Continuous Probability Review

Lizzy Huang
Duke University

Week 2

After a week of Bayes' rule and playing with the Binomial distribution, we now move onto the continuous counterpart. While discrete probability distributions only require simple summation, in the continuous case, we will need to do some integrals. However, we will not expect you to know how to perform these calculations by hands. Our focus is always on how to use R to do the job. But before we implement code in R, we need to make sure the mathematical logic is clear.

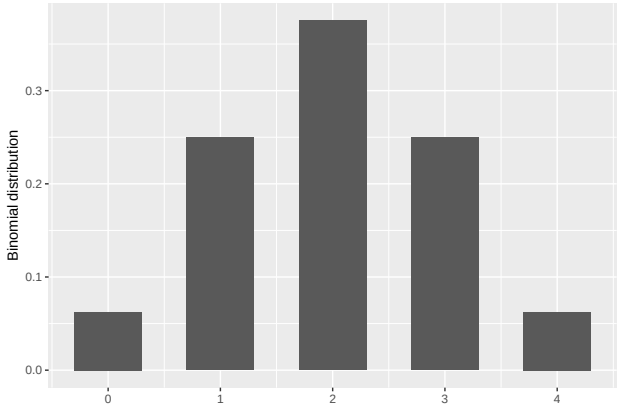
In this **Review**, we will present summary of probability theories for continuous random variables and probability distributions. We will do so by making comparison with the discrete probability distribution, and show that they are analogous.

Discrete Random Variable vs Continuous Random Variable

We have reviewed the concept of random variable in last week. One example that we have seen is the Binomial random variable X, which takes the Binomial probability distribution

$$p(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, 2, \cdots, n.$$

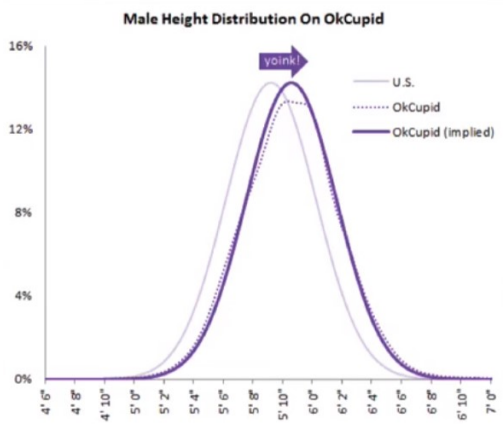
This is a discrete random variable, because the values X can take, denoted as k, can only be discrete numbers. The $p(k)$ immediately gives us the probability for $X = k$. We can visualize these probabilities for different k values using histogram. The following shows the Binomial distribution when $n = 4$ and $p = 0.5$. The height of each bar represents the value of $p(k)$, $k = 0, 1, 2, 3, 4$.



However, many quantites of interest do not just take discrete numbers. For example, when we study the adult male's height data from OKCupid, we should assume the variable "height" can take continuously infinite many values. Namely, a continuous random variable is a variable that can take a continuum of values.

Normal Distribution

In Course 1 *Introduction to Probability and Data*, we see that the male height distribution from OKCupid can be approximated by the Normal distribution, or the bell-curve.



For Normal distribution, so far we have talked about the z-score, the 68-95-99.7% rule, the z-table, and how to use the z-table to look up the p-value. But we never show why these work. So today we will go into more details to understand some critical properties of Normal distribution.

In general, any "bell-curves" or any Normal distribution curves we see can be described by the following family of functions

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

Similar to the Binomial distribution with n and p as the parameters, we have 2 parameters in the Normal distribution μ and σ. Different combinations of values of μ and σ give us different shapes of "bell-curves". Here μ turns out to be the mean of the Normal random variable, and σ is the standard deviation of the Normal random variable. The abbreviation of Normal distribution is

$$\mathsf{N}(\mu, \sigma^2)$$

Unlike the Binomial distribution formula, which gives us the probability of a Binomial random variable taking different values, the Normal distribution formula (??}) does not give us any probability. For example, $f(4) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{4-\mu}{\sigma}\right)^2\right)$ **is not the probability of the Normal random variable** X **taking the value** 4**. It is either not the probability of the Normal random variable** $X \leq 4$**.** We will address this issue when we talk about probability density functions.

