Special Continuous Distributions: Normal and Gamma Families

Cheng Peng

Contents

1 Introduction

We have introduced the procedures for characterizing the general continuous random variables and used the simplest well-known uniform distribution and some not-well-known distributions as examples to illustrate the properties/requirements of continuous distributions.

The following is a summary.

- **Requirements for a probability density function (pdf)**

If function $f(x)$ is a density function some random variable X *if and only if* both of the following conditions are satisfied

1. $f(x) \geq 0$;

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

- **Definition of cumulative distribution function (CDF)**

Let $f(x)$ be a pdf of random variable X, and the CDF of X is defined to be

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

Note that the above CDF is defined based on the cumulative probability. Therefore, we should consider using the CDF to address problems probability associated with continuous random variables.

- **Requirements for cumulative distribution function (CDF)**

if function $F(x)$ is a CDF of a random variable X *if and only if* all of the following conditions are satisfied.

1. $0 \leq F(x) \leq 1$;

2. $F(x)$ is non-decreasing;

3. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

- **Relationship between pdf and CDF**

The relationship between pdf and CDF can be easily observed from the definition of CDF: $F'(x) = f(x)$.

- **Expectations and Properties**

Let $f(x)$ be the density function of X and $g(X)$ is a function of random variable X. Apparently, $G(X)$ is also a random variable.

1. Expectation of X is defined as $\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx$.

2. The expectation of $g(X)$ is defined to be $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$.

As a special case of the above property 2. We can find the variance of X in the following

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

The next example is related to a very well-known distribution. We need to use two important rules of integral: integral by part and substitution.

Example 1: A random variable Y has the density function

$$f(y) = \begin{cases} \theta e^{-\theta y} & \text{if } y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

a). Prove that $f(y)$ is a valid density function.

b). Find the CDF of Y.

c). Find $E[Y]$ and $V[Y]$.

d). Find $E[e^{3Y/2}]$.

Solution See the *board work* in the class.

a). $\int_0^{\infty} f(y)dy = \int_0^{\infty} \theta e^{-\theta y}dy = -\int_0^{\infty} e^{-\theta y}d(\theta y) = -\theta e^{-\theta y}|_0^{\infty} = -(0 - 1) = 1$. This implies that the given function is a valid density function for $\theta > 0$.

We have used substitution in the above derivation (the first equation): if letting $w = -\theta y$ (substitution!), then $\int_0^{\infty} e^{-\theta y}d(-\theta y) = \int_0^{\infty} e^w dw = e^w|_0^{\infty}$.

b). The CDF for Y is defined to be

$$F(y) = \int_{-\infty}^{y} f(x)dx = \int_{0}^{y} \theta e^{-\theta x} dx = -\int_{0}^{y} e^{-\theta y} d(-\theta y) = -e^{-\theta y}|_{0}^{y} = -(e^{-\theta y} - 1) = 1 - e^{-\theta y}.$$

for $y \geq 0$. $F(x) = 0$ if $y < 0$. We have used substitution in the above derivation (the third equation) implicitly.

c). $E[Y] = \int_{0}^{\infty} y\theta e^{-\theta y} dy = -\int_{0}^{\infty} ye^{-\theta y} d(-\theta y)$ (substitution) $= -y \int_{0}^{\infty} yd(e^{-\theta y}) = -[ye^{-\theta y}|_{0}^{\infty} - \int_{0}^{\infty} e^{-\theta y} dy]$ (integral by parts) $= -[(0 - 0) - \int_{0}^{\infty} e^{-\theta y} dy] = (1/\theta) \int_{0}^{\infty} \theta e^{-\theta y} dy = 1/\theta.$

Note that the variance $V[Y] = E[Y^2] - \mu^2$. To find the variance, we only need to calculate the second moment.

$$E[Y^2] = \int_{0}^{\infty} y^2 \theta e^{-\theta y} dy = -\int_{0}^{\infty} y^2 e^{-\theta y} d(-\theta y) dy = -\int_{0}^{\infty} y^2 d(e^{-\theta y}) dy = -[y^2 e^{-\theta y}|_{0}^{\infty} - \int_{0}^{\infty} e^{-\theta y} dy^2] = \int_{0}^{\infty} e^{-\theta y} (2y) dy = 2$$

