Sampling distribution

- When we draw a random sample **typically** the way the units in the sample are distributed is very close to the way elements are distributed in the population. So, sample stastics are close to population parameters.
- With low probability we may get a sample that deviates significantly from the population. Then the sample statistics will also deviate significantly from the population parameters.
- The **sampling distribution of a statistic** is the distribution of the statistic when samples of the same size N are drawn i.i.d. with replacement. Imagine drawing with replacement and calculating the statistic repeatedly, say n times, from the population, as $n \to \infty$.
- So the sampling distribution is a theoretical construct.
- For example when statistic is the mean. The sampling distribution is approx. normal irrespective of the population distribution, with mean equal to the population mean μ and standard deviation (often called standard error or standard error of the mean) $\sigma_{se} = \frac{\sigma}{\sqrt{N}}$. This is a consequence of the central limit theorem.

Notation

- Population values: greek letters. Ex. μ (mean), σ (std-dev), σ^2 (variance), ρ (Pearson correlation coefficient).
- Estimates for a distribution: greek or latin letter (corresponding to the parameter) with a cap. Ex. $\hat{\mu}$, $\hat{\sigma}$, $\hat{p}$.
- Values for a sample: latin letters, Ex random variable name with bar, $\bar{X}$, for mean; s for std. dev. etc.
- Sampling distribution values will have subscript sd Ex. $\bar{X}_{sd}$, s_{sd}.
- N (capital N) - size of a single sample.
- n (small n) - no. of samples.

Convergence

- Convergence of a sequence x_n to a limit x is based the how the distance between x_n and x changes.
- If X_n is a sequence of RVs with distribution $F_n(x)$, $F_n : \mathbb{R} \to [0, 1]$ (cdf) then if n is fixed $F_n(x)$ gives a real number associated with x. Similarly, if x is fixed then $F_n(x)$ is a sequence of real numbers x_n and we can ask whether it converges to some limit $F(x)$? And whether this is true for almost all values of x.
- If we are concerned with RVs one measure of distance between two RVs can be the distance between their distributions.
- This gives us the concept of **convergence in distribution**.

Convergence in distribution

Definition 2 (Convergence in distribution or weak convergence)

A sequence $X_1, X_2, \ldots$ of real valued RVs converges in distribtion or weakly (written $\xrightarrow{d}$) to RV X if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for every $x \in \mathbb{R}$ at which F is continuous. F_n and F are cumulative distribution functions of X_n and X respectively. $\triangleleft$

Intuitively, we expect later values in the sequence to be better modelled by the distribution F.

Examples of convergence in distribution

Example 1

Let $\{X_i\}$ *be a sequence of random variables drawn i.i.d from* $U(-1,1)$ *then the sequence* Z_n, $Z_n = \frac{\sum_{i=1}^{n} X_i}{\sqrt{n}}$ *(normalized sums)* $Z_n \xrightarrow{d} N(0, \frac{1}{3})$.

Example 2

If sequence X_n *is the fraction of heads after tossing an unbiased coin* n *times. Then* X_1 *has a Bernoulli distribution,* $X_n, n > 1$ *have binomial distributions. As* $n \to \infty$ *the distribution will be increasingly closer to a normal distribution. The sequence* Z_n, $Z_n = \sqrt{n}(\frac{X_n - \mu}{\sigma})$ *will converge to SND. That is:* $Z_n \xrightarrow{d} N(0,1)$.

Example of convergence in distribution contd. I

Example 3

Let X_n be distributed iid as:

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

Define $Y_n = n(1 - \max_{1 \leq i \leq n} X_i)$

Its distribution function is:

$F_n(y) = P(Y_n \leq y) = P(n(1 - \max_{1 \leq i \leq n} X_i) \leq y)$

Example contd.

$$
\begin{aligned}
P(n(1 - \max_{1 \le i \le n} X_i) \le y) &= P(\max_{1 \le i \le n} X_i \ge 1 - \frac{y}{n}) \\
&= 1 - P(\max_{1 \le i \le n} X_i < 1 - \frac{y}{n}) \\
&= 1 - P(X_1 < 1 - \frac{y}{n}, \ldots, X_n < 1 - \frac{y}{n}) \\
&= 1 - P(X_1 < 1 - \frac{y}{n}) \times \ldots \times P(X_n < 1 - \frac{y}{n}) \\
&= 1 - P(X_1 \le 1 - \frac{y}{n}) \times \ldots \times P(X_n \le 1 - \frac{y}{n}) \\
&= 1 - F_{Y_1}(1 - \frac{y}{n}) \times \ldots \times F_{Y_n}(1 - \frac{y}{n}) \\
&= 1 - (F_{Y_n}(1 - \frac{y}{n}))^n
\end{aligned}
$$

Example contd.

So,

$$F_{Y_n} = \begin{cases} 0 & y < 0 \\ 1 - (F_{X_n}(1 - \frac{y}{n}))^n & 0 \leq y < n \\ 1 & y \geq n \end{cases}$$

Now, $\lim_{n \to \infty} (1 - \frac{y}{n})^n = e^{-y}$.

$$\lim_{n \to \infty} F_{Y_n}(y) = F_Y(y) = \begin{cases} 0 & y < 0 \\ 1 - e^{-y} & y \geq 0 \end{cases}$$

That is an exponential distribution.

Central limit theorem

Theorem 3 (Central Limit Theorem (CLT))

Let $\{X_1, X_2 \ldots\}$ *be a sequence of iid RVs with* $E[X_i] = \mu$ *and* $Var[X_i] = \sigma^2 < \infty$. *Let* $S_N = \frac{X_1 + \cdots + X_N}{N}$ *be the sample mean. Then as* $N \to \infty$ *the RVs* $\sqrt{N}(S_N - \mu)$ *converge in distribution to the normal distribution* $N(0, \sigma^2)$. *That is:*

$$\sqrt{N}(S_N - \mu) \xrightarrow{d} N(0, \sigma^2)$$

or equivalently

$$S_N \xrightarrow{d} N(\mu, \frac{\sigma^2}{N})$$

CLT for us - 1

Let $X_1, \ldots, X_N$ be a random sample from an infinite or sufficiently large population with any distribution with mean μ and variance $\sigma^2 < \infty$. If N is *sufficiently large* then:

- the sample mean $\bar{X}$ follows an approximate normal distribution
- with $E[\bar{X}] = \hat{\mu} = \mu$
- and variance $\sigma^2(\bar{X}) = \hat{\sigma^2} = \frac{\sigma^2}{N}$

Equivalently, $\bar{X} \xrightarrow{d} N(\mu, \frac{\sigma^2}{N})$ as $n \to \infty$ or $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{N}}} \xrightarrow{d} N(0,1)$ as $N \to \infty$

CLT for us - 2

What is sufficiently large?

- If the population distribution is symmetric, unimodal or continuous then a sample size as small as $N = 4$ or $N = 5$ is enough.
- If the distribution is skewed then $N \geq 30$ can be adequate.
- If the distribution is extremely skewed then $N \geq 50$ or even $N \geq 100$ may be needed.