Chapter 7: Continuous probability distri

Introduction

A continuous random variable takes any value within an interval. Examples of a continuous RV: height, weight, distance, time, etc. A histogram can be used to graphically represent the observed distribution of a continuous random variable.

Example (from page 87-89 of your text): $X =$ is height of 27 female students

![Histogram](image)

This is a histogram displaying the distribution of height of 27 female students. There are only 6 classes.

Figure 7.1

Proportion of female students with height between 164.5 and 169.5 cm:

$$P(164.5 \leq X \leq 169.5) = \frac{\text{area of the rectangle with base from 164.5 to 169.5}}{\text{total area of all rectangles}}$$

If we have height $X$ of a large number of female students (e.g., all female students in higher education), then there will be more classes in the histogram, the top becomes smoother and begins to look like a curve.

Figure 7.2

If we draw (scale) Figure 7.2 (in the vertical direction) so that the total area under the curve is 1 square unit, then you should not call the vertical axis "number of students".

In this scaled graph, probability and area can be treated equivalently. For example:

$$P(\text{a randomly selected female student has a height between 164.5 and 169.5 cm}) = \frac{1}{\text{Area under the curve between 164.5 and 169.5}}$$

Figure 7.3

The vertical axis is scaled and the total area under the curve is 1. Continuous prob distri for the variable height.
Note: Continuous probability distributions are usually depicted with smooth lines (curves).

A continuous random variable

- Takes any value within a range
- Example: X: weight of newborn babies
X can take, theoretically, any value between 5 to 13 lb.
- X has a continuous probability distribution, which is usually represented using a smooth curve.

\[
\text{Weight (lb)}
\]

\[ P(X = K) = 0 \rightarrow \text{the probability that } X \text{ takes some particular value is always zero.}\]

Example: \[ P(X = 6) = 0 \]

For any continuous random variable, only probability of an interval makes sense; for example, \[ P(8 \leq X \leq 11) \]

- For any continuous probability distribution
  (a) The area under the curve is 1
  (b) The area of any interval should be between 0 and 1

\[ 0 \leq P(K_1 \leq X \leq K_2) \leq 1 \]

\[ P(a < x < b) = P(a \leq x < b) = P(a \leq x \leq b) \]

There are several types of continuous probability distributions in this chapter, we will consider the two most important ones:

1. The Normal distribution (the Gaussian distribution)
2. The Rectangular distribution (the Uniform distribution)
Normal Distribution

- Is the most important distribution in statistics, because
  
a) many naturally occurring variables tend to have a normal distribution. For example, weight, height, IQ, scores on some tests
  
b) Normal distributions have good properties useful in statistical inference. Sums and averages of variables have approximate normal distributions (this is subject to chapter 8).

- Properties of a normal distribution
  
a) It is a bell-shaped distribution.
  
- $\mu$ = the mean = measure of center
- $\sigma$ = standard deviation = measure of variation

b) It has two parameters

- It is symmetric about the mean $\mu$

- The two shaded areas are equal.

- For any normal distribution, mean = median = mode.

c) Total area under the curve is 1

d) The possible values of a normal random variable: $-\infty$ to $+\infty$
Notation: If $X$ is normally distributed with mean $\mu$ and standard deviation $\sigma$, then we write $X \sim N(\mu, \sigma^2)$.

Example: $X \sim N(50, 25)$

$\mu_x = 50$

$\sigma_x^2 = 25 \Rightarrow \sigma_x = 5$

---

The empirical rule: For any normal distribution, the following is approximately true.