Therefore,

$$V[Y] = E[Y^2] - \mu^2 = \frac{2}{\theta^2} - (\frac{1}{\theta})^2 = \frac{1}{\theta^2}.$$

d). $E[e^{3Y/2}] = \int_{0}^{\infty} e^{3y/2} \theta e^{-\theta y} dy = \int_{0}^{\infty} \theta e^{3y/2 - \theta y} dy = \int_{0}^{\infty} \theta e^{-(\theta - 3/2)y} dy = \frac{\theta}{\theta - 3/2} \int_{0}^{\infty} (\theta - 3/2) e^{-(\theta - 3/2)y} dy = \frac{\theta}{\theta - 3/2}.$

2 Normal Distribution

(Section 4.5)

Before we introduce the distribution of normal random variables, we introduce the following integral from calculus

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

In Calculus III, we can convert this problem to double integral using change variable in the polar coordinate system. We will not drive this result. You can simply consider this result as theorem through out this course.
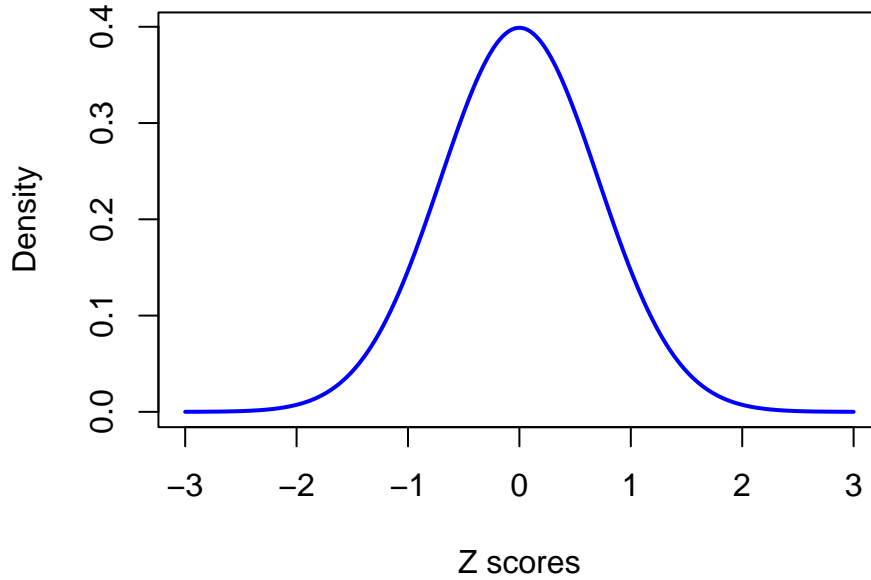
2.1 Standard Normal Distribution

The standard normal distribution is a special normal distribution that has a mean of zero and a standard deviation of 1. It density is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \text{for} \quad -\infty < z < \infty.$$

The density curve sketched in the following

```
z = seq(-3,3,0.01)
fz = (1/(sqrt(2*pi)))*exp(-z^2)
plot(z, fz, type = "l", lwd = 2,
     xlab = "Z scores",
     ylab = "Density",
     col = "blue",
     main = "Standard Normal Density Curve")
```

3

Standard Normal Density Curve

There are different online normal distribution calculators you can use to find probabilities and quantiles of a normal distributions. Here is the one created by R Shiny apps. https://pengdsci.github.io/ISLA/ISLA-03.html.

To validate the density, we need to show

1. $f(z) \geq 0$, and

2. $\int_{-\infty}^{\infty} f(z)dz = 1$.

Since both of the multiplicative terms in $f(z)$ are non-negative, therefore, $f(z) \geq 0$. To calculate in the second integral, we only need to a simple algebraic manipulation.

$$\int_{-\infty}^{\infty} f(z)dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\left(\frac{z}{\sqrt{2}}\right)^2} dz = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-\left(\frac{z}{\sqrt{2}}\right)^2} d(\frac{z}{\sqrt{2}})$$

$$\stackrel{y=z/\sqrt{2}}{=} \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2} d(y) = \frac{1}{\sqrt{\pi}} \times \sqrt{\pi} = 1.$$

This shows that $f(z)$ is a valid density function. We can use the definitions of the expectation and the variance to derive the expectation and variance of the standard normal distribution to be 0 and 1, respectively.

2.2   General Normal Distribution

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

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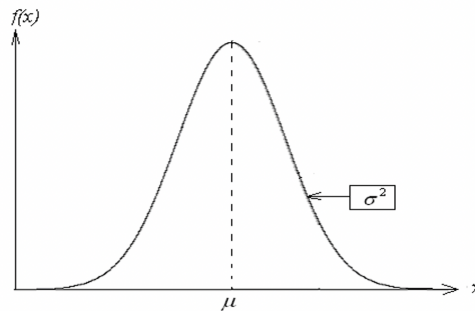
with $-\infty < x < \infty$.