z-Score

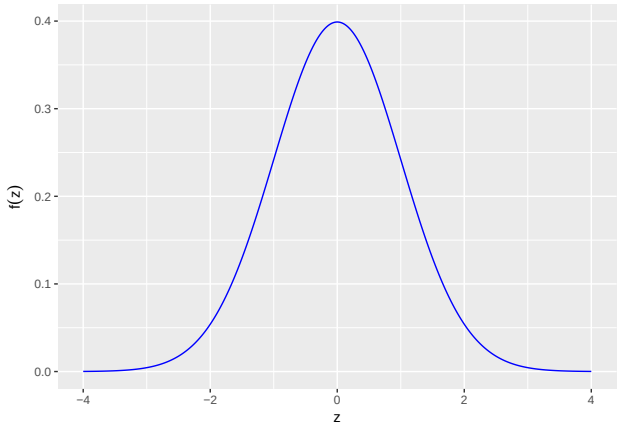
In the Normal distribution formula (1), we can recover the z-score that we are familiar with. Recall that the z-score is defined to be

$$z = \frac{x - \text{mean}}{\text{standard deviation}} = \frac{x-\mu}{\sigma}.$$

It is actually part of (1) inside the exponential function. Since all Normal distributions have "bell-shaped" curves, imagine that we shift and rescale these curves into a standard one, which is called the standard Normal distribution. This standard Normal distribution has mean $\mu = 0$, and standard deviation $\sigma = 1$.

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right). \tag{2}$$

The curve of the standard Normal distribution is symmetric about the y-axis ($x = 0$), with inflection points[1] at -1 and 1. The distance between the symmetrical axis and the 2 inflection points is exactly 1, which is the standard deviation of the Normal distribution.



To standardize any Normal distribution (1) into the standard Normal distribution (2), we use the z-score to substitute $\frac{x-\mu}{\sigma}$. Therefore, the z-score is the link between any Normal distribution to the standard Normal distribution. When we use the z-score, we are switching from the Normal distribution we currently have to the standard Normal distribution, so that we can use the results obtained from the standard Normal distribution, such as the z-table, and the corresponding p-values.

R Code

We can obtain the value of the Normal distribution by using the `dnorm` function in R. This function does not provide the probability of a Normal random variable is below a given value. Instead, it only provides the value of the function $f(x)$ for any given value x. In the following, we set the mean $\mu = 0$ and the standard deviation $\sigma = 0.05$. The value $f(0.1) \approx 1.0798$, which is larger than 1. Apparently, this cannot be any probability.

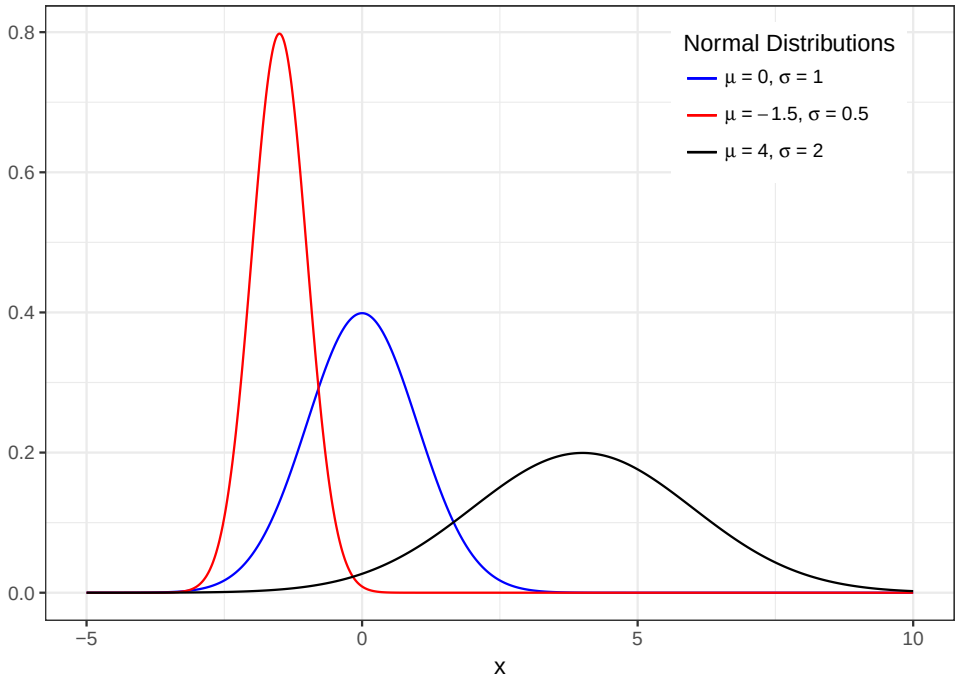
```
norm_value = dnorm(0.1, mean = 0, sd = 0.05)
norm_value
```

```
## [1] 1.079819
```

