Continuous Random Variables and Their Distributions

Cheng Peng

Contents

1 Introduction to Continuous Random Variables

(Sections 4.1 – 4.3)

Continuous random variables take on an uncountably infinite number of values.

Example 1. The following are examples of the continuous random variable.

1. The amount of electricity generated by a nuclear plant in a day
2. The lifetime of an electronic component

The **cumulative distribution function (CDF)** of a random variable Y is defined as

$$F(y) = P(Y \leq y)$$

where $-\infty < y < \infty$.

We have discussed the CDF for discrete random variables in the previous section. The following is another example.

Example 2: Toss a fair coin three times. Let Y denote the number of heads appearing. Noting that

$P(Y=0)=1/8, P(Y=1)=3/8, P(Y=2)=3/8, P(Y=3)=1/8.$

then the CDF is given by

$F(Y\leq 0)=1/8.$

$F(Y\leq 1)=P(Y=0)+P(Y=1)=1/8+3/8=1/2.$

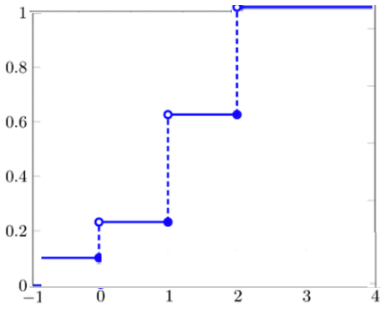
$F(Y\leq 2)=P(Y=0)+P(Y=1)+P(Y=2)=1/8+3/8+3/8=7/8.$

$F(Y\leq 3)=P(Y=0)+P(Y=1)+P(Y=2)=1/8+3/8+3/8+1/8=1.$

The CDF is explicitly given by

$$F_X(x)=\begin{cases}1/8, & x\leq 0;\\ 4/8, & 0<x\leq 1;\\ 7/8, & 1<x\leq 2;\\ 1, & 2<x\leq 3;\\ 1, & x>3.\end{cases}$$

and the plot of the CDF is given by



Properties of the CDF

$$\lim_{y\to\infty} F(y)=1$$

$$\lim_{y\to-\infty} F(y)=0$$

Furthermore, a CDF is a non-decreasing function of y (see the above figure) and is discontinuous for discrete random variables.

1.1 CDF of Continuous Random Variable

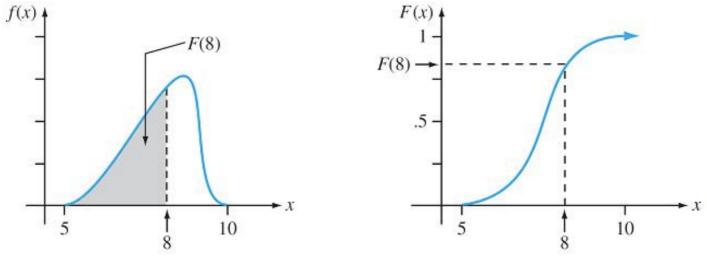
A random variable X having a continuous CDF $F(x)$ is a continuous random variable. The probability density function (pdf) of a continuous random variable Y is defined as ,

$$f(x)=\frac{dF(x)}{dx}=F'(x)$$

where $-\infty<x<\infty$. Thus,

$$F(x) = \int_{-\infty}^{x} f(u)du.$$

The geometry of the above integral is depicted in the following figure.



The area of the shaded region is the CDF of X evaluated at $u = 8$. On the right-hand side is the curve of area vs x.

1.2 Properties of CDF and CDF of Continuous R.V.

Properties

Let $f(x)$ be a real-valued function. If

1. $f(x) \geq 0$ and
2. $\int_{-\infty}^{\infty} f(x)dx = 1$,

$f(x)$ is a probability density function (PDF) of random variable X.

