Chapter 3

Continuous Distributions

3.1 Continuous-Type Data

In Chapter 2, we discuss random variables whose space S contains a countable number of outcomes (i.e. of discrete type). In Chapter 3, we study random variables whose space S contains an interval of numbers (i.e. **of continuous type**).

For a set of continuous-type data, we can group the data into classes of equal width and then construct a histogram of the grouped data.

1. Determine the minimum observation m and the maximum observation M. The **range**
$$R = \text{maximum} - \text{minimum}.$$

2. Divide $[m, M]$ into $k = 5$ to $k = 20$ **class intervals** of equal width, say
$$(c_0, c_1), \quad (c_1, c_2), \quad \cdots, \quad (c_{k-1}, c_k),$$
where $c_0 = m$ and $c_k = M$.

3. The boundaries $c_0, c_1, \cdots, c_k$ are called **class boundaries or cutpoints**. The **class mark** is the midpoint of a class. The **class limits** are the smallest and the largest possible observed values in a class.

The frequency table, frequency histogram, and relative frequency histogram may be used to analyze the grouped data. A **relative frequency histogram** or **density histogram** consists of rectangles, each with base the class interval and area the relative frequency. The height of rectangle with base (c_{i-1}, c_i) is

$$h(x) = \frac{f_i}{n(c_i - c_{i-1})}, \qquad \text{for} \quad c_{i-1} < x \le c_i, \quad i = 1, 2, \cdots, k,$$

where f_i is the number of observed data in the class (c_{i-1}, c_i) and n is the total number of observations.

Ex 1, p.112. (scanned file)

If the data $x_1, x_2, \cdots, x_n$ have a sample mean $\bar{x}$ and sample standard deviation s, and the histogram of these data is "bell shape", then approximately

- 68% of the data are in the interval $(\bar{x} - s, \bar{x} + s)$;
- 95% of the data are in the interval $(\bar{x} - 2s, \bar{x} + 2s)$;
- 99.7% of the data are in the interval $(\bar{x} - 3s, \bar{x} + 3s)$;

The **relative frequency polygon** may be used to study the data.

Ex 2, p.114. (scanned file)

Sometimes class intervals are not required to have equal width.

Ex 3, p.116. (scanned file)

Homework
§3.1 3, 7, 9

Attachment: Scanned textbook pages of Section 3-1

3.2 Exploratory Data Analysis

To explore an unknown distribution, we take a sample of n observations $x_1, x_2, \cdots, x_n$. We can do the same thing as a frequency table and histogram can, but keep the original data values, through a **stem-and-leaf display**.

See Table 3.2-1, p.122 (scanned file). In the "**depths**" column, the frequencies are added together from the low end and the high end until the row is reached that contains the middle value in the ordered display.

If we order the leaves in each row, then we get the **ordered stem-and-leaf display**. We may refine the stems to get a more accurate display.

Ex Tables 3.2-2, 3.2-3, and Ex 3.2-1, p.122 (scanned file)

. . . .

The ordered data of a sample of n observations $x_1, \cdots, x_n$ is called the **order statistics** of the sample. We can label the data by ranks. For example, y_1 (min), $y_2, \cdots, y_n$ (max).

The (100p)-th **sample percentile** of the sample is the $(n+1)p$-th order statistic.

The following three quartiles are of great importance:
25th percentile: **the first quartile** $\widetilde{q}_1$,
50th percentile: **the second quartile (median)** $\widetilde{m} = \widetilde{q}_2$,
75th percentile: **the third quartile** $\widetilde{q}_3$.

The difference between the first and the third quartiles $(\widetilde{q}_3 - \widetilde{q}_1)$ is called the **interquartile range**, **IQR**.

Ex 2, p.124 (scanned file)

The **five number summary** are used to describe a sample of data.

minimum, 1st quartile, 2nd quartile, 3rd quartile, maximum.

It is very convenient to use **box-and-whisker diagram** (i.e. **box plot**) to summarize the data.