(a) 68% of the observations lie within 1 standard deviation from the mean

\[
p(\mu - \sigma < X < \mu + \sigma) = 0.68
\]

(b) 95% of the observations lie within 2 standard deviations from the mean

\[
p(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95
\]

(c) 99.7% of the observations lie within 3 standard deviations from the mean

\[
p(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997
\]
Summary of the empirical rule:

- 68% within μ ± 1σ
- 95% within μ ± 2σ
- 99.7% within μ ± 3σ

Example:

Suppose X, weight of a fully grown dog, is normally distributed with μ = 44 lb and σ = 8 lb. Find the probability that a randomly selected dog weighs less than 50 lb.

\[ P(X < 50) = \text{Area beneath the curve to the left of 50}. \]

Note:

There are infinite normal distributions with different means (μ) and standard deviations (σ). So, it is impossible to construct probability tables for every normal distribution. We have probability tables for a special kind of normal distribution, the standard normal distribution, which has a mean (μ) of zero and a standard deviation (σ) of one. Then we translate the original prob question into an equivalent normal prob question involving the standard normal distribution.
The Standard Normal Random Variable (Z) is a normal RV with mean \( \mu = 0 \) and std. dev. \( \sigma = 1 \). It is represented by the letter \( Z \):

\[
Z \sim N(0,1)
\]

Any normal RV can be transformed into a standard normal RV using the following transformation:

If \( X \sim N(\mu_X, \sigma_X^2) \),

Then \( Z = \frac{X - \mu_X}{\sigma_X} \sim N(0,1) \).

How to use the Standard Normal Probability Table (C.3(a))

Let \( Z \sim N(0,1) \), Table C.3(a) gives cumulative prob. for standard normal RV.

**Rule 1.** The table gives areas to the left of a positive number only.

\[
P(Z < a) = \text{get it directly from the table if } a \text{ is positive.}
\]

**Example**

a) \( P(Z < 1.33) = 0.9082 \)

b) \( P(Z < 1.7) = 0.9554 \)

c) \( P(Z < 1.42) = 0.9222 \)

d) \( P(Z < 0) = 0.5 \)
Rule 2.

\[ P(Z > a) = 1 - P(Z < a) \]

\text{if } a \text{ is positive}

\text{Example}

\begin{align*}
\text{a) } P(Z > 1.33) &= 1 - P(Z < 1.33) \\
&= 1 - 0.9082 \\
&= 0.0918
\end{align*}

\begin{align*}
\text{b) } P(Z > 1.7) &= 1 - P(Z < 1.7) \\
&= 1 - 0.9554 \\
&= 0.0446
\end{align*}

\begin{align*}
\text{c) } P(Z > 1.42) &= 1 - P(Z < 1.42) \\
&= 1 - 0.9222 \\
&= 0.0778
\end{align*}

\begin{align*}
\text{d) } P(Z > 0) &= 0.5
\end{align*}

Rule 3.

\[ P(Z < a) = 1 - P(Z < -a) \]

\text{if } a \text{ is negative}

\text{Example}

\begin{align*}
\text{a) } P(Z < -1.33) &= 1 - P(Z < 1.33) \\
&= 1 - 0.9082 \\
&= 0.0918
\end{align*}

\begin{align*}
\text{b) } P(Z < -1.7) &= 1 - P(Z < 1.7) \\
&= 1 - 0.9554 \\
&= 0.0446
\end{align*}

\begin{align*}
\text{c) } P(Z < -1.42) &= 1 - P(Z < 1.42) \\
&= 1 - 0.9222 \\
&= 0.0778
\end{align*}
Rule 4: \[ P(z > a) = P(z < -a) \] if \( a \) is negative

Example:

a) \[ P(z > 1.33) = P(z < -1.33) \]
   \[ = 0.9082 \]

b) \[ P(z > 1.7) = P(z < -1.7) \]
   \[ = 0.9554 \]

c) \[ P(z > 1.42) = P(z < -1.42) \]
   \[ = 0.9222 \]

Symmetry:

All normal distributions are symmetric about the mean. What does this mean? It means that

for \( z \sim \mathcal{N}(0,1) \) \( \Rightarrow \)

\[ P(z > a) = P(z < -a) \] for any \( a \)

