation.

More precisely, a continuous random variable is measured over an infinite set of values and by just setting values for the probability of X being equal to individual outcomes, as we do for discrete random variables will lead to us setting zero and nonzero probabilities to specific outcomes of an uncertain continuous quantities. This would result in uncountable disjoint outcomes with nonzero probability. The sum of an uncountable number of positive values is infinite then axioms of probability will collapse or be unreasonable such as the probability over all events is less than 1.

Data Types – Continuous

- Recall that Qualitative Data types are descriptive and are associated to characteristics and are not measured nor counted. They are in fact observed frequencies. Therefore, we cannot numerically measure these data types.

- Quantitative Data Types are measured and can be quantified, so we can place a continuous model and probability mass function to them.

Probability Density Function

- The definition is very similar to the discrete version.
- The probability density function follows the same rules:

➢ Let X be a random variable, and $f_X(x)$ is called a probability density function if

$$0 \leq f_X(x) \leq 1$$

for all x and

$$\int_{x \in S} f_X(x)\, dx \leq 1 \quad for\ all\ x \in S.$$

Probability Density Function

This formal definition is just says that there is a function / curve which traces out the shape of a histogram of a continuous variable such that the area underneath the curve is positive and equals 1.

- If you consider the rate in which $f_X(x)$ changes then you are adding up the area (probability) and this creates what is called as the cumulative distribution function for X. Thus, we denote it as $F_X(x)$ and is defined by

$$F_X(x) \coloneqq \int_{-\infty}^{x} f_X(z)dz = P(X \leq x)$$

Cumulative Distribution Function

Some key properties of the cumulative distribution function for any continuous distribution and any continuous random variable X:

$$\lim_{x \to -\infty} F_X(x) = 0$$

$$\lim_{x \to \infty} F_X(x) = 1$$

$$F_X(b) \geq F_X(a), \quad if \quad b \geq a\,.$$

- This means that the $P(X \leq x) = F_X(x)$, when x is very small (first ordered least likely observation) is zero, and cumulates as x increases to cover all observations. Therefore, the third properties implies that F is non-decreasing.

Example of a Continuous R.V

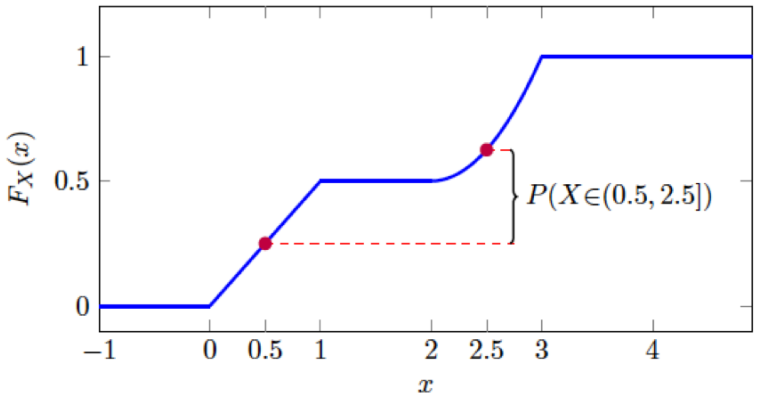
Example: Consider a continuous random variable X which has been found to have a cumulative distribution function given by:

$$F_X(x) := \begin{cases} 0, & x < 0, \\ \frac{x}{2}, & 0 \leq x < 1, \\ \frac{1}{2}, & 1 \leq x < 2, \\ \frac{1}{2}(1 + (x-2)^2), & 2 \leq x < 3, \\ 1, & x \geq 3. \end{cases}$$

Question: Check the assumptions that $F_X(x)$ satisfies a cumulative distribution function. What is the probability that X is between ½ and 2.5.

Example of a Continuous R.V cont.

Question: Check the assumptions that $F_X(x)$ satisfies a cumulative distribution function. What is the probability that X is between ½ and 2.5.