Ex 3 and 4, p.125-126 (scanned file)

Fences are used to pick up the extreme values (outliers) of the sample. In a box-and-whisker diagram, construct **inner fences** to the left and right of the box at the distance of 1.5 times the IQR, and **outer fences** at the distance of 3 times the IQR. The observations beyond inner fences are denoted as circles or a dot (•).

Ex 5, p.127 (scanned file)

The box-and-whisker diagram can be used to obtain some functions of order statistics, like **midrange**, **trimean**, **range**, and **interquartile range** (p.127, scanned file).

Homework
§3.2 5, 7, 9, 11

Attachment: Scanned textbook pages of Section 3-2

3.3 Random Variables of the Continuous Type

The relative frequency histogram $h(x)$ of a continuous r.v. X is defined so that the area below the graph and between $[a, b]$ is approximately $P(a < X < b)$. For many continuous r.v. X , there exists a **probability density function (p.d.f.)** $f(x)$ of X such that

1. $f(x) > 0$, $x \in S$.
2. $\int_S f(x)dx = 1$.
3. If $(a, b) \subseteq S$, then $P(a < X < b) = \int_a^b f(x)dx$.

For convenience, we often define $f(x)$ on $(-\infty, \infty)$ by letting $f(x) = 0$ for $x \notin S$. Then $f(x) \geq 0$ for all x, and $\int_{-\infty}^{\infty} f(x)dx = 1$.

Ex 1, p.132 (scanned file)

The **(cumulative) distribution function (c.d.f.)** of a random variable X of the continuous type, defined in terms of the p.d.f. of X, is

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

The c.d.f. $F(x)$ is increasing function with

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

Ex 2, p.133 (scanned file)

Ex 3, p.134 (scanned file)

Let X be a continuous random variable with p.d.f. $f(x)$.

- The **expected value/mean of** X is

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

- The **variance of** X is

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

- The **standard deviation of** X is

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

- The **moment-generating function (m.g.f.)** of X (if exists) is

$$M(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)dx, \qquad -h < t < h.$$

Ex 5, 6, p.136 (scanned file)

The **(100p)th percentile** is a number π_p such that the area under $f(x)$ to the left of π_p is p, that is,

$$p = \int_{-\infty}^{\pi_p} f(x)dx = F(\pi_p).$$

Recall that the 50th percentile is called the **median**, and the 25th and 75th percentiles are called the **first** and the **third quartiles**, respectively.

For empirical data, if $y_1 \le y_2 \le \cdots \le y_n$ are the order statistics associate with the sample $x_1, x_2, \cdots, x_n$, then y_r is called the **quantile of order** $r/(n+1)$ as well as the $100r/(n+1)$**th percentile**. The plot of (y_r, π_p) for several values of r is called the **quantile-quantile plot** or the **q-q plot** (See Fig 3.6-3 on p.165, scanned file.)

Ex 7, 8, p.137-139 (scanned file)

Homework
§3.3 1, 7, 23, 25

Attachment: Scanned textbook pages of Section 3-3

3.4 The Uniform and Exponential Distributions

A r.v. X has a **uniform distribution** if its p.d.f. is equal to a constant on its support.

In particular, if the support is $[a,b]$, then we say that $X \sim U(a,b)$, and

$$\textit{p.d.f.}: \qquad f(x) = \frac{1}{b-a}, \qquad a \le x \le b.$$

$$\textit{c.d.f.}: \qquad F(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \le x < b, \\ 1, & b \le x. \end{cases}$$

$$\text{mean:} \qquad \mu = E(X) = \frac{a+b}{2}.$$

$$\text{variance:} \qquad \sigma^2 = E(X^2) - E(X)^2 = \frac{(b-a)^2}{12}.$$

$$\textit{m.g.f.}: \qquad M(t) = \begin{cases} \frac{e^{tb}-e^{ta}}{t(b-a)}, & t \neq 0, \\ 1, & t = 0 \end{cases}$$

The curves of p.d.f. and c.d.f. for $U(a,b)$ are given in Fig 3.4-1, p.141 (scanned file).

An important uniform distribution is $U(0,1)$.