For \( x \sim \mathcal{N}(\mu, \sigma^2) \) \( \Rightarrow \)

\[ P(x > \mu + a) = P(x < \mu - a) \] for any \( a \)

Example:

\( z \sim \mathcal{N}(0,1) \)

\[ P(z > 2) = P(z < -2) \]

\[ P(z > 1.5) = P(z < -1.5) \]

Example:

\( x \sim \mathcal{N}(10, 2^2) \)

\[ P(x < 10 - 2) = P(x > 10 + 2) \]

\[ P(x > 14) = P(x < 10) \]
Example: Suppose X is weight of premature babies and has a normal distribution with a mean 8.5 lbm and standard deviation 2 lbm.

(a) Find the proportion of premature babies with birth weight less than 4 lbm.

(b) Find the proportion of premature babies with birth weight greater than 8.4 lbm.

(c) Find \( P(X > 4 \text{ lbm}) \)

(d) Find \( P(X < 8.4 \text{ lbm}) \)

Solution:

\[ X \sim N(\mu = 8.5, \sigma^2 = 2^2) \]

(a) \( P(X < 4) = P\left( \frac{X - \mu}{\sigma} < \frac{4 - 8.5}{2} \right) = P(Z < -0.5) \)

\[ = 1 - P(Z < 0.5) \]
\[ = 1 - 0.6915 \]
\[ = 0.3085 \]
\[ \approx 31\% \text{ of premature babies have birth weight < 4 lbm.} \]

(b) \( P(X > 8.4) = P\left( \frac{X - \mu}{\sigma} > \frac{8.4 - 8.5}{2} \right) = P(Z > 0.2) \)

\[ = 1 - P(Z > 0.2) \]
\[ = 1 - 0.5833 \]
\[ = 0.4167 \]
\[ \approx 41.6\% \text{ of the babies have weight > 8.4 lbm.} \]

(c) \( P(X > 4) = P\left( \frac{X - \mu}{\sigma} > \frac{4 - 8.5}{2} \right) = P(Z > -0.5) \)

\[ = P(Z < 0.5) \]
\[ = 0.6915 \approx 69\% \text{ have weight > 4 lbm} \]

(d) \( P(X < 8.4) = P\left( \frac{X - \mu}{\sigma} < \frac{8.4 - 8.5}{2} \right) = P(Z < 0.2) \)

\[ = 0.5793 \]
\[ \approx 96\% \text{ have weight < 8.4 lbm.} \]
Percentiles of a normal distribution

Note: Recall the meaning of percentiles from Chapter 4.

Example:

\[ X : \]

\[ 80\% \]

Smallest observation

\[ k_1 \]

Largest observation

\[ 95\% \]

\[ 5\% \]

\[ X : \]

Smallest

\[ k_2 \]

Largest

\[ X : \]

Smallest

\[ k_3 \]

Largest

\[ k_1 \] is the 80th percentile of \( X \)

\[ P(X \leq k_1) = 0.8 \]

\[ k_2 \] is the 95th percentile of \( X \)

\[ P(X \leq k_2) = 0.95 \]

\[ k_3 \] is the 25th percentile of \( X \)

\[ P(X \leq k_3) = 0.25 \]

\[ k_3 \] is the lower quartile

Percentile of a standard normal distribution (\( Z_k \))

Let \( Z \sim N(0, 1) \)

\[ Z = 0 \]

\[ Z = 0.67 \]

\[ Z = 1.645 \]

\[ Z = -1.645 \]

\[ Z = -0.67 \]

0 is the 50th percentile of \( Z \)

\[ P(Z \leq 0) = 0.5 \]

0.67 is the 75th percentile of \( Z \)

\[ P(Z \leq 0.67) = 0.75 \]

1.645 is the 95th percentile of \( Z \)

\[ P(Z \leq 1.645) = 0.95 \]

-1.645 is the 5th percentile of \( Z \)

\[ P(Z \leq -1.645) = 0.05 \]