R also provide other Normal-distribution related functions which calculate the quantiles, the probabilities, and so on. We can simply type `?dnorm` in the R Console for more information.

The following figure gives several Normal distributions with different sets of μ and σ.

[1] Inflection point: a value where the second derivative of a function is 0 or does not exist.



Probability Mass Function (pmf) vs Probability Density Function (pdf)

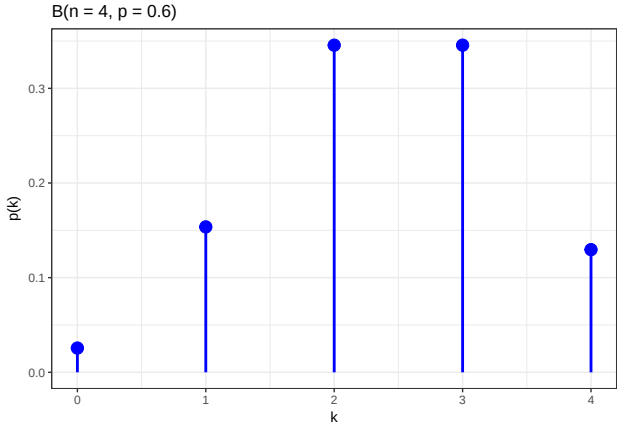
Suppose X is a discrete random variable. Similar to the Binomial distribution, $X = k$ for any values k that X may take, is an event, and $P(X = k)$ calculates the probability for this event to happen. We use a shorter notation $p(k)$ to mean $P(X = k)$, since most of the time, we care more about which value k we are talking. $p(k) = P(X = k)$ for a discrete random variable X is called a probability mass function (pmf). This is an analogue to the concept "mass" in physics.

A point mass is a discontinuous segment in a probability distribution. since each value $p(k)$ for a discrete probability distribution is not continuous, every value of k gives us a point mass. For example, the Binomial distribution $\mathsf{B}(4, 0.6)$, which has the form

$$p(k) = P(X = k) = \binom{4}{k}(0.6)^k(1 - 0.6)^{4-k}, \qquad k = 0, 1, 2, 3, 4,$$

the point masses of this distribution are

$$p(0) = 0.0256, \quad p(1) = 0.1536, \quad p(2) = 0.3456, \quad p(3) = 0.3456, \quad p(4) = 0.1296.$$