Ex 1, p.142 (scanned file)

Now we explore the exponential distribution. Suppose in a process the number of changes occurring in a unit time interval has a Poisson distribution with mean λ. This is a discrete random variable. However, the waiting times W between successive changes is a continuous random variable. The space of W is $[0,\infty)$. The distribution function of W is

$$\begin{aligned} F(w) &= P(W \le w) = 1 - P(W > w) = 1 - P(\text{no changes in [0,w]}) \\ &= 1 - e^{-\lambda w}. \end{aligned}$$

The p.d.f. of W is $F'(w) = \lambda e^{-\lambda w}$.

We often let $\lambda = 1/\theta$ (so θ is the average waiting time of successive changes), and say that the r.v. X has an **exponential distribution** if its p.d.f. is

$$\textit{p.d.f.}: \qquad f(x) = \frac{1}{\theta} e^{x/\theta}, \qquad 0 \le x < \infty.$$

The exponential distribution has the following characteristics:

$$\begin{aligned} m.g.f.: &\quad M(t) = E(e^{tX}) = \frac{1}{1-\theta t}, \quad t < \frac{1}{\theta}. \\ \text{mean:} &\quad \mu = M'(0) = \theta. \\ \text{variance:} &\quad \sigma^2 = M''(0) - [M'(0)]^2 = \theta^2. \end{aligned}$$

The curves of p.d.f. and c.d.f. for an exponential distribution ($\theta = 5$) are given in Fig 3.4-2, p.144 (scanned file).

Ex 2, p.143 (scanned file)

Ex 3-4, p.144-145 (scanned file)

Homework
§3.4 1, 3, 11, 13

Attachment: Scanned textbook pages of Section 3-4

3.5 The Gamma and Chi-Square Distributions

In the (approximate) Poisson process with mean λ, the waiting time until the first change has an exponential distribution. Now let W denote the **waiting time until the αth change occurs**. The c.d.f. of W when $w \geq 0$ is

$$\begin{aligned}
\text{c.d.f.:} \qquad F(w) &= P(W \leq w) = 1 - P(W > w) \\
&= 1 - P(\text{fewer than } \alpha \text{ changes occur in [0,w]}) \\
&= 1 - \sum_{k=0}^{\alpha-1} \frac{(\lambda w)^k e^{-\lambda w}}{k!}, \\
\text{p.d.f.:} \qquad f(w) &= F'(w) = \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!} e^{-\lambda w}.
\end{aligned}$$

We have

$$1 = \int_0^\infty f(w)dw = \frac{1}{(\alpha-1)!} \int_0^\infty (\lambda w)^{\alpha-1} e^{-\lambda w} d(\lambda w) = \frac{1}{(\alpha-1)!} \int_0^\infty y^{\alpha-1} e^{-y} dy$$

Define the **gamma function**:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy.$$

It is continuous and $\Gamma(\alpha) = (\alpha-1)!$ when α is a positive integer.

A r.v. X has a **gamma distribution** if its

$$\text{p.d.f.:} \qquad f(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta}, \qquad 0 \leq x < \infty.$$

The waiting time until the αth change occurs in the Poisson process has a Gamma distribution with $\theta = 1/\lambda$. The gamma distribution X has the characteristics

$$\begin{aligned}
\text{m.g.f.:} \qquad & M(t) = \frac{1}{(1-\theta t)^\alpha}, \qquad t < 1/\theta, \\
\text{mean:} \qquad & \mu = \alpha\theta, \\
\text{variance:} \qquad & \sigma^2 = \alpha\theta^2.
\end{aligned}$$

Ex 1-2, p.150-151 (scanned file)

Fig 3.5-1 (p151) shows the p.d.f.'s for several Gamma distributions (scanned file)

Let r be a positive integer. The gamma distribution X with $\theta = 2$ and $\alpha = r/2$ has a **chi-square distribution with r degrees of freedom**, denoted by $X \sim \chi^2(r)$.

$$\begin{aligned}
\text{p.d.f.:} &\quad f(x) = \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} e^{-x/2}, \qquad 0 \le x < \infty, \\
\text{m.g.f.:} &\quad M(t) = (1-2t)^{-r/2}, \qquad t < \frac{1}{2}, \\
\text{mean:} &\quad \mu = \alpha\theta = r, \\
\text{variance:} &\quad \sigma^2 = \alpha\theta^2 = 2r.
\end{aligned}$$

Fig 3.5-2 (p152) shows the p.d.f.'s for some chi-square's.