-0.67 is the 25th percentile of \( Z \)

\[ P(Z \leq -0.67) = 0.25 \]

Note: Make sure that you understand how to get percentiles of a standard normal distribution from the standard normal probability table (Table A.8(a)).
Let \( X \sim N(\mu_X, \sigma_X^2) \)

To find the \((j \times 100)\)th percentile of \( X \), let us call it \( K \) \((0 \leq j \leq 1)\)

a) First find the \((j \times 100)\)th percentile of a standard normal distribution, call it \( Z_K \)

b) Then \( K \) is obtained as follows

\[
K = \mu_X + \sigma_X Z_K
\]

**Example**

Let \( X \sim N(430, 22^2) \)

\[
\mu_X = 430, \sigma_X = 22
\]

(a) Find the 75th percentile of \( X \)

Let \( K \) be the 75th percentile of \( X \)

\[
P(X \leq K) = 0.75
\]

\[
K = \mu_X + \sigma_X Z_{0.75}
\]

\[
= 430 + 22 \times 0.67
\]

\[
= 444.44
\]

(b) Find the 25th percentile of \( X \)

\[
P(X \leq K) = 0.25
\]

\[
K = \mu_X + \sigma_X Z_{0.25}
\]

\[
K = 430 + 22 (-0.67)
\]

\[
K = 415.26
\]

(c) Find the 92nd percentile of \( X \)

\[
P(X \leq K) = 0.92
\]

\[
K = \mu_X + \sigma_X Z_{0.92}
\]

\[
K = 430 + 22 (1.41)
\]

\[
K = 461.02
\]

This means 92% of the observations are less than or equal to 461.
Do the following exercises

Let \( X \sim N(45, 6) \)

a) Find the 97th percentile
b) Find the 3rd percentile
c) Find the 80th percentile

---

**Normal Approximation to Binomial** (Section 7.7 in your text)

**Recall:** a) If a random variable \( X \) has a binomial distribution, then its probability distribution is given by

\[
p[X=x] = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n
\]

\( n \) = # of trials
\( p \) = prob of success in a single trial
Mean of \( X \): Expected Value of \( X = \mu_X = np \)
Standard deviation of \( X = \sigma_X = \sqrt{np(1-p)} \)

b) To calculate binomial probabilities we use the table (Table D.1).

**Example:** Let \( X \sim B(n=50, p=0.3) \)

\( P(X \leq 20) = 0.9522 \)
\( P(X > 75) = 1 - P(X \leq 75) = 1 - 0.0007 \)
\( P(X < 15) = P(X \leq 14) = 0.4468 \)

**Problem:** How about if \( X \sim B(n=80, p=0.48) \)?

Can we use Table D.1 to find?

\( P(X \leq 70) \) \( P(X > 75) \)

**No.** Because the table does not contain cumulative probabilities for \( n = 80 \) or \( p = 0.48 \).
The only way to answer those questions is by using the formula (model) given above but it is time-consuming.

But there is a simple solution to this problem if certain conditions are satisfied, we can approximate binomial probabilities using normal distribution.
To approximate a binomial RV $X$ by a normal RV $Y$, the following two conditions must be satisfied:

(a) $n*p > 5$
(b) $n*(1-p) > 5$

If these two conditions are met, then:

$X \sim B(n, p) \rightarrow Y \sim N(\mu_Y, \sigma_Y^2)$

$\mu_X = np \rightarrow \mu_Y = np$

$\sigma_X = \sqrt{np(1-p)} \rightarrow \sigma_Y = \sqrt{np(1-p)}$

We are approximating a discrete distribution by a continuous distribution, so we need to use continuity correction.