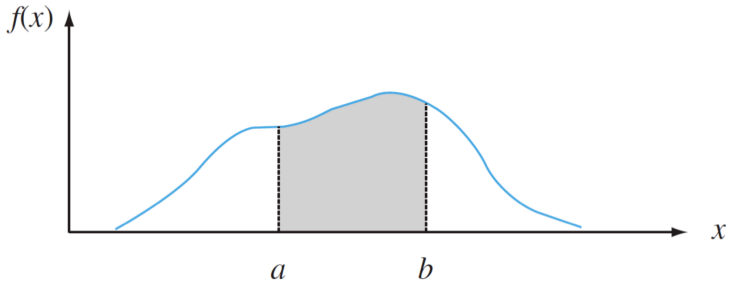
Events and Probability of Continuous random variable:

An event based on the continuous RV is defined to be an interval (including the union of a set of individual values and subintervals).

Example 3: Let X be a continuous random variable with probability density function (pdf) $f(x)$. Define event $E = (a, b)$, then

$$P[a < X < b] = \int_{a}^{b} f(x)dx = F(b) - F(a)$$

This event and the probability distribution are depicted in the following figure.



The probability that X is between a and b is the area of the shaded region.

Example 1: The length of time to failure (in hundreds of hours) for a transistor is a random variable Y with distribution function given by

$$F(y) = \begin{cases} 0, & y < 0; \\ 1 - e^{-y^2}, & y \geq 0. \end{cases}$$

a Show that $F(y)$ has the properties of a distribution function.

b Find the .30-quantile, $\phi_{0.30}$, of Y.

c Find $f(y)$.

d Find the probability that the transistor operates for at least 200 hours.

e Find $P(Y > 100|Y \leq 200)$.

Solution: see the *board-work* in class.

1.3 Expectation and Variance

Let X be a continuous random variable.

1. The expectation or the expected value or the mean of X is defined as

$$\mu = E[X] = \int_{-\infty}^{\infty} yf(y)dy$$

2. The expected value of a function of X, denoted by $g(x)$, is given by

$$E[g(X)] = \int_{-\infty}^{\infty} g(y)f(y)dy$$

3. The variance of a random variable Y is given by

$$V[X] = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)dy$$

The results for the expectation for discrete random variables still hold for continuous random variables (because of the linearity of the integral operator). Let c be a constant and $g(X), g_1(X)$, and $g_2(x)$ be functions of random variable X. Then

1. $E[c] = c$,
2. $E[cg(X)] = cE[g(X)]$,
3. $E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$,
4. $\sigma^2 = V[X] = E[(X-\mu)^2] = E[Y^2] - \mu^2$

Example 2: As a measure of intelligence, mice are timed when going through a maze to reach a reward of food. The time (in seconds) required for any mouse is a random variable Y with a density function given by

$$f(y) = \begin{cases} b/y^2, & y \geq b; \\ 0, & \text{elsewhere.} \end{cases}$$

where b is the minimum possible time needed to traverse the maze.

a. Show that $f(y)$ has the properties of a density function.

b. Find $F(y)$.

c. Find $P(Y > b + c)$ for a positive constant c.

d. If c and d are both positive constants such that $d > c$, find $P(Y > b + d|Y > b + c)$.

e. Find $E[Y]$.

f. Find $V[Y]$.

Solution See the *board-work* in class.

We will present concrete examples when discussing special continuous random variables in the subsequent sections.

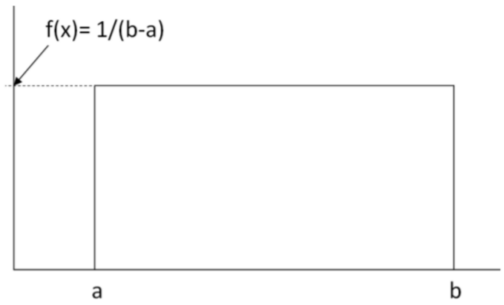
2 Uniform Probability Distribution

(Section 4.4)