Solution: The assumptions that F_X satisfies a cdf follows since every component of the function is positive and less than 1. Also, it is always increasing. Plus, if add up the area under the curve, it sums to 1. Finally, the probability that X is between ½ and 2.5 is given by

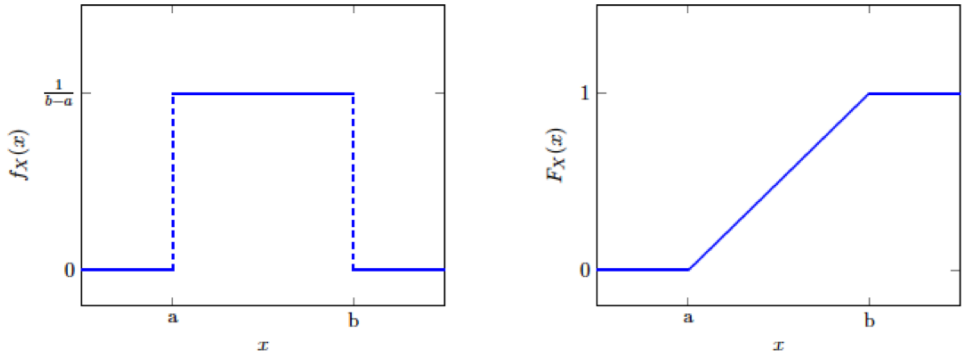
$$P\left(\frac{1}{2} \leq X \leq 2.5\right) = F_X(2.5) - F_X(0.5) = 0.375 \quad \text{(use R to calculate)}$$

Types of Distributions

- Uniform Distribution
- Exponential Distribution
- Normal or Gaussian Distribution
- Beta Distribution
- Plenty more to come ☺

Uniform Distribution

A uniform random variable models an experiment in which every outcome within a continuous interval is equally likely. As a result the probability mass function is constant over the interval. In R, we can produce the two distributions, pdf and cdf respectively.



The mathematical definition of the uniform function defined over the interval a to b is given by:

$$f_X(x) := \begin{cases} \dfrac{1}{b-a}, & a \le x \le b, \\ 0\,, & elsewhere \end{cases}$$