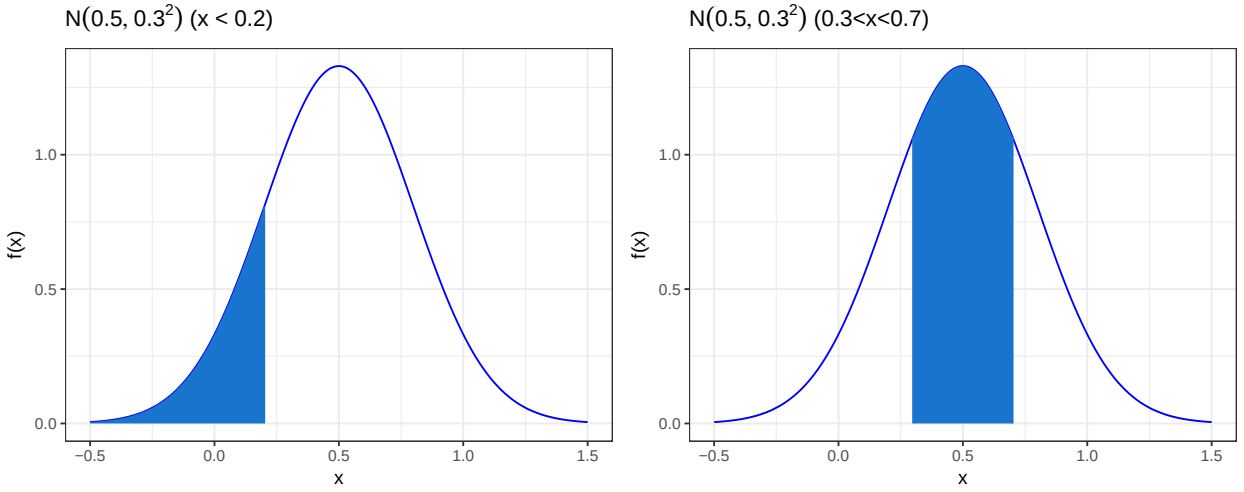


However, when X is a continuous random variable, it no longer makes sense to discuss the probability of X taking exactly some value. Imagine X is the time when a bus comes. There is no way for us to actually measure the probability that the bus would come at exactly 10am sharp, since any measurement of time will not be accurate for a variable that takes continuum of values. Instead, we would ask for the probability that the bus would come within 10 minutes, or between 9:50am to 10am. Therefore, the probability distribution of a continuous random variable no longer provides the probability, instead, it gives us the "density", so that we can "accumulate" this "density" within an interval, to get the probability. Hence, a probability distribution for continuous random variables is called a probability density function (pdf).

For example, when we discussed the p-values in Course **Inferential Statistics**, we say the one-sided p-value for the hypothesis testing of proportion p

$$H_0 : p = 0.5, \qquad H_a : p < 0.5$$

is represented by the lower tail of the "area" under the curve of a Normal distribution. Here, the Normal distribution curve no longer gives the probability, the *area under the Normal distribution curve* gives the probability.



Probability Density Function Notation

Although probability density functions (pdf's) will not provide us the probability, the areas under their curves do. So we still denote them by the lower case p. For the independent variable of this function, we usually prefer to use the corresponding lower case letter to the upper case letter which denotes the random variable. For example, a Normal distribution for a Normal random variable X, can be denoted as

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

Sometimes, we may also specify the parameters and denote the function as

$$p(x; \mu,\ \sigma).$$

However, when the independent variable also uses p as its symbol, we will use $\pi(p)$, instead of $p(p)$ to denote the pdf of the random variable. Here, π is just a symbol for the function, not the π convention we usually use for the following transcendental number.

```
pi
```

```
## [1] 3.141593
```

Learners need to distinguish it from the mathematical context.

In the discrete case, we know that for any discrete random variable X:

$$\sum_{\text{all } k} P(X = k) = 1.$$

As an analogue of the discrete case, we also have

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

This can be interpreted as the very familiar saying, that **the area under the curve of a probability density function is always 1**, since we know that the graphical meaning of integral is "area under the curve".

Cumulative Distribution Function (new)

Since pdf does not provide us the probability, we use a new type of functions which actually represent the area under the pdf's, the cumulative distribution function (cdf).

By definition, the cumulative distribution function $F(x)$ of a continuous probability distribution function $p(t)$ is defined as

$$F(x) = \int_{-\infty}^{x} p(t)\, dt.$$

The graphical meaning of integral is indeed the area under the curve of $p(t)$. Therefore, the cdf defined above represents the area under $p(t)$ from $-\infty$ up to the given value x. That is, the lower tail area under $p(t)$, that is, the probability of the continuous random variable X up to the given value x. Hence, we can interpret cdf $F(x)$ as

$$F(x) = P(-\infty < X \leq x) = \int_{-\infty}^{x} p(t)\, dt.$$

Using cdf, we can calculate the probability of a continuous random variable X between two points L and U, the lower endpoint and the upper endpoint of an interval$[L, U]$

$$P(L \leq X \leq U) = F(U) - F(L) = \int_{L}^{U} p(t)\, dt.$$

The definition of cumulative distribution function also proves that the probability for a continuous random variable X to take any value k is 0.

$$P(X = k) = P(k \leq X \leq k) = F(k) - F(k) = 0.$$

Percentile

In Course **Introduction to Probability and Data**, we introduced the concept of percentile.