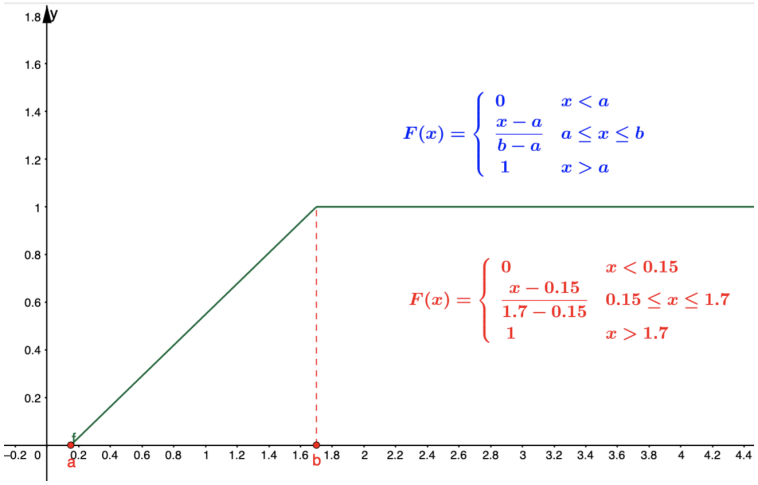
Suppose X can take on any value within the interval $[a, b]$, and each value in the interval is "equally likely" to be chosen (to be more exact, all non-overlapping sub-intervals of the same length that partition the interval $[a, b]$ are equally likely to be selected). Then X has a uniform probability distribution, and its pdf is given by

$$f(x) = \begin{cases} 1/(b-a) & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

The PDF plot is given by



The CDF and its plot are given below.



Expectation and Variance

The mean and variance of the uniform distribution defined on the interval $[a, b]$ are given by

$$\mu = E[X] = \int_a^b x \times \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b x dx = \frac{1}{b-a}(\frac{b^2}{2} - \frac{a^2}{2}) = \frac{a+b}{2}$$

and

$$\sigma^2 = V[X] = \int_a^b [x - (a+b)/2]^2 \frac{1}{b-a} dx = \cdots = \frac{(b-a)^2}{12}$$

A General Remark: Finding probability and finding quantile are two practical questions for all distributions.

Example 4: [*Finding Probabilities*] Suppose the time a friend of yours will show up for a social appointment is uniformly distributed between 5 minutes before and half an hour after the appointed time.

1. What is the probability that he will not be late?
2. What is the probability that he will be at least ten minutes late?
3. What is the probability that he will be at least 20 minutes late, given that he still has not yet arrived five minutes after the appointed time?

Solution: Let X be the *difference between the scheduled time and the arrival time*, then, based on the definition, X is a uniform random variable on $[-5, 30]$. Therefore, the probability density function is given by $f(x) = 1/[30 - (-5)] = 1/35$ for $ -5 < x < 30$; otherwise, $f(x) = 0$.

1. The event "not late" is $[-\infty, 0]$ (arriving at the place before or on time). Therefore, $P[-\infty < X < 0] = P(-5 < X < 0) = [0 - (-5)]/[30 - (-5)] = 5/35 = 1/7$.
2. The event "at least 10 minutes late" $= X > 10$. Therefore, $P(X > 10) = P(10 < X < 30) = [30 - 10]/[30 - (-5)] = 20/35 = 4/7$.
3. This is a conditional probability (pay attention to the "magic word" - **given that**)! We define two events: $A =$ not yet arrived five minutes after the appointed time $= X > 5 = [5, 30]$, B = at least 20 minutes late $= X > 20 = [20, 30]$. Therefore, the joint event A and $B = [20, 30]$ (overlapped part). Note that $P[A \text{ and } B] = (30 - 20)/35 = 10/35 = 2/7$, $P(A) = (30 - 5)/35 = 25/35 = 5/7$. Therefore, using the definition of conditional probability

$$P(B|A) = (2/7)/(5/7) = 2/5.$$

3 Normal Distribution

(Section 4.5)

Let X be a normal random variable or is normally distributed with mean μ and variance σ^2 if the pdf of X is given by

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

with $-\infty < x < \infty$.

