2 Convergence results for sequences of random variables

2.1 Introduction

This chapter will consider some convergence results that can be used for approximating the distribution of random variables under certain situations and/or assessing the behaviour of aggregates of a random sample when the sample size is large.

First we consider the binomial distribution, Bin(n, p), and how this can be approximated by the Poisson distribution when the parameter n is large and the parameter p is small.

We will also present a justification for the use of the arithmetic average of a random sample when one is attempting to estimate the value of the population mean $\mu$ from which the sample was drawn; we do this using a limit theorem known as the Weak Law of Large Numbers.

Much of the theory of statistical estimation is built upon the assumption that the observed random sample is generated by a collection of i.i.d. r.v.’s drawn from the normal distribution; in many cases this assumption may be fine, but in many others it might seem hard to justify. Based upon the central limit theorem, it will be argued that the assumption of approximate normality may hold true for certain aggregated quantities of the random variables even if the random variables themselves are not normal.

2.2 Poisson approximation to the Binomial Distribution

Recall that the p.m.f. of a r.v. $X$ from the Binomial distribution, with parameters $n$ and $p$, i.e. $X \sim \text{Bin}(n, p)$, takes the form

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k \in \mathbb{R} \subset \{0, 1, \ldots, n\}$$

with distribution function $F_X(x) = 0$ for $x < 0$, and

$$F_X(x) = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$$

for $x \geq 0$, where $\lfloor x \rfloor$ is the integer part of $x$.

Tabulated values for $F_X(x)$ are often only available for up to $n = 20$ (c.f. Lindley & Scott). To proceed beyond $n = 20$, we could explicitly compute $F_X(x)$, according to the above formula, however this could be a time consuming and/or a computationally expensive exercise. It may also be the case that the behaviour of $F_X(\cdot)$ is required in the limit as $n \longrightarrow \infty$ as part of a derivation or proof of another result.
It turns out that for:
(i) large $n$,
(ii) small $p$ (i.e. $p \ll 1$),
(iii) $np$ of moderate size (i.e. $0 \leq np \leq 20$)
we can approximate a $\text{Bin}(n,p)$ distribution by a $\text{Poisson}(np)$ distribution (which may be an easier distribution to handle). To explain the theoretical basis for this conclusion, we will first introduce the notion of convergence in distribution.

**Definition 2.2.1 (convergence in distribution)**
Let the sequence $Z_1, Z_2, \ldots,$ and $Z_n$ be a collection of random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.
If $F_{Z_n}(x) \rightarrow F_Z(x)$ as $n \rightarrow \infty$ at each point $x \in \mathbb{R}$ at which $F_Z(x)$ is continuous, then $Z_n \xrightarrow{d} Z$.
We say that $Z_n$ **converges in distribution** to $Z$.

**Definition 2.2.2 (generating functions)**
For a random variable $X$,
(i) the **probability generating function** (p.g.f) is given by
$$G_X(z) = E[z^X]$$
for all values of $z$ for which the R.H.S exists;

(ii) the **moment generating function** (m.g.f.) is given by
$$M_X(t) = E[e^{tX}]$$
for all values of $t$ for which the R.H.S. exists.

In the case in which $X$ is discrete and non-negative, we can invoke the law of the unconscious statistician to express the p.g.f. as $G_X(z) = \sum_{k=0}^{\infty} z^k p_X(k)$. For $X$ continuous, then the m.g.f. can be expressed as $M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) \, dx$. Also, for those values of $z$ and $t$ for which the p.g.f. and m.g.f. exists, there is a one-to-one correspondence between the two transformations due to the following identities:

$$G_X(z) = E[z^X] = E[(e^{\ln z})^X] = E[e^{X\ln z}] = M_X(\ln z)$$
$$M_X(t) = E[e^{tX}] = E[(e^t)^X] = G_X(e^t).$$
These identities can be used to established that, where they exist, there is one-to-one correspondence between distributions and their m.g.f.'s (or equivalently their p.g.f.'s). This means that it is possible to check convergence in distribution of a sequence of random variables by considering their corresponding generating functions.
Proposition 2.2.3 (convergence of distribution≡convergence of m.g.f.)
Suppose the sequence $X_1, X_2, \ldots$, and $X$ are r.v.'s on $(\Omega, \mathcal{F}, \mathbb{P})$. Then (under certain mild conditions)

$$F_{X_n}(x) \rightarrow F_X(x) \text{ at continuity points } x \in \mathbb{R}$$

$$\Leftrightarrow M_{X_n}(t) \rightarrow M_X(t)$$

or

$$G_{X_n}(z) \rightarrow G_X(z).$$

This next proposition provides the basis for our approximation.