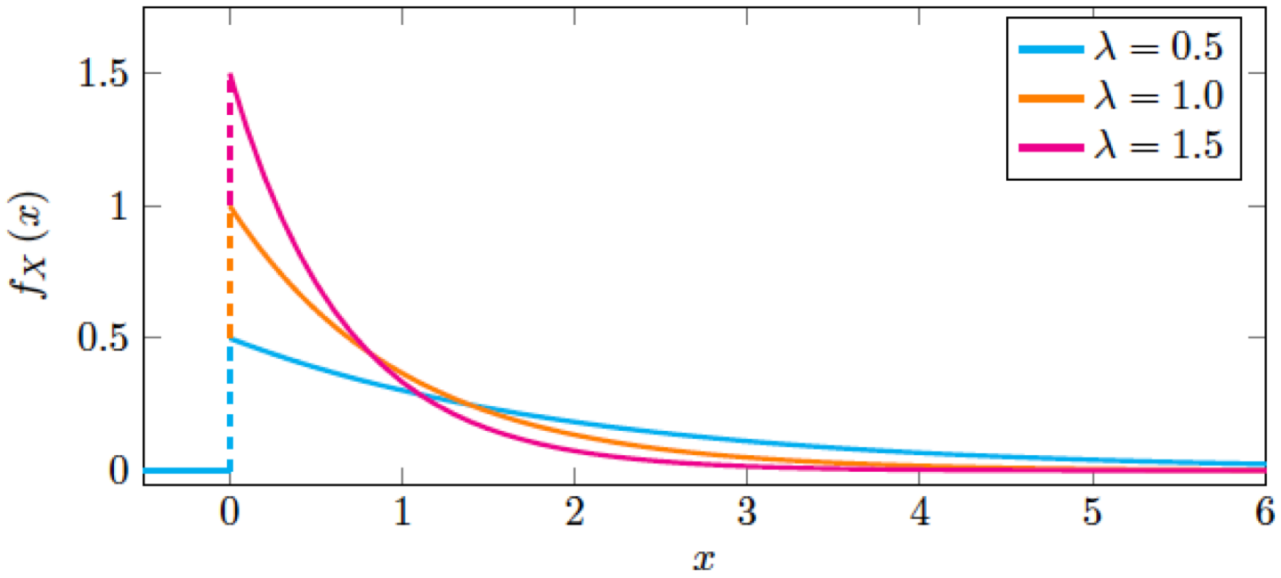
Exponential Distribution

- Exponential random variables are often used to model the time that passes until a certain event occurs. Examples include decaying radioactive particles, telephone calls, earthquakes and many others. It is quite a useful and versatile distribution in modelling real data, and another application is in inter-arrival times of calls at the same call centre occurring between 8 pm and midnight over two days in September 1999.
- An important property of an exponential random variable is that it is memoryless.
- The mathematical definition of the exponential distribution with parameter λ is given below:

$$f_X(x) := \begin{cases} \lambda\, e^{-\lambda x}, & if\ x \geq 0 \\ 0\,, & elsewhere \end{cases}$$

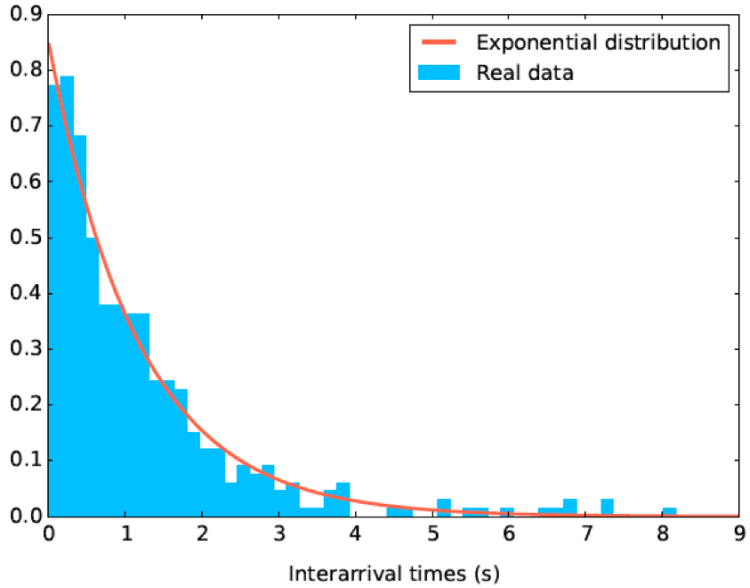
Exponential distribution cont.

Here are the pdfs of three exponential random variables with different parameters below.



Example

The inter-arrival times of calls at the same call centre occurring between 8 pm and midnight over two days in September 1999. Here is a histogram of the real data and a fitted exponential distribution.



Normal or Gaussian Distribution

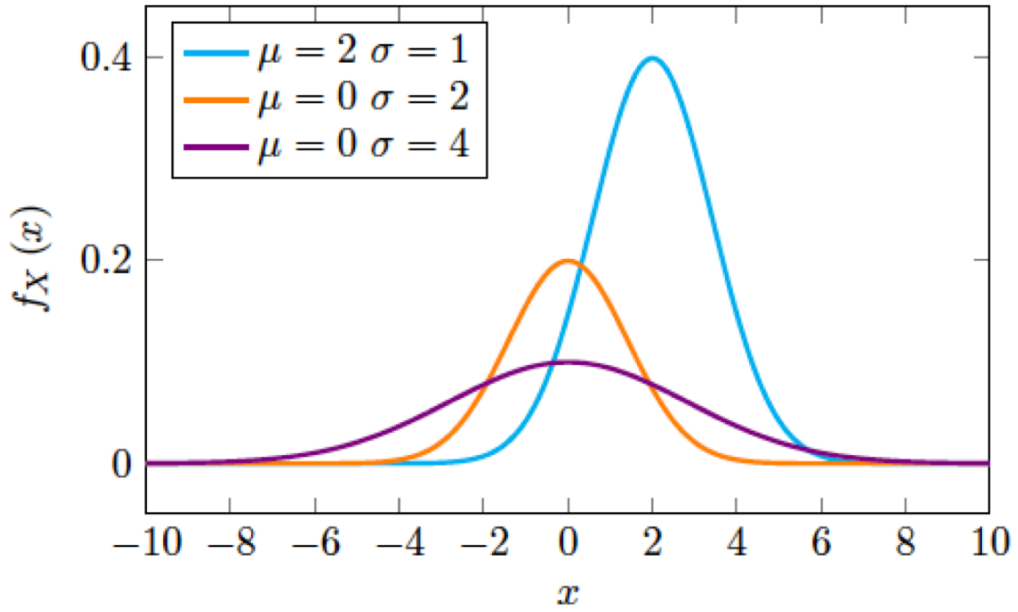
The Gaussian or Normal random variable is arguably the most popular random variable in all of probability and statistics. It is often used to model variables with unknown distributions in the natural sciences, physical sciences and finance. This is motivated by the fact that sums of independent random variables often converge to Gaussian distributions. This phenomenon is captured by the Central Limit Theorem specifically and allows us to estimate population parameters such as the mean.