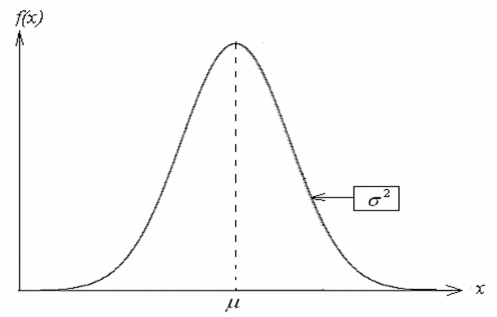
After some algebra, we can check the expectation and variance of X in the following.

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

and

$$V[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \sigma^2$$

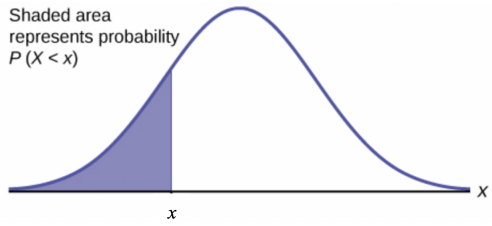
The density curve of the normal distribution has the following form



Special Case - Standard Normal Distribution:

When $\mu = 0$ and $\sigma = 1$, the general normal distribution reduces to the standard normal distribution.

The two basic types of questions: finding probability and finding quantile



Finding Probabilities

Finding $P(X < x)$ (i.e., the left-tail area) for given x.

Finding Percentiles

Finding x from $P(X < x) = p$ for given p (the left-tail area).

There are no formulas for finding probability and quantile based on the normal distribution. We can use either **software programs** or the **standard normal table** to answer the above two types of questions.

In R, there are functions to find probabilities and quantiles.

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
# left-tail probability
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
```

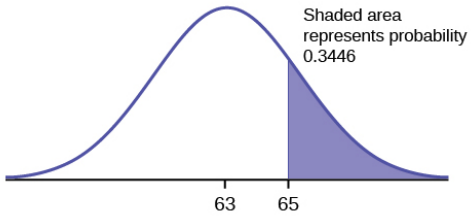
```
# quantile for given left-tail probability
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

Example 5. The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

1. Find the probability that a randomly selected student scored more than 65 on the exam.
2. Find the probability that a randomly selected student scored less than 85.
3. Find the 90th percentile (that is, find the score k that has 90% of the scores below k and 10% of the scores above k).
4. Find the 70th percentile (that is, find the score k such that 70% of scores are below k and 30% of the scores are above k).

Solution: We will use R functions to answer the above questions.

1. We want to find the right-tail area. R function `pnorm` can be used to find either left-tail or right-tail area. By default, it yields the left-tail area.




```
pnorm(q = 65, mean = 63, sd = 5, lower.tail = FALSE)
```

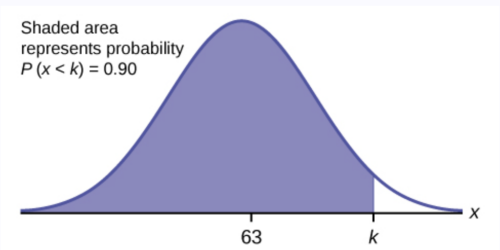
```
## [1] 0.3445783
```

2. This probability is equal to the left-tail area and can be found in the following R function

```
pnorm(q = 85, mean = 63, sd = 5)
```

```
## [1] 0.9999946
```

3. We need to use R quantile function `qnorm()` to find the quantile.




```
qnorm(p = 0.9, mean = 63, sd = 5)
```

```
## [1] 69.40776
```

4. 70th percentile can be found using the following R command.

```
qnorm(p = 0.7, mean = 63, sd = 5)
```

```
## [1] 65.622
```

Example 6. A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

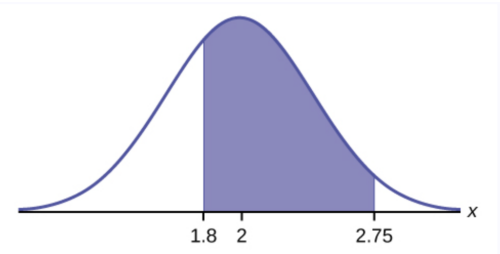
1. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.
2. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Solution. We still use R functions to answer these two questions.