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

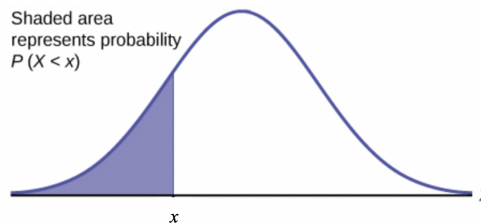$$V[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \sigma^2$$

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 2. 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.
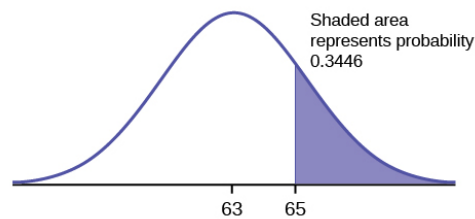

```
pnorm(q = 65, mean = 63, sd = 5, lower.tail = FALSE)
```
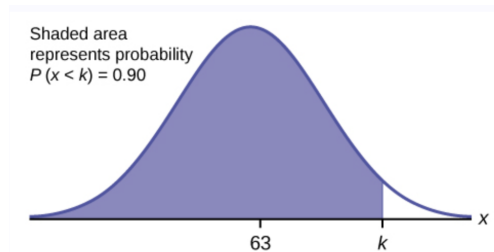
```
## [1] 0.3445783
```

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)
```

6

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

Example 3. 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.
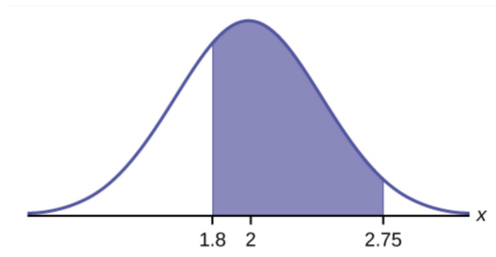
1. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.

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
```


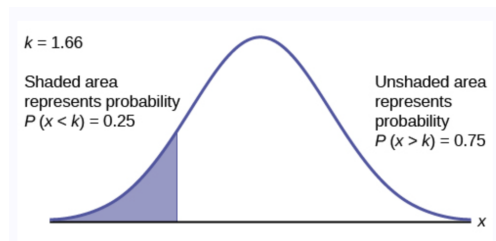The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

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.

3 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.

3.1 The Gamma Function

The gamma function is defined by

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

It has the property that

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

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

.

3.2 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.
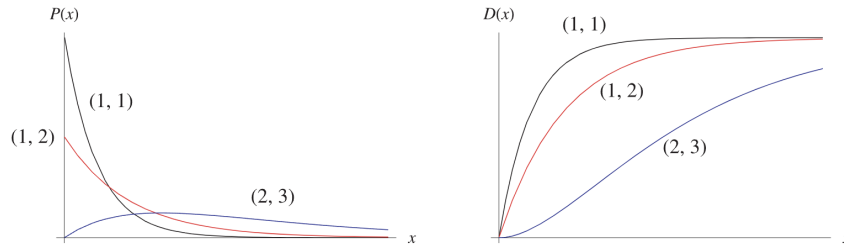
3.3 Expectation and Variance

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.

3.4 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 $V[X] = 2\nu$.

3.5 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.