> Percentile is the percentage of observations that fall below a given data point. Graphically, percentile is the area below the probability density function curve to the left of that given point, if the observations vollow a continuous distribution.

Now we see that, "percentile", the term that we used in Course 1, is mathematically given by the comulative distribution function $F(x)$

$$\begin{aligned} F(x) =& P(-\infty < X \leq x) = \int_{-\infty}^{x} p(t)\, dt \\ &= \text{area below the curve of the pdf } p(t) \text{ to the left of the given point } x \\ &= \text{percentile of the pdf } p(t) \text{ fall below the given point } x. \end{aligned}$$

For example, the percentile of a pdf fall below a given point 1, is given by $F(1)$.

R Code

In this course, we do not expect you to manually calculate integrals. The good news is, we can use R functions to calculate the probability of traditional probability density functions. For example, suppose X is a standard Normal random variable. The probability of X falling **below** -1, which is given mathematically by

$$F(-1) = P(-\infty < X \leq -1) = \int_{-\infty}^{-1} \mathsf{N}(0, 1^2)\, dt,$$

can be obtained by

```
cdf_neg1 = pnorm(-1, mean = 0, sd = 1, lower.tail = TRUE)
cdf_neg1
```

```
## [1] 0.1586553
```

Similarly, the probability of X falling **below** 1, which is given mathematically by

$$F(1) = P(-\infty < X \leq 1) = \int_{-\infty}^{1} \mathsf{N}(0, 1^2)\, dt,$$

can be obtained by

```
cdf_pos1 = pnorm(1, mean = 0, sd = 1, lower.tail = TRUE)
cdf_pos1
```

```
## [1] 0.8413447
```

Then the probability of X lying between -1 and 1, which is

$$F(1) - F(-1) = P(-1 \leq X \leq 1) = \int_{-1}^{1} \mathsf{N}(0, 1^2)\, dt,$$

is simply the difference between the 2 cdf values

```
cdf_pos1 - cdf_neg1
```

```
## [1] 0.6826895
```

The answer is 0.68, which is part of our 68-95-99.7% rule for the Normal distribution.

Conditional Probability: the Continuous Case (new)

Recall in the discrete case, the conditional probability formula is

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}.$$

This can be viewed as the definition of the conditional probability: the probability of event A conditioning on event B.

In the continuous case, we no longer play with probability mass functions. Instead, we have probability density functions. We can also talk about the probability density function of a continuous random variable X conditioning on another continuous random variable Θ

$$p(x \mid \theta) = \frac{p(x, \theta)}{p(\theta)}.$$

The numerator of the right-hand side is a probability density function of two independent variables x and θ. This is called the joint (probability) distribution. We will provide more details about joint distribution in Week 3.

Preview of New Distributions in Week 2

Besides the Binomial distribution $\mathsf{B}(n, p)$, and the Normal distribution $\mathsf{N}(\mu, \sigma^2)$, we will introduce more probability distribution, including both discrete probability mass functions and continuous probability density functions. Our goal in this week is to impose prior information for the parameters of these distributions. In reality, we often only know the general distribution that the data follow, but are not 100% sure which exact distribution we should pick inside the entire family of distributions. That is, we may know the data follow the Binomial distribution, without knowing what the probability of success p is. We hope to derive posterior information (posterior distribution) for these parameters, by using the prior information (prior distribution) that we impose, and the information from the data (likelihood). So we need new distributions, which will work well with the Binomial distribution family, and the Normal distribution family.

Here is a summary of the distributions we will cover in Week 2.