1. We find probability $P(1.8 < X < 2.75)$.

```
pnorm(q = 2.75, mean = 2, sd = 0.5) - pnorm(q = 1.8, mean = 2, sd = 0.5)
```

```
## [1] 0.5886145
```

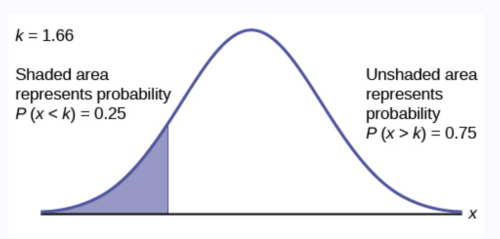


The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

2. This is percentile question. R function `qnorm()` will be used.

```
qnorm(p = 0.25, mean = 2, sd = 0.5)
```

```
## [1] 1.662755
```



The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

4 Gamma Distribution

(Section 4.6)

The gamma distribution contains two special and practically important members: exponential and χ^2 distributions. The PDF of gamma involves a special gamma function.

4.1 The Gamma Function

The gamma function is defined by

$$\Gamma(r) = \int_0^\infty r^{r-1} e^{-t} dt, r > 0.$$

It has the property that

$$\Gamma(r+1) = r\Gamma(r), r > 1.$$

For a positive integer n, $\Gamma(n+1) = n\Gamma(n) = \cdots = n!$.

.

4.2 Definition of Gamma Density

Gamma distribution has two different forms (reparameterization). Our textbook uses the following form

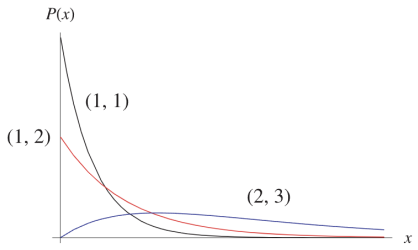
$$f(x) = \begin{cases} \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise} \end{cases}$$

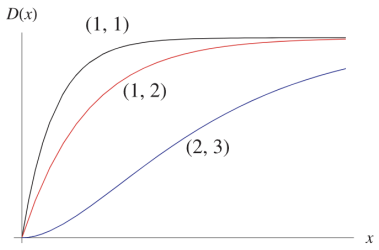
Density curves with various values of parameters α (shape) and β (scale).

The CDF is defined as

$$F(x) = \int_0^y \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)} dy$$

The following figures show the density curves and their corresponding CDF curves based on different values of gamma parameters.





4.3 Expectation and Variance

Using the definition of expectation and variance, we can derive the formulas of the expectation and variance of gamma distribution in the following:

$$E[X] = \int_0^\infty x \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} dx = \alpha\beta$$

and

$$V[X] = \int_0^\infty (x - \alpha\beta)^2 \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} dx = \alpha\beta^2$$

The detailed derivation of the above two formulas can be found on page 187 of the textbook.

4.4 Special Cases

- If $\alpha = 1$, the gamma distribution is reduced to the well-known exponential distribution with the following density function.

$$f(x) = \begin{cases} \frac{e^{-x/\beta}}{\beta} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise} \end{cases}$$

The expectation and variance of the exponential distribution are $E[X] = \beta$ and $V[X] = \beta^2$.

The exponential distribution has been widely used in reliability and survival analysis as a base model since it is mathematically simple.

- If $a = \nu/2$ and $\beta = 2$, the gamma distribution is reduced to the well-known χ^2 distribution with ν degrees of freedom. We define the χ^2 from the normal distribution in subsequent notes.

