5 Convergence Results and the Normal Distribution

5.1 Convergence

Definition 1 (convergence in distribution)**.** We say that a sequence $(X_n)_{n\geq 1}$ of random variables with distribution functions $(F_n(\cdot))_{n\geq 1}$ converges in distribution to a random variable X with distribution function $F(\cdot)$ if $F_n(x) \to F(x)$ at all points x where $F(x)$ is continuous (i.e. doesn't jump). In that case we write $X_n \xrightarrow{D} X$ or $F_n \xrightarrow{D} F$ and we also say that the corresponding sequence of distributions functions F_n converges in distribution to F.

In formulas:

$$X_n \xrightarrow{D} X \quad \Leftrightarrow \quad \lim_{n\to\infty} F_n(x) = F(x) \text{ at all points of continuity of } F(\cdot).$$

Remark 1. The reason that we exclude the points where F jumps is that, according to our conventions, probability distribution functions have to be *right continuous*, and one can find examples of sequences of right continuous functions whose *pointwise* limit is not right continuous anymore: for example, if

$$F_n(x) = \begin{cases} 1, & x \geq \frac{1}{n} \\ 0, & x < \frac{1}{n}, \end{cases}$$

then, *pointwise*, F_n converges to the function

$$\widetilde{F}(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0, \end{cases}$$

which is not a distribution function (since it is not right continuous at $x = 0$, it is in fact left continuous there). However, if we define

$$F(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0, \end{cases}$$

then $F_n \to F$ in distribution, and all is well, since we simply exclude the jump-point $x = 0$ from consideration.

Theorem 1 (continuity)**.** Suppose that $F_1, F_2, \ldots$ is a sequence of distribution functions with corresponding characteristic functions $\Psi_1, \Psi_2, \ldots$.

1. If $F_n \xrightarrow{D} F$ for some distribution F with characteristic function Ψ then

$$\Psi_n(\theta) \longrightarrow \Psi(\theta)$$

   for all θ.

2. Conversely, if $\Psi(\theta) = \lim_{n\to\infty} \Psi_n(\theta)$ exists for all (real) θ and is continuous at $\theta = 0$, then Ψ is the characteristic function of some distribution function F, and $F_n \xrightarrow{D} F$.

Definition 2 (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

$$\text{for all } \varepsilon > 0,\ \mathbb{P}(|Z_n - c| \geq \varepsilon) = \mathbb{P}(\{\omega : |Z_n(\omega) - c| \geq \varepsilon\}) \longrightarrow 0$$

as $n \to \infty$.
We write $Z_n \xrightarrow{p} c$.

Remark 2.
(i) If $Z_n \xrightarrow{p} c$, then for large n, the probability of the set of $\omega's$ whose Z_n-value differs from c by at least ε is small.
(ii) Since

$$\mathbb{P}(|Z_n - c| \geq \varepsilon) = 1 - \mathbb{P}(|Z_n - c| < \varepsilon) = 1 - \mathbb{P}(c - \varepsilon < Z_n < c + \varepsilon) \longrightarrow 0$$

$$\Leftrightarrow \mathbb{P}(c - \varepsilon < Z_n < c + \varepsilon) \longrightarrow 1,$$

then the distribution of Z_n becomes more concentrated around c as n increases.

Example 1.
Suppose $Z_n \sim Exp(n)$ 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 $Z_n \xrightarrow{p} 0$.
Solution:

$$F_{Z_n}(x) = \int_0^x ne^{-ny}dy = -e^{-ny}|_0^x = 1 - e^{-nx}.$$

For $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}(|Z_n - 0| \geq \varepsilon) = \mathbb{P}(|Z_n| \geq \varepsilon) = \mathbb{P}(Z_n \geq \varepsilon) = 1 - \mathbb{P}(Z_n < \varepsilon) \\ = 1 - F_{Z_n}(\varepsilon) = 1 - (1 - e^{-n\varepsilon}) = e^{-n\varepsilon} \longrightarrow 0\end{aligned}$$

as $n \longrightarrow \infty$.

5.2 Limit Theorems

We next present two remarkable limit theorems for sequences of independent identically distributed random variables, the *weak law of large numbers* and the *central Limit Theorem* which have important applications in probability and statistics. Indeed, in statistics they justify for example the usual procedure of taking the empirical mean $\frac{1}{n}\sum_{j=1}^{n} x_j$ of a large number of successive realizations $x_1, x_2, \ldots$ of a random variable X to estimate its expectation $\mathbb{E}[X]$.