Proposition 2.2.4 (a limit theorem relating Bin and Poisson distns)
Suppose $X_n \sim \text{Bin}(n, p_n)$ where $p_n := \frac{\mu}{n}$ for $n \in \mathbb{Z}^+$ such that $n > \mu$. Then

$$F_{X_n}(x) \rightarrow F_X(x)$$

where $X \sim \text{Poisson}(\mu)$.

Proof
The p.g.f. of $X_n$ is given by

$$G_{X_n}(z) = (p_nz + (1 - p_n))^n = \left(\frac{\mu z}{n} + 1 - \frac{\mu}{n}\right)^n = \left(1 + \frac{z - 1}{n}\right)^n$$

which, upon passing to the limit as $n \rightarrow \infty$, is equal to $e^{\mu(z-1)}$. [This follows from the standard result that $(1 + \frac{x}{n})^n \rightarrow e^x$.]

However, $e^{\mu(z-1)}$ is the p.g.f. of a r.v. from the Poisson$(\mu)$ distribution. Hence, by Proposition 2.2.3,

$$F_{X_n}(x) \rightarrow F_X(x)$$

where $X \sim \text{Poisson}(\mu)$. $\square$

From the above proposition, we conclude that if $n$ is large, and (by the way $p_n$ is defined) $p_n \ll 1$, then $F_{X_n}(x)$ is approximately equal to $F_X(x)$ for all $x \in \mathbb{R}$. In other words, for (i) large $n$, (ii) $p$ small, $X \sim \text{Bin}(n, p)$ can be approximated by a random variable from the Poisson$(np)$ distribution.

Example 2.2.5 (application of the Poisson approx. to the Bin distribution)
Suppose $X \sim \text{Bin}(40, 0.05)$. From the statistical package Spotfire S+, we find that

$$\mathbb{P}(X \leq 2) = F_X(2) = 0.6767.$$

> pbinom(2, 40, 0.05)

[1] 0.6767358

3
Since, in our case, \( n = 40 \), and the tables [e.g. Lindley and Scott pp. 4-23] only go up to \( n = 20 \), then the quantity under consideration is not immediately accessible without the use of a statistical package or suitable computer program in which to carry out the calculation. Let us consider using an approximation, thus working with the Poisson distribution instead. To use the approximation, note that \( n = 40 \) is ‘quite large’, \( p = 0.05 \) is ‘quite small’, and \( \mu = n \times p = 40 \times 0.05 = 2 \).

Hence, taking \( Y \sim \text{Poisson}(2) \) as our approximating random variable, then from tables, we find that
\[
P(Y \leq 2) = F_Y(2) = 0.6767.
\]

### 2.3 Weak Law of Large Numbers

In this section, we introduce another intuitively appealing ‘mode of convergence’ concept, which, in conjunction with the Weak Law of Large Numbers (see later), can be used to justify particular statistical estimation procedures.

#### Definition 2.3.1 (convergence in probability)
Let \( Z_1, Z_2, \ldots, \) be an arbitrary sequence of r.v.’s and \( c \) a finite constant. Then we say that \( Z_n \) converges to \( c \) in probability if

\[
\text{for all } \varepsilon > 0, \ P(|Z_n - c| \geq \varepsilon) = P(\{\omega : |Z_n(\omega) - c| \geq \varepsilon\}) \longrightarrow 0
\]
as \( n \rightarrow \infty \).

We write \( Z_n \overset{p}{\longrightarrow} c \).