Definition *(Gaussian Distribution).* The pdf of a Gaussian or Normal random variable with mean and standard deviation, $\mu \; and \; \sigma$ is given by

$$f_X(x) \coloneqq \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

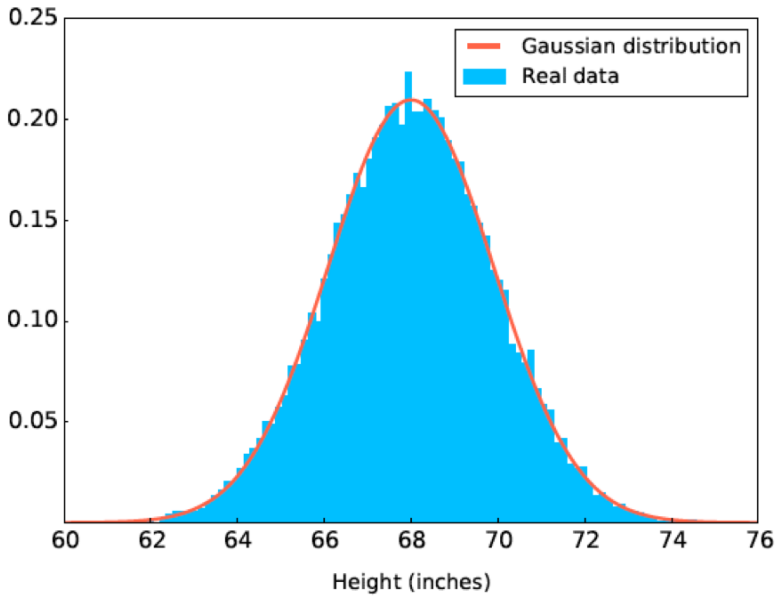
Normal (Gaussian) Distribution

Below are three Normal distribution plots with different means and standard deviations.



Example: Normal (Gaussian) Distribution

Consider the heights (inches) of a population of 25,000 tree branches and if we try to fit a Normal Distribution to it then we see that it is very well approximated by a Gaussian random variable.



That is a crazy function!

- The probability density function is an ugly looking function and faces an annoying consequence which is there is no closed form (expression/formula) for the cumulative distribution function of a Gaussian random variable. This leads to us conveniently most of the time transform the raw data to a standard Normal distribution by

$$Z := \frac{X - \mu}{\sigma}$$

This gives us a Normal Distribution of the random variable Z with mean = 0 and standard deviation 1.

Since we cannot write a closed form of the cumulative distribution function then we represent it by

$$\mathrm{F_X(x)} := \Phi(x),$$

and we use a table or R to calculate / approximate the answer.

Beta Distribution

Beta distributions allow us to parametrize unimodal continuous distributions supported on the unit interval. This is useful in Bayesian statistics, which we will adventure more into in the coming weeks again.