The expectation and variance of χ^2_ν is simple: $E[X] = \nu$ and $V[X] = 2\nu$.

4.5 Use of Technology

The exponential and χ^2 distributions are special members of the gamma family. R has standalone sets of functions for these distributions.

Special attention should be paid to the form (reparameterization) of the gamma and exponential distributions. R uses the same form as what we used in this note. We will use several examples to show how to use these R functions.

4.5.1 Related R Functions for Gamma Distributions

(Help document: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/GammaDist.html)

```
dgamma(x, shape, rate = 1, scale = 1/rate, log = FALSE)
pgamma(q, shape, rate = 1, scale = 1/rate, lower.tail = TRUE, log.p = FALSE)
qgamma(p, shape, rate = 1, scale = 1/rate, lower.tail = TRUE, log.p = FALSE)
rgamma(n, shape, rate = 1, scale = 1/rate)
```

Example 7. Engineers designing the next generation of space shuttles plan to include two fuel pumps —one active, the other in reserve. If the primary pump malfunctions, the second is automatically brought online. Suppose that the time to failure of the first pump, denoted by X, is a gamma distribution with a mean of 200 and a variance of 400. What are the chances that such a fuel pump system would not remain functioning for the full 50 hours?

Solution. Since $E[X] = \alpha\beta = 200$ and $V[X] = \alpha\beta^2 = 20000$, we solve the shape ($\alpha$) and scale ($\beta$) and obtain $\alpha = 2$ and $\beta = 100$. Therefore, the density function of this gamma distribution is given by

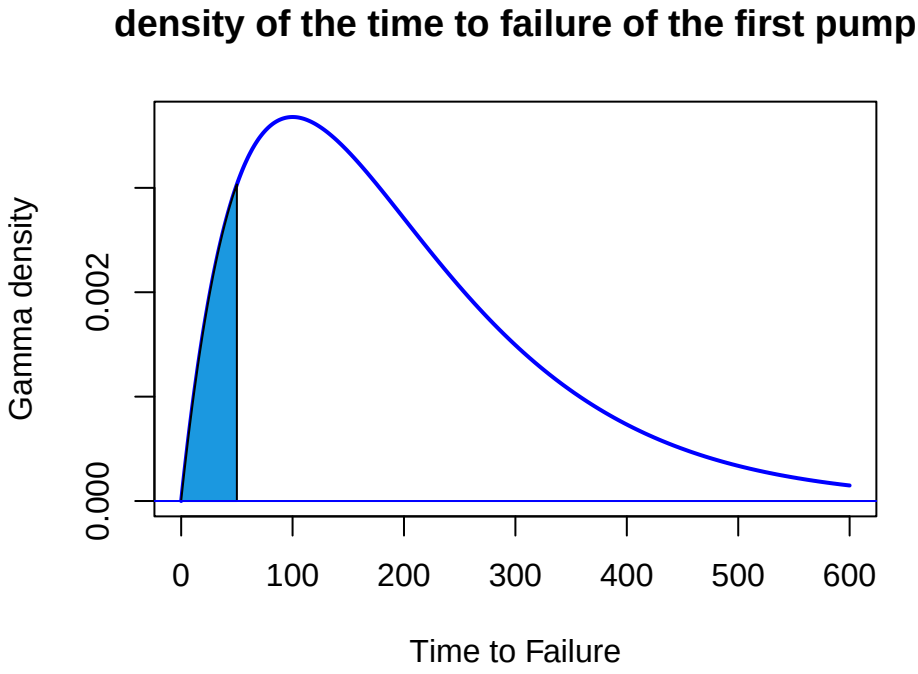
$$f(x) = \frac{1}{100^2\Gamma(2)} e^{-x/100} x^{2-1} = \frac{1}{10000} y e^{-x/100}, \text{ for } x \geq 0.$$

The probability $P(X < 50)$ can be found using R function `pgamma()`.

```
pgamma(q=50, shape = 2, scale = 100)
```

```
## [1] 0.09020401
```

The area of the shaded region in the following density curve is the probability.