Table IV in Appendix can be used to find the c.d.f.'s of chi-square distributions (scanned file).

Ex 3-4, p.152-153 (scanned file)

For a chi-square distribution $X \sim \chi^2(r)$, the number $\chi^2_\alpha(r)$ is defined such that

$$P[X \ge \chi^2_\alpha(r)] = \alpha.$$

So $\chi^2_\alpha(r)$ is the $100(1-\alpha)$th percentile of the chi-square distribution with r degrees of freedom.

Ex 5, p.153 (scanned file)

Ex 6, p.154 (scanned file)

Homework

§3.5 1, 7, 9, 15, 17

Attachment: Scanned textbook pages of Section 3-5

3.6 The Normal Distribution

A random variable X has a **normal distribution** if its p.d.f. is defined by

$$p.d.f.: \qquad f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad -\infty < x < \infty,$$

where the parameters $\mu \in (-\infty, \infty)$ and $\sigma > 0$. We say that $X \sim N(\mu, \sigma^2)$.

The r.v. X has characteristics

$$m.g.f.: \qquad M(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right), \qquad -\infty < t < \infty,$$

$$\text{mean:} \qquad E(X) = M'(0) = \mu,$$

$$\text{variance:} \qquad \text{Var}(X) = M''(0) - [M'(0)]^2 = \sigma^2.$$

Ex 1-2, p.158-159 (scanned file)

....

If $Z \sim N(0,1)$, we say that Z has a **standard normal distribution**. Clearly, Z has mean 0, variance 1, m.g.f. $M(t) = e^{t^2/2}$, and

$$c.d.f.: \qquad \Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw.$$

Table Va gives the c.d.f. $P(Z \le z)$ for standard normal distribution $N(0,1)$ and positive z. Table Vb provides $P(Z > z)$. In fact, the p.d.f. curve of $N(0,1)$ is symmetric about the y-axis. So if $z > 0$, then

$$\Phi(-z) = 1 - \Phi(z).$$

Thus Table Va is enough.

Ex 3-4, p.159-160 (scanned file)

In statistical application, we often want to find z_α for $Z \sim N(0,1)$ such that

$$P(Z \ge z_\alpha) = \alpha.$$

That is, z_α is the $100(1-\alpha)$th percentile (or called the upper 100α percent point) of the standard normal distribution (See Fig 3.6-2 in scanned file). By the symmetry of the p.d.f. of $N(0,1)$,

$$P(Z \ge z_{1-\alpha}) = 1 - \alpha = 1 - P(Z \ge z_\alpha) = P(Z \ge -z_\alpha).$$

Thus

$$z_{1-\alpha} = -z_\alpha.$$

Ex 5, p.161 (scanned file)

Thm 3.1. *If* X *is* $N(\mu, \sigma^2)$*, then* $Z := \frac{X-\mu}{\sigma}$ *is* $N(0, 1)$*.*

The proof is skipped. It is easy to verify that $E(X) = 0$ and $\mathrm{Var}(Z) = 1$. By this theorem, any normal distribution can be changed to a standard normal distribution.

Ex 6-7, p.162-163 (scanned file)

. . . .

(optional) A normal distribution is also related to chi-square.

Thm 3.2. *If* $X \sim N(\mu, \sigma^2)$*, then the r.v.* $V := \frac{(X-\mu)^2}{\sigma^2} = Z^2 \sim \chi^2(1)$*.*

. . . .

q-q plot can be used to examine the coherence between empirical data and conjectured model.

Ex 9, p.164-165 (scanned file)

Homework

§3.6 1, 3, 5, 7, 13

Attachment: Scanned textbook pages of Section 3-6