#### Remarks 2.3.2

(i) If \( Z_n \overset{p}{\longrightarrow} c \), then for large \( n \), the probability of the set of \( \omega \)'s whose \( Z_n \)-value differs from \( c \) by at least \( \varepsilon \) is small.

(ii) Since
\[
P(|Z_n - c| \geq \varepsilon) = 1 - P(|Z_n - c| < \varepsilon) = 1 - P(c - \varepsilon < Z_n < c + \varepsilon) \longrightarrow 0
\]
\[
\Leftrightarrow P(c - \varepsilon < Z_n < c + \varepsilon) \longrightarrow 1,
\]
then the distribution of \( Z_n \) becomes more concentrated around \( c \) as \( n \) increases.

#### Example 2.3.3
Suppose \( Z_n \sim \text{Exp}(n) \) where the parameter \( n \) is a positive integer, i.e. \( f_{Z_n}(x) = ne^{-nx} \) where \( n \in \mathbb{Z}^+ \), for \( x \geq 0 \).

Show that \( Z_n \overset{p}{\longrightarrow} 0 \).

**Solution:**
\[
F_{Z_n}(x) = \int_0^x ne^{-ny}dy = -e^{-ny}|_0^x = 1 - e^{-nx}.
\]

For \( \varepsilon > 0 \),
\[
P(|Z_n - 0| \geq \varepsilon) = P(|Z_n| \geq \varepsilon) = P(Z_n \geq \varepsilon) = 1 - P(Z_n < \varepsilon) = 1 - F_{Z_n}(\varepsilon) = 1 - (1 - e^{-n\varepsilon}) = e^{-n\varepsilon} \longrightarrow 0
\]
as \( n \rightarrow \infty \).
Some Notation
Suppose that $X_1, X_2, \ldots$ is an i.i.d. sequence of r.v’s with mean $E[X_i] = \mu$, and variance $\text{var}(X_i) = \sigma^2$, for each $i$.

For each $n$, we can define

$$S_n = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i.$$ 

Then it is easy to show that $E[S_n] = n\mu$, and $\text{var}(S_n) = n\sigma^2$.

So for $\mu \neq 0$, distribution of $S_n$ drifts away from $\mu$, and becomes more spread out!

On the other hand, $E[S_n/n] = \mu$, $\text{var}(S_n/n) = \sigma^2/n$.

So, the distribution of $S_n/n$ remains ‘centred’ at $\mu$, and becomes more concentrated as $n \to \infty$.

Before proceeding onto the main result of this section, we first introduce the following tool.

**Lemma 2.3.4 (Chebyshev’s inequality)**

For any random variable $Y$ with finite mean and variance, and for $\varepsilon > 0$,

$$P(|Y - E[Y]| \geq \varepsilon) \leq \frac{\text{var}(Y)}{\varepsilon^2}$$

**Proof**

Suppose $Y$ has p.d.f. $f_Y(\cdot)$ and set $A = \{y : |y - E[Y]| \geq \varepsilon\}$.

$$\text{var}(Y) = E[(Y - E[Y])^2] = \int (y - E[Y])^2 f_Y(y)dy$$

$$= \int_A (y - E[Y])^2 f_Y(y)dy + \int_{A^c} (y - E[Y])^2 f_Y(y)dy$$

$$\geq \int_A (y - E[Y])^2 f_Y(y)dy \geq \int_A \varepsilon^2 f_Y(y)dy = \varepsilon^2 \int_A f_Y(y)dy = \varepsilon^2 P(Y \in A)$$

$$= \varepsilon^2 P(|Y - E[Y]| \geq \varepsilon) \Rightarrow P(|Y - E[Y]| \geq \varepsilon) \leq \frac{\text{var}(Y)}{\varepsilon^2}.$$ 

The previous comments and the above lemma suggest that the following result should be true.