But, first, here is an ancillary result.

Lemma 1 (Chebyshev's inequality)**.**
For any random variable Y with finite mean and variance, and for $\varepsilon > 0$,

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

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

$$\begin{aligned}\text{var}(Y) = E[(Y - E[Y])^2] = \textstyle\int (y - E[Y])^2 f_Y(y)dy \\ = \textstyle\int_A (y - E[Y])^2 f_Y(y)dy + \int_{A^c}(y - E[Y])^2 f_Y(y)dy \\ \geq \textstyle\int_A (y - E[Y])^2 f_Y(y)dy \geq \int_A \varepsilon^2 f_Y(y)dy = \varepsilon^2 \int_A f_Y(y)dy = \varepsilon^2\mathbb{P}(Y \in A) \\ = \varepsilon^2\mathbb{P}(|Y - E[Y]| \geq \varepsilon) \Rightarrow \mathbb{P}(|Y - E[Y]| \geq \varepsilon) \leq \text{var}(Y)/\varepsilon^2.\end{aligned}$$

$\square$

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

Theorem 2 (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 μ and finite variance σ^2 and set $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{S_n}{n} \xrightarrow{p} \mu.$$

Proof. By Lemma 1,

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

But

$$E\left[\frac{S_n}{n}\right] = \mu \quad \text{and} \quad \operatorname{var}\left(\frac{S_n}{n}\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

therefore

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

Hence

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \longrightarrow 0 \text{ as } n \longrightarrow \infty,$$

i.e.

$$\frac{S_n}{n} \xrightarrow{p} \mu. \qquad \square$$

Definition 3 (order of a function)**.**
A function $h(\cdot)$ is said to be of order $o(t)$ as $t \to a$ if

$$\lim_{t \to a} \frac{h(t)}{t} = 0.$$

Write as $h(t) \sim o(t)$ or $h(t) = o(t)$ as $t \to a$.

Theorem 3 (central limit theorem)**.** Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite mean μ and finite non-zero variance σ^2, and let $S_n = X_1 + \cdots + X_n$. Then

$$\frac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{D} N(0,1) \qquad \text{as } n \longrightarrow \infty.$$

Proof. Let $Y_i = (X_i - \mu)/\sigma$, and let $\Psi_Y(\cdot)$ be the characteristic function of Y_i. Then $\mathbb{E}[Y] = 0$ and $\mathbb{E}[Y^2] = 1$, so we have

$$\Psi_Y(\theta) = 1 - \frac{1}{2}\theta^2 + o(\theta^2) \quad \text{as} \quad \theta \to 0.$$

Also, the characteristic function $\Psi_n(\cdot)$ of

$$U_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}}\Sigma_{i=1}^n Y_i$$

satisfies

$$\Psi_n(\theta) = \left[\Psi_Y(\theta/\sqrt{n})\right]^n = \left[1 - \frac{\theta^2}{2n} + o(\theta^2/n)\right]^n \to e^{-\frac{1}{2}\theta^2} \qquad \text{as } n \to \infty,$$

which is the characteristic function of a $N(0,1)$-random variable. □

5.3 Multivariate normal distribution

5.3.1 General Description

The normal distribution is of fundamental importance in univariate sampling theory.

To recall, suppose that $X \sim N(\mu, \sigma^2)$. Then X has a p.d.f. given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

for $x \in \mathbb{R}$.

A generalization of this distribution for a $p \times 1$ random vector $\mathbf{X}$ is:

Definition 4 (MVN distribution)**.** A $p \times 1$ random vector $\mathbf{X}$ is said to have a *multivariate normal* (MVN) distribution if its joint p.d.f. is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{1}$$

for $\mathbf{x} \in \mathbb{R}^p$, where $\mathbf{\Sigma}$ is a $p \times p$, symmetric, positive-definite, matrix[1] and $\boldsymbol{\mu} \in \mathbb{R}^p$. We write $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$.

[1] A symmetric $p \times p$ real matrix M is said to be positive definite if $z'Mz$ is positive for all non-zero column vectors z of p real numbers where z' denotes the transpose of z.

Remark

Suppose that $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $p = 1$. Then, in this case, $\boldsymbol{\Sigma} = \sigma_{11} = \sigma_1^2$, and so

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1}(x_1) = \frac{1}{(2\pi)^{1/2}|\sigma_1^2|^{1/2}} \exp\left\{-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2\right\}$$

for $x_1 \in \mathbb{R}$ i.e. $X_1 \sim N(\mu_1, \sigma_1^2)$. Thus, the MVN *really* is a generalization of the univariate normal.