### Binomial Probability

<table>
<thead>
<tr>
<th>$P(X &gt; r)$</th>
<th>$= P(Y \geq r + 0.5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X \geq r)$</td>
<td>$= P(Y &gt; r - 0.5)$</td>
</tr>
<tr>
<td>$P(X &lt; r)$</td>
<td>$= P(Y &lt; r - 0.5)$</td>
</tr>
<tr>
<td>$P(X \leq r)$</td>
<td>$= P(Y \leq r + 0.5)$</td>
</tr>
</tbody>
</table>

### Example

$x \sim B(n=100, p=0.37)$

Find:
1) $P(X \geq 45)$
2) $P(X < 34)$

### Solution

$n = 100, p = 0.37 \Rightarrow$ no table values

So, we use normal approx., but need to check conditions:

First:

- $np = 37 > 5$
- $n(1-p) = 63 > 5$ \( \Rightarrow \) both conditions are met

$\mu_X = np = 37$

$\sigma_X = \sqrt{np(1-p)} = 4.82$

Approximate $X \sim B(n=100, p=0.37)$ by $Y \sim N(\mu_Y = 37, \sigma_Y^2 = 4.82^2)$
\[ P(X > 45) = P(Y > 45 + 0.5) \]
\[ = P(Y > 45.5) \]
\[ = P(Y - \mu_Y \geq 45.5 - 37) = P(z \geq 1.76) \]
\[ = 1 - P(z < 1.76) \]
\[ = 1 - 0.9608 \]
\[ = 0.0392 \]

\[ P(X \leq 34) = P(Y \leq 34 + 0.5) \]
\[ = P(Y \leq 34.5) = P \left( \frac{Y - \mu_Y}{\sigma_Y} \leq \frac{34.5 - 37}{0.82} \right) \]
\[ = P(z < -0.152) \]
\[ = 1 - P(z < 0.152) \]
\[ = 1 - 0.6985 \]
\[ = 0.3015 \]

See also the example on page 96 of your text for a good example.
Rectangular (Uniform) Distribution

For a continuous random variable $X$, if the probability of every interval of equal length (between $X=a$ and $X=b$) is the same, then $X$ is said to have a uniform distribution on the interval $[a, b]$.

Using symbols: $X \sim \text{Uniform} [a, b]

The probability distribution $f_X$ looks like

![Diagram of a rectangle with base $b-a$ and height $\frac{1}{b-a}$, where the area is $1$.]

Example: Let $X$ be the delivery time for a pizza. Let us assume that $X$ has a uniform distribution on the interval 15 to 45 minutes.

$X \sim \text{Uniform} [15, 45]

This means that the delivery is equally likely to happen any time in that interval.

The probability distribution $f_X$ looks like

![Diagram of a rectangle with base 45-15 and height $\frac{1}{45-15}$, where the area is $1$.]

(a) Find the probability that your next pizza order takes less than 30 minutes.

$p(X<30) = \text{area of the shaded region}

= \text{height} \times \text{base}

= \frac{1}{30} \times 15

= 0.5

(b) $p(X<20) = \frac{1}{30} \times (20-15) = \frac{5}{30} = \frac{1}{6}$
Do the following exercises

1) Let $Y$ be the arrival time for school bus. It is known to have a uniform distribution on the interval 20 to 50 minutes.

$Y \sim \text{uniform } [20, 50]$

The probability distribution is

$\begin{array}{|c|}
\hline
\text{20} & \text{50} \\
\hline
\end{array}$

Find

a) $P(Y \leq 35)$
b) $P(Y > 40)$
c) $P(30 \leq Y \leq 40)$
d) $P(30 \leq Y \leq 30)$
e) $P(40 \leq Y \leq 50)$

The answer for d and e should be the same.

Remember: If $Y$ has a uniform distribution on $[a, b]$, it means that any two intervals of equal length must have the same probability.

In this case, the probability that next bus will arrive in 30 minutes is equally likely to arrive between 20 to 30 minutes and 40 to 50 minutes. These two intervals have the same length.

2) Read and do the example on section 7.6 (page 95).