**Theorem 2.3.5 (weak law of large numbers)**

Let $X_1, X_2, \ldots, \ldots$ be a sequence of i.i.d. r.v.’s. each with finite mean $\mu$ and finite variance $\sigma^2$. Then

$$\frac{S_n}{n} \overset{p}{\to} \mu.$$ 

**Proof**

By Lemma 2.3.4,

$$P \left( \left| \frac{S_n}{n} - E \left[ \frac{S_n}{n} \right] \right| \geq \varepsilon \right) \leq \frac{\text{var}(S_n/n)}{\varepsilon^2}$$

i.e.

$$P \left( \left| \frac{S_n}{n} - \mu \right| \geq \varepsilon \right) \leq \frac{\sigma^2}{n\varepsilon^2}.$$ 

5
Hence
\[ P\left( \left| \frac{S_n}{n} - \mu \right| \geq \varepsilon \right) \rightarrow 0 \text{ as } n \rightarrow \infty. \]
i.e.
\[ \frac{S_n}{n} \xrightarrow{p} \mu. \]

2.4 The Central Limit Theorem
Suppose \( X_i \sim N(\mu, \sigma^2), i = 1, 2, \ldots, n, \) is a collection of independent r.v.’s. Then we know that
\[ S_n = \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2). \]
But now suppose that the \( X_i \)'s are a collection of i.i.d. r.v.’s drawn from any distribution with finite mean \( \mu \) and finite variance \( \sigma^2 \). What can we say about the distribution of \( S_n = \sum_{i=1}^{n} X_i \) in this case? It turns out that if \( n \) is large, then
\[ \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2) \text{ approximately.} \]
The basis for this approximation comes from the CENTRAL LIMIT THEOREM.

**Theorem 2.4.1 (central limit theorem)**
Let \( X_1, X_2, \ldots \) be an i.i.d. sequence of r.v.’s with finite mean \( \mu \) and finite non-zero variance \( \sigma^2 \). Then
\[ Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma \sqrt{n}} = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} Z \sim N(0, 1) \]
irrespective of the distribution of the \( X_i \)'s.

Before we present the proof, let us now introduce an additional concept that will give us a mechanism for describing the speed at which a quantity converges to zero.

**Definition 2.4.2 (order of a function)**
A function \( h(\cdot) \) is said to be of order \( o(t) \) if
\[ \lim_{t \to 0} \frac{h(t)}{t} = 0. \]

Also recall the following result first introduced in *Statistics: Theory and Practice* relating to the moment generating function for a linear combination of independent random variables. This will be shown to provide a neat short cut in the development of our proof.
Proposition 2.4.3 (m.g.f. of a linear combination of indep. r.v.’s)
If \( X_1, X_2, \ldots, X_n \) are indep. r.v.’s and \( a_1, a_2, \ldots, a_n \) are constants, then
\[
M_{a_1X_1+a_2X_2+\ldots+a_nX_n}(t) = M_{X_1}(a_1t)M_{X_2}(a_2t)\ldots M_{X_n}(a_nt).
\]

Now back to the result we were originally trying to prove.

Proof of Theorem 2.4.1
\[
Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i
\]
where \( Y_i = \frac{X_i - \mu}{\sigma}, i = 1, \ldots, n \), is an i.i.d. sequence of r.v.’s; it follows that \( E[Y_i] = 0 \) and \( \text{var}(Y_i) = 1 \), and consequently \( E[Y_i^2] = \text{var}(Y_i) + 0^2 = 1 \) (check for yourself). Suppose \( Y \) has the same distribution as \( Y_i, i = 1, \ldots, n \). Then the moment generating function can be approximated in the following way:
\[
M_Y(s) = E[e^{sY}] = E \left[ 1 + sY + \frac{s^2Y^2}{2!} + \ldots \right] = 1 + sE[Y] + \frac{s^2}{2}E[Y^2] + o(\frac{s^2}{2})
\]
Hence
\[
M_{Z_n}(t) = E[e^{tZ_n}] = E[e^{\frac{t}{\sqrt{n}} \sum_{i=1}^{n} Y_i}] = M_{\sum_{i=1}^{n} Y_i} \left( \frac{t}{\sqrt{n}} \right) = \left[ M_Y \left( \frac{t}{\sqrt{n}} \right) \right]^n
\]
noting that the fourth equality follows from Proposition 2.4.3 (with \( a_1 = a_2 = \ldots = a_n = 1 \)), and the limit follows from the fact that \( (1 + \frac{x}{n})^n \longrightarrow e^x \) as \( n \longrightarrow \infty \).