5.3.2 Bivariate normal distribution

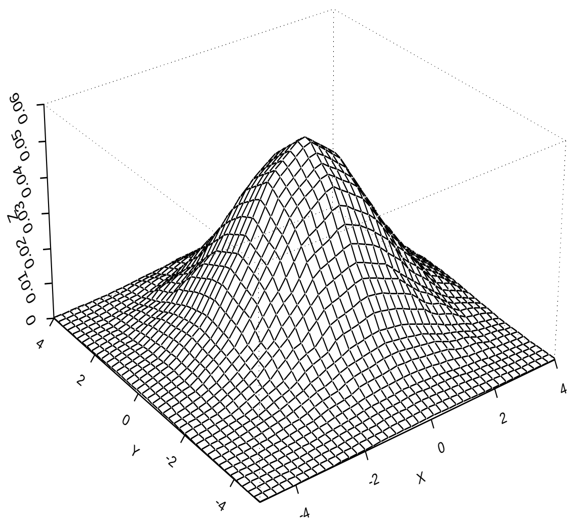

Figure 1: Bivariate Normal density

The Bivariate normal distribution is just the MVN for $p = 2$. So, here, $\boldsymbol{\mu} = (\mu_1, \mu_2)'$, and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

where $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$ (the correlation coefficient between X_1 and X_2).

It can be shown that

$$f_{(X_1,X_2)}(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\} \tag{2}$$

for $(x_1,x_2) \in \mathbb{R}^2$ and provided that $|\rho| < 1$.

Remarks

(i) The p.d.f. of (2) is specified by 5 parameters, $\mu_1, \mu_2, \sigma_1, \sigma_2$ and ρ.

(ii) x_1, x_2 only appear in the argument of the $\exp(\cdot)$ function. So the contour lines of $f_{(X_1,X_2)}(\cdot,\cdot)$ are given by

$$\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) = k > 0.$$

These are ellipse equations.

If $\rho < 0$, then the major axis has negative slope, and for $\rho > 0$, a positive slope; e.g. for $\mathbf{\Sigma} = \begin{bmatrix} 1.5 & -1 \\ -1 & 2.5 \end{bmatrix}$, $\rho = -1/\sqrt{1.5 \times 2.5} < 0$, and $\mathbf{\Sigma} = \begin{bmatrix} 1.5 & 1 \\ 1 & 2.5 \end{bmatrix}$, $\rho = 1/\sqrt{1.5 \times 2.5} > 0$.

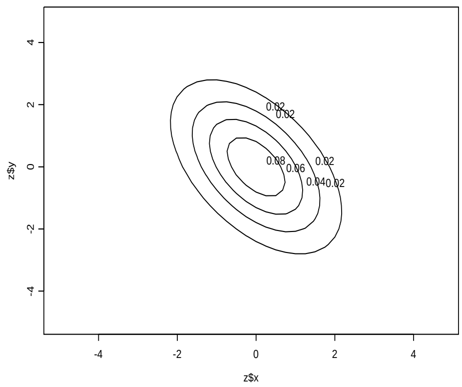

Figure 2: Major axis: negative slope

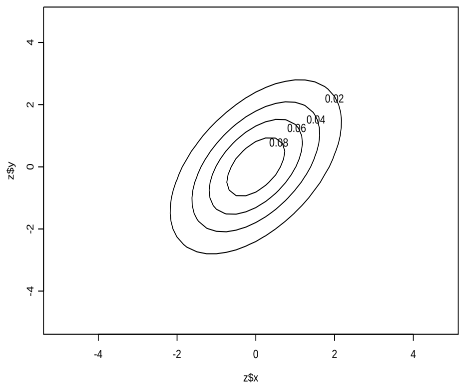

Figure 3: Major axis: positive slope

(iii) $\mathbf{\Sigma}$ is positive definite if, and only if, $|\rho| < 1$. If $\rho = 1$, then rows (or columns) of $\mathbf{\Sigma}$ are no longer linearly independent.
(iv) For this distribution, it is the case that $\rho = 0$ implies that X_1 and X_2 are independent, since

$$
\begin{aligned}
&f_{(X_1,X_2)}(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-0}} \\
&\times \exp\left\{-\frac{1}{2(1-0)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 0\right]\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right\} \times \frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left\{-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\} \\
&= f_{X_1}(x_1) f_{X_2}(x_2)
\end{aligned}
$$

with $R_{X_1} = \mathbb{R}$ and $R_{X_2} = \mathbb{R}$, thus $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$.