Definition (Beta distribution). The pdf of a beta distribution with parameters a and b is defined as

$$f_X(x) := \begin{cases} \dfrac{x^{a-1}\,(1-x)^{b-1}}{\beta(a,b)} & if \quad 0 \le x \le 1 \\ 0\,, & elsewhere \end{cases}$$

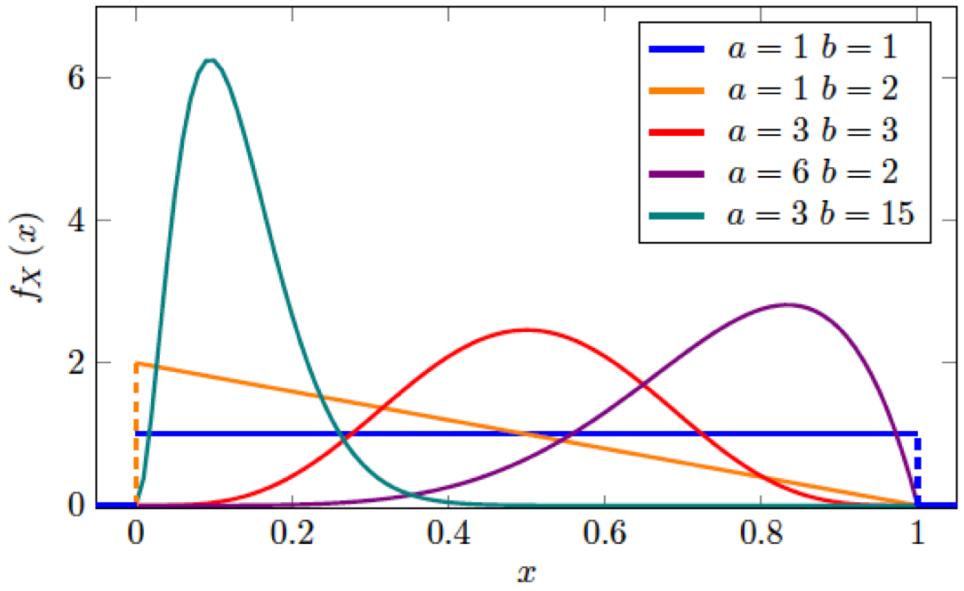
where

$$\beta(a,b) := \int_u u^{a-1}\,(1-u)^{b-1}\,du$$

NOTE: You not have to remember this distribution's formulae.

Beta Distribution

$\beta(a, b)$ is a special continuous function and is called the beta function. A few probability density functions are generated below and observe what happens in the scenario when $a = 1, b = 1$.



Expectations

The definition is analogous and can be built from the discrete scenario. Again, instead of sums by adding up discrete blocks of area (probability), we use integration to absorb the continuous nature of the random variables.

Definition (Expectation for continuous random variables). Let X be a continuous random variable with probability density function $f_X(x)$. The expected value of a function $g(X)$ where $g: R \to R$ is given by

$$E\big(g(X)\big) \coloneqq \int g(x) f_X(x)\, dx$$

Note that in this course, I will not make you do any integration by hand. We will use R!

Variance

The definition again can be written for the continuous scenario since we have a definition for the expectation of a random variable. Therefore,

Definition 4.2.8 (Variance and standard deviation). The variance of X is the mean square deviation from the mean,

$$Var(X) = E\big((X - E(X))^2\big) = \mathrm{E}(\mathrm{X}^2) - \mathrm{E}(\mathrm{X})^2$$

We denote the **standard deviation of a random variable X** by

$$\sigma = \sqrt{Var(X)}.$$

Linearity of Expectations

Theorem 4.1.5 (Linearity of expectation). For any constant $a \in R$, any function $g: R \to R$ and any continuous (or discrete) random variable X

$$E(a\,X) = a\,E(X)$$

For any constants $a, b \in R$ and any functions $g_1, g_2 : R \to R$, and any continuous random variables X and Y

$$E\big(ag_1(X) + bg_2(Y)\big) = aE\big(g_1(X)\big) + bE(g_2(Y))$$

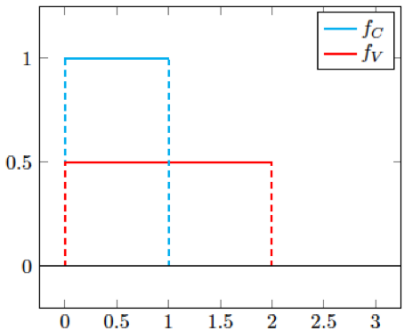
Example: Coffee Beans

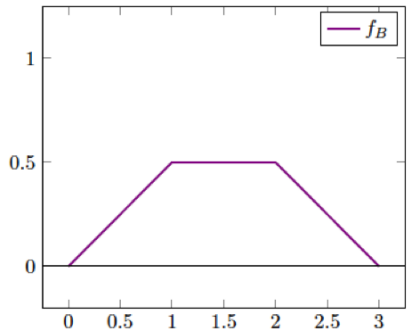
A company that makes coffee buys beans from two small local producers in Colombia and Vietnam. The amount of beans they can buy from each producer varies depending on the weather. The company models these quantities C and V as independent random variables, which is safe assumption and where they have uniform distributions in [0, 1] and [0, 2] (the unit is tons) respectively.