4.5.2 Related R Functions for Exponential Distributions

(Help Document: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Exponential.html)

```
dexp(x, rate = 1, log = FALSE)
pexp(q, rate = 1, lower.tail = TRUE, log.p = FALSE)
qexp(p, rate = 1, lower.tail = TRUE, log.p = FALSE)
rexp(n, rate = 1)
```

Example 8. The number of miles that a particular car can run before its battery wears out is exponentially distributed with an average of 10,000 miles. The owner of the car needs to take a 5000-mile trip. What is the probability that he will be able to complete the trip without having to replace the car battery?

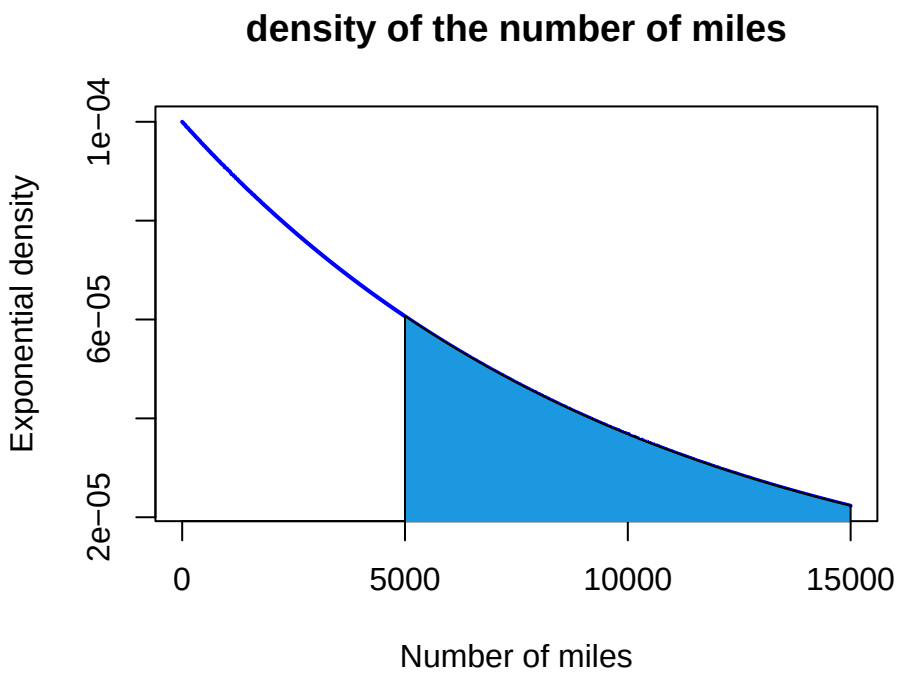
Solution: In R, the exponential related R function uses the following form

$$f(x) = \lambda e^{-\lambda x}, \text{ for } x \geq 0.$$

Based on the above density form, we have $E[X] = 1/\lambda$. Since $1/\lambda = 10000$, $\lambda = 1/10000$. Therefore, the desired probability $P(X > 5000)$ (upper tail) can be found using the following R function.

```
pexp(5000, rate = 1/10000,  lower.tail = FALSE)
```

```
## [1] 0.6065307
```



4.5.3 Related R Functions for χ^2_{df} Distributions

(Help Document: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Chisquare.html)

```
dchisq(x, df, ncp = 0, log = FALSE)
pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qchisq(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
rchisq(n, df, ncp = 0)
```

The χ^2 distribution is one of the most commonly used distributions and will be used to define the χ^2 test. We will not present examples in this note. But it will be used later.

5 Calculus Review: Integrals of Functions

Finding the integral of a function is an **opposite** process of finding a derivative of a function - **antiderivative**.

5.1 Antiderivatives

An antiderivative (sometimes also called `inverse derivative`) of a function f is a differentiable function $F(x)$ whose derivative is equal to the original function $f(x)$.