Discrete Probability Distribution

Distribution	Formula	Mean	Variance	Parameter Meaning
Binomial $\mathsf{B}(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$	n: number of trials p: probability of success
Poisson $\mathsf{Pois}(\lambda)$	$\frac{\lambda^k}{k!} e^{-\lambda}$	λ	λ	λ: rate

Continuous Probability Distribution

Distribution	Formula	Mean	Variance	Parameter Meaning
Normal $\mathsf{N}(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$	μ	σ^2	μ: mean σ: standard deviation

Distribution	Formula	Mean	Variance	Parameter Meaning
Beta $\mathsf{Beta}(\alpha, \beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	α, β: shapes
Gamma $\mathsf{Gamma}(k, \theta)$	$\frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}$	$k\theta$	$k\theta^2$	k: shape θ: scale
Gamma $\mathsf{Gamma}(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\alpha = k$: shape $\beta = 1/\theta$: rate

Remark

1. Medians and modes are important features of distributions. We do not provide the values here, since they are not used as often as means and variances / standard deviations. However, we provide Wikipedia links of each distribution in **Resources** section in this course, which have listed detailed information for each distribution.
2. The Beta distribution shares similar form with the Binomial distribution, but they are not the same. Beta distribution in nature has good "interaction" with the Binomial distribution due to their similar form mathematically, which we will discuss in this Week.

- Standard Uniform distribution $\mathsf{Unif}[0, 1]$ is a special case of the Beta distribution $\mathsf{Beta}(1, 1)$.

3. There are 2 definitions of Gamma distribution, and they are equivalent to each other. In Week 2 lecture, we use the first definition with parameters k and θ. However, you will see the second definition with parameters α and β is more convenient for Bayesian updating. Therefore, we also introduce them in Week 2 Lab.

- Exponential distribution $\mathsf{Exponential}(\lambda) = \lambda e^{-\lambda x}$ is a particular case of the Gamma distribution $\mathsf{Gamma}(\alpha = 1, \beta = \lambda)$.

4. We have listed some "shape" parameters, in both the Beta distribution, and the Gamma distribution. The reason that they are called "shape" parameters is because different values for these parameters will change the general shape of the curve of the distributions. However, they in general do not possess general meaning unless they are used in special cases.
5. $\Gamma(\alpha)$, $\Gamma(\beta)$, $\Gamma(k)$ are all from the Gamma function (not Gamma distribution)

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx.$$

This function do not have explicit closed form in general, unless z is a strictly positive natural number n:

$$\Gamma(n) = (n-1)! \qquad n = 1, 2, \cdots$$

We can use the R function `gamma` for calculation of Gamma function

```
# gamma(4) = (4-1)! = 3*2*1 = 6
gamma(4)
```

```
## [1] 6
```

R Functions for Distributions

We also list here the R functions we can use for these distributions.

1. Binomial $\mathsf{B}(n, p)$: `dbinom` (pmf)

2. Poisson $\mathsf{Pois}(\lambda)$: `dpois` (pmf)
3. Normal $\mathsf{N}(\mu, \sigma^2)$: `dnorm` (pdf), `pnorm` (cdf)
4. Beta $\mathsf{Beta}(\alpha, \beta)$: `dbeta` (pdf), `pbeta` (cdf)
5. Gamma $\mathsf{Gamma}(k, \theta)$ or $\mathsf{Gamma}(\alpha, \beta)$: `dgamma` (pdf), `pgamma` (cdf)

There are some examples.

```
# Poisson `Pois(k = 1, lambda = 2)`
dpois(1, lambda = 2)
```

```
## [1] 0.2706706
```

```
# Beta `Beta(x = 0.5, alpha = 2, beta = 3)`
dbeta(0.5, shape1 = 2, shape2 = 3)
```

```
## [1] 1.5
```

```
# Gamma `Gamma(x = 1, k = 2, theta = 3)`
dgamma(1, shape = 2, scale = 3)
```

```
## [1] 0.07961459
```

```
# Gamma `Gamma(x = 1, alpha = 2, beta = 3)
dgamma(1, shape = 2, rate = 3)
```

```
## [1] 0.4480836
```