Question: Find the distribution for $B := C + V$ and sketch the distribution.

Answer:

$$f_B(b) := \begin{cases} 0, & if \quad b \leq 0 \\ \frac{b}{2}, & if \quad 0 \leq b \leq 1 \\ \frac{1}{2}, & if \quad 1 \leq b \leq 2 \\ \frac{3-b}{2}, & if \quad 2 \leq b \leq 3 \end{cases}$$





Example: Coffee Beans

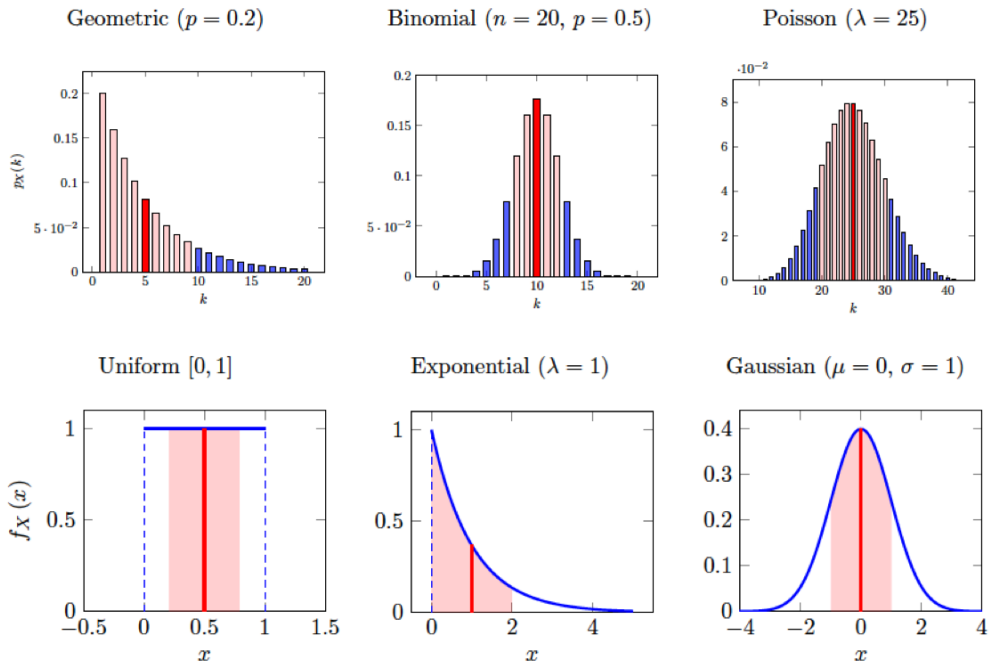
Compute the expected total amount of beans that can be bought.

If C is uniform in [0, 1], so E(C) = 1. Also, V is uniform in [0, 2], so E(V) = 0.5. By linearity of expectation, we have

$$E(B) = E(C + V) = E(C) + E(V) = 1.5$$ ☺

Comparison to Discrete

Below are some graphs of both discrete and continuous distributions which we have looked at so far to see the comparison between the cases.



Summary of Expectations and Variances

Random variable	Parameters	Mean	Variance
Bernoulli	p	p	$p(1-p)$
Geometric	p	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Binomial	n, p	np	$np(1-p)$
Poisson	λ	λ	λ
Uniform	a, b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gaussian	μ, σ	μ	σ^2

That’s it folks for Week 7.

We will continue continuous distributions next week by looking:

- Bivariate continuous distributions, and
- Conditional Probabilities
- Bayesian Statistics in the continuous setting