Example 1. Let $f(x) = 2x$. From the power rule of derivative, we know that $[x^2]' = 2x$. This means that $F(x) = x^2$ is the antiderivative of $f(x) = 2x$. Note also that, $[x^2 + 5]' = 2x$, that is $G(x) = x^2 + 5$. Therefore, the antiderivative of a function not unique. In general, the difference between two antiderivatives of the same original function is a constant.

5.2 Rules and Properties of Integral

5.2.1 Basic Rules

The following are rules of integrals. C is a real number and called **coefficient of integral**.

1. $f(x) = a$, then $F(x) = \int f(x)dx = ax + C$
2. $f(x) = x^k$ (k is a constant and $k \neq -1$), then $F(x) = \int x^k dx = x^{k+1}/(k+1) + C$.
3. $f(x) = 1/x = x^{-1}$, then $F(x) = \int (1/x)dx = \ln(x) + C$
4. $f(x) = e^x$, then $F(x) = \int e^x dx = e^x + C$

4.1. $f(x) = a^x$ ($a > 0$ and $a \neq 1$), then $F(x) = a^x \ln(a)$

5. $f(x) = \ln(x)$, then $F(x) = \int \ln(x)dx = x\ln(x) - x + C$

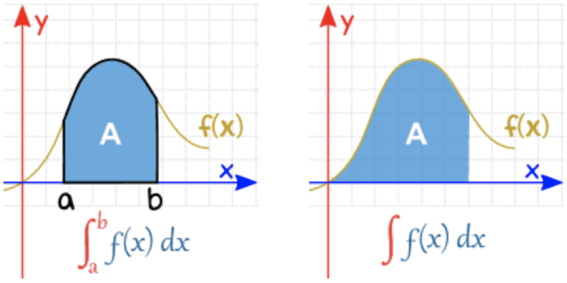
5.2.2 Properties of Integrals

1. Multiplying a constant: $\int af(x)dx = a\int f(x)dx$
2. Additive property: $\int [f(x) + g(x)]dx = \int f(x)dx + \int g(x)dx$
3. Difference property: $\int [f(x) - g(x)]dx = \int f(x)dx - \int g(x)dx$
4. Integral by parts: $\int f(x)dg(x) = f(x)g(x) - \int g(x)df(x)$
5. Change of variable (substitution): $\int f[g(x)]g'(x)dx = \int f[g(x)]dg(x) = \int f(u)du$, where substitution $u = g(x)$.

5.2.3 Definite Integrals

This is the type integrals we used to calculate the probability of an event (i.e., the union of intervals) defined based on a continuous distribution.

A **Definite Integral** has start and end values: in other words there is an interval $[a, b]$. a and b (called limits, bounds or boundaries) are put at the bottom and top of the integral sign $\int$,



Definite Integral (from **a** to **b**)

Indefinite Integral (no specific values)

Let $F(x) = \int_{-\infty}^{x} f(t)dt$, then definite integral $\int_a^b f(t)dt = F(b) - F(a)$. This is called the **Fundamental Theorem of Calculus**. It is used to calculate the definite integral for a given function and the two integral limits.

Some Properties and Rules of Integrals

1. $\int_a^b [f(x) \pm g(x)]dx = \int_a^b f(x)dx \pm \int_a^b g(x)dx$
2. $\int_a^b f(x)dx = -\int_b^a f(x)dx$
3. $\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$ for any constant c (c is not necessarily between a and b).

5.3 Practice Exercises

There two set of exercises you can practice. Ignore all problems involves trigonometric functions.

1. Exercises with answer keys. https://github.com/pengdsci/WCUSTA504/raw/main/topic03/Basic-Integration-Problems.pdf
2. Worksheets: https://pengdsci.github.io/WCUSTA504/Worksheet%20Bundle.pdf

- Worksheet #20: Fundamental Theorem of Calculus.
- Worksheet #21: Definite Integrals