3.5.1 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 4. 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 20000. 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
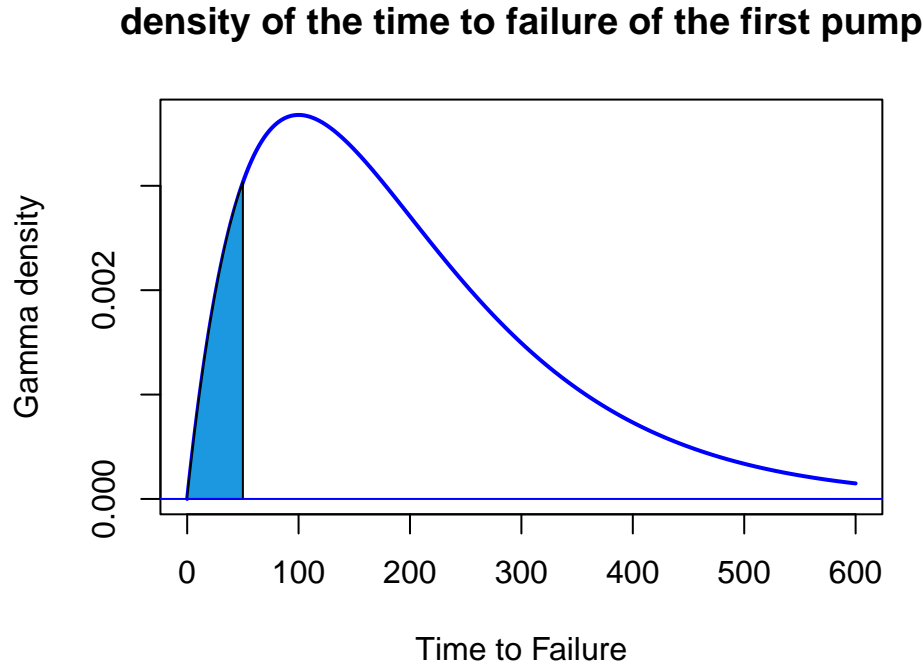
$$f(x) = \frac{1}{100^2\Gamma(2)}e^{-x/100}x^{2-1} = \frac{1}{10000}ye^{-x/100}, \text{ for } x \geq 0.$$

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.

density of the time to failure of the first pump

3.5.2 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 5. 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?
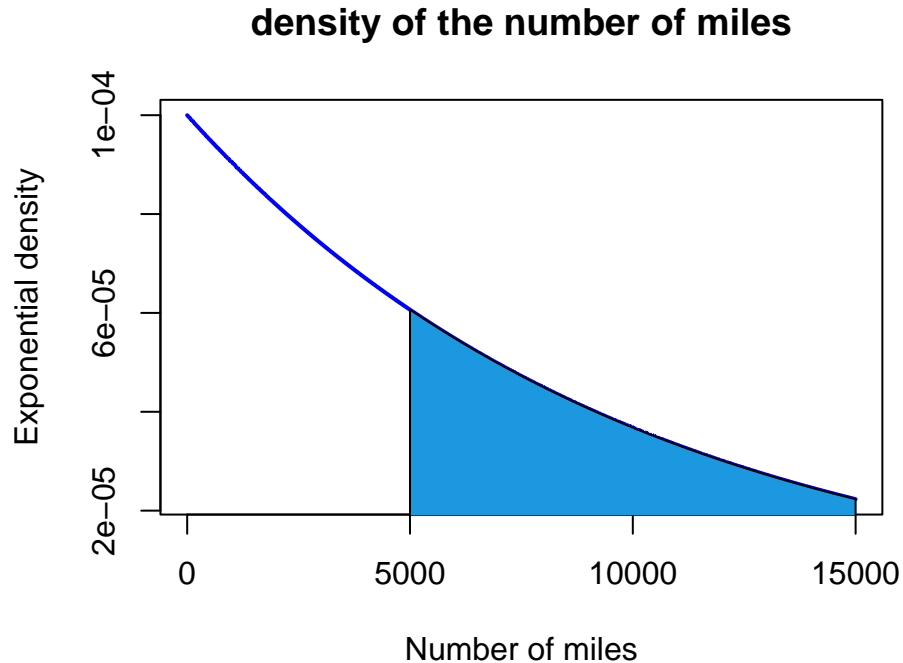
Solution: In R, the exponential related R function uses the following form

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

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
```

density of the number of miles

3.5.3   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.

4   Calculus Review: Integrals of Functions

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

4.1   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.

4.2 Rules and Properties of Integral

4.2.1 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 $F(x) = \int \ln(x)dx = x\ln(x) - x + C$
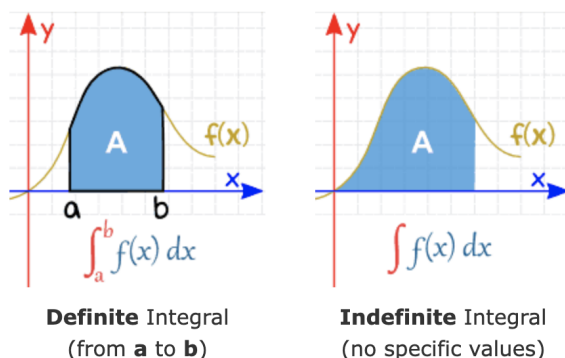
4.2.2 Properties of Integrals

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 $u = g(x)$.

4.2.3 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).

4.3   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