But \( e^{t^2/2} \) is the m.g.f. of a r.v. from the \( N(0,1) \) distribution. Hence \( Z_n \overset{d}{\longrightarrow} Z \sim N(0,1) \). \( \square \)

Remarks 2.4.4
Since \( \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \left( \frac{1}{n} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \right) \sim N(0,1) \) approximately for large \( n \), then it now follows (as hinted at earlier) that
\( \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2) \) approximately also, for large \( n \).

Example 2.4.5 (How large does \( n \) need to be for CLT normal approximations?)
Consider a sequence of i.i.d. random variables, \( X_1, X_2, \ldots, \) each drawn from the \( \text{Exp}(\beta) \) distribution. If we set \( \beta = 0.01 \), then, for each \( i, \mu = E[X_i] = 100 \) and \( \sigma^2 = \text{var}(X_i) = 10000 \). We can plot the probability density function of
\[
\frac{\sum_{i=1}^{n} X - n\mu}{\sigma\sqrt{n}}
\]
for various values of \( n \).
It is clear that by the time we get to $n = 300$, the probability density function is fairly close to that of the standard normal, $N(0,1)$. However, the peak of the p.d.f. isn’t quite centred at 0, and perhaps the p.d.f. is a little asymmetrical. However, the CLT is actually a more direct statement about the shape of the cumulative distribution function. So again, let us plot the c.d.f. of

$$\frac{\sum_{i=1}^{n} X - n\mu}{\sigma\sqrt{n}}$$

for various values of $n$ (labelled by the 1’s in the figures) and compare against the c.d.f. of the standard normal (labelled by the 2’s in the figures).
It appears that there is fairly good correspondence for \( n \geq 20 \). As a rule of thumb, \( n = 25 \) or greater can be taken to be a value beyond which the approximation is reasonably robust. For \( n \geq 30 \), one would be in very safe territory.

**Example 2.4.6 (application of CLT)**

Let \( X_1, X_2, \ldots \) be an i.i.d. sequence of r.v.’s from the Uniform(0, 1) distribution. Find the *approximate* probability that \( \sum_{i=1}^{25} X_i \) lies between 10 and 12.

**Solution:**

\( n = 25 \) might just be considered to be sufficiently large for the approximation via the CLT to be valid.

\[
E[X_i] = (0 + 1)/2 = 1/2, \quad \text{var}(X_i) = (1 - 0)^2/12 = 1/12.
\]

Therefore \( E[Y_{25}] = E[\sum_{i=1}^{25} X_i] = 25 \times 0.5 = 12.5 \), and \( \text{var}(Y_{25}) = \text{var}(\sum_{i=1}^{25} X_i) = 25 \times 1/12 = 25/12 \).

Hence

\[
P(10 \leq Y_{25} \leq 12) = P \left( \frac{10 - E[Y_{25}]}{\sqrt{\text{var}(Y_{25})}} \leq \frac{Y_{25} - E[Y_{25}]}{\sqrt{\text{var}(Y_{25})}} \leq \frac{12 - E[Y_{25}]}{\sqrt{\text{var}(Y_{25})}} \right)
\]

\[
= P \left( \frac{10 - 12.5}{\sqrt{25/12}} \leq \frac{Y_{25} - 12.5}{\sqrt{25/12}} \leq \frac{12 - 12.5}{\sqrt{25/12}} \right) \approx P(-1.73205 \leq Z \leq -0.34641)
\]

where \( Z \sim N(0, 1) \).

Let \( \Phi(z) = P(Z \leq z) \). By the symmetry of the \( N(0, 1) \) distribution, \( \Phi(z) = 1 - \Phi(-z) \), and so

\[
P(10 \leq Y_{25} \leq 12) = \Phi(-0.34641) - \Phi(-1.73205) = (1 - \Phi(0.34641)) - (1 - \Phi(1.73205))
\]

\[
= \Phi(1.73205) - \Phi(0.34641) \approx 0.9582 - 0.6368 = 0.3214.
\]

[see Table 4 pp.34-35 Lindley and Scott.]