5.4 Distributions Arising from the Normal Distribution

Statisticians are frequently faced with a collection of experiments. They might be prepared to make a general assumption about the unknown distribution of these variables without specifying the numerical value of certain parameters. Frequently they might suppose that the sample $X_1, \cdots, X_n$ is a collection of $N(\mu, \sigma^2)$ variables for some fixed but unknown values of μ and σ^2; this assumption is often a very close approximation to reality. They might then proceed to estimate the values of μ and σ^2 by using functions of $X_1, \cdots, X_n$. They will commonly use the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

as a guess at the value of μ. And the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

as a guess of the value of σ^2. We can check that

$$\mathbb{E}[\overline{X}] = \mu \qquad \text{and} \qquad \mathbb{E}[S^2] = \sigma^2.$$

If $X_1, X_2, \cdots$ are independent $N(\mu, \sigma^2)$ variables then

$$\overline{X} \sim N(\mu, \sigma^2/n),$$

as follows by computing the variance of $\overline{X}$, which turns out to be σ^2/n.

The pair $\overline{X}$ and S^2 are related as follows:

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}(X_i-\overline{X}+\overline{X}-\mu)^2 \\ &= \sum_{i=1}^{n}(X_i-\overline{X})^2 + 2(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\overline{X}) + n(\overline{X}-\mu)^2.\end{aligned}$$

Since $\sum_i (X_i - \overline{X}) = n\overline{X} - n\overline{X} = 0$, it follows that

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 + \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2.$$

and therefore

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2. \tag{2}$$

We will now show that the left hand side has a χ^2_n-distribution. Since $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$,

$$Z_i := \frac{X_i - \mu}{\sigma} \sim N(0,1)$$

We know for example that $Z_1^2 \sim \chi^2_1$. Similarly, we can show that the sum of squares of n independent $N(0,1)$ random variables Z_i is χ^2 with n degrees of freedom:

$$Z_1^2 + \cdots + Z_n^2 \sim \chi^2_n. \tag{3}$$

Moreover

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{and so} \qquad \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2_1.$$

It can also be shown that the random variables S^2 (sample variance) and $\overline{X}$ (sample mean) are independent when, as assumed in this section, the random sample $X_1, \cdots, X_n$ is a collection of independent $N(\mu, \sigma^2)$ variables for some fixed but unknown values of μ and σ^2.

Using (2), it can be shown that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

5.4.1 Student t Distribution

If $Z \sim N(0,1)$ and $Y \sim \chi^2_r$ *independently* of each other then

$$T = \frac{Z}{(Y/r)^{1/2}}$$

is defined to have a t distribution with r degrees of freedom. This is denoted by

$$T = \frac{Z}{(Y/r)^{1/2}} \sim t_r.$$

Its density can be computed to be equal to

$$f_T(t) = \frac{\Gamma\left(\frac{1}{2}(r+1)\right)}{\sqrt{\pi r}\,\Gamma\left(\frac{1}{2}r\right)}\left(1+\frac{t^2}{r}\right)^{-\frac{1}{2}(r+1)}, \qquad -\infty < t < \infty.$$

Consider the sampling distributions

$$Y = \frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1} \quad \text{and} \quad Z = \frac{\sqrt{n}}{\sigma}(\overline{X}-\mu) \sim N(0,1).$$

Note that both Y and Z have distributions that do not depend on σ. We have shown that $\overline{X}$ and S^2 are independent so that the random variable $T = \frac{Z}{\sqrt{Y/(n-1)}}$ has the t distribution with $n-1$ degrees of freedom. That is

$$T \;=\; \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}.$$

5.4.2 F distribution

Another important distribution is the F distribution. Let U and V be independent variables with χ^2_r and χ^2_s distributions respectively. Then

$$F = \frac{U/r}{V/s}$$

is said to have the F distribution with r and s degrees of freedom; written $F_{r,s}$. Its density function is given by

$$f(x) = \frac{r\Gamma\left(\frac{1}{2}(r+s)\right)}{s\Gamma\left(\frac{1}{2}r\right)\Gamma\left(\frac{1}{2}s\right)} \frac{(rx/s)^{\frac{1}{2}r-1}}{\left[1+(rx/s)^{\frac{1}{2}(r+s)}\right]}, \qquad \text{for } x > 0.$$

Note that if $T \sim t_r$, then $T^2 \sim F_{1,r}$. For example

$$\frac{(\overline{X}-\mu)^2}{S^2/n} \sim F_{1,n-1}.$$