4 Sampling

4.1. Sample Mean and Central Limit Theorem

Suppose we take a series of measurements from some population (e.g., height, duration, etc.) Suppose the quantity we are measuring is distributed with mean μ and variance σ^2.

The sequence of measurements can be modeled as a sequence of RVs $X_1, X_2, \dots, X_n$ that are i.i.d.

The n-th sample mean is the RV

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

We know that

$$E[\overline{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n} E[X_i] = \frac{1}{n}\cdot n\cdot \mu = \mu$$

$$Var(\overline{X}_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var(X_i) = \frac{1}{n^2}\cdot n\cdot \sigma^2 = \frac{1}{n}\sigma^2$$

That is,

- the mean stays the same
- the standard deviation approaches 0

This is the reason behind the weak law of large numbers:

$$P[\,|\overline{X}_n - \mu| > \varepsilon\,] \longrightarrow 0 \qquad (n \to \infty)$$

for all possible bounds $\varepsilon > 0$.

What is the Shape of the Distribution of $\overline{X}_n$?

Problem $\overline{X}_n$ is squeezed by the division by n.

Consider instead

$$Y_i := \frac{X_i - \mu}{\sigma}.$$

Then $E[Y_i] = 0$, $Var(Y_i) = Var\left(\frac{X_i - \mu}{\sigma}\right) = \frac{1}{\sigma^2} Var(X_i) = 1$.

The X_i are i.i.d, so also the Y_i are i.i.d.

Let $$U_n := \frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}} = \sqrt{n} \cdot \overline{Y}_n = \sqrt{n}\, \frac{\overline{X}_n - \mu}{\sigma}$$

Then $E[U_n] = \sqrt{n} \cdot E[\overline{Y}_n] = \sqrt{n} \cdot 0 = 0$

$Var(U_n) = Var(\sqrt{n}\, \overline{Y}_n) = n\, Var(\overline{Y}_n) = n \cdot \frac{1}{n} \cdot 1 = 1$

The Central Limit Theorem (CLT)

The CLT says that the distributions of the U_n, (i.e., the cdfs) converge towards the cdf of the standard normal.

Theorem (Lindeberg-Lévy) [Central Limit Theorem]

Let X_i be i.i.d. RVs with mean μ and variance σ^2 and let

- $U_n := \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$
- F_n be the cdf of U_n (i.e., $F_n(x) = P[U_n \leq x]$)
- Φ be the cdf of $N(0,1)$.

Then

$$\lim_{n \to \infty} F_n(x) = \Phi(x) \quad \text{f.a. } x \in \mathbb{R}$$

Convergence in Distribution

This kind of convergence is called "convergence in distribution", which is the weakest kind of convergence among RVs.

For instance, the Weak Law of Large Numbers says that $\overline{X}_n \longrightarrow \mu$ "in probability", which implies convergence in distribution.

The CLT says, $F_n(x) \longrightarrow \Phi(x)$, but this may be fast for some x and slow for others.

In practice, convergence is faster for x close to 0, that is, close to the mean, and slow if $|x|$ is large, i.e., far away from the mean.

Interpretation and Application of the CLT

Let X_i be i.i.d. RVs with mean μ and variance σ^2.

Let $S_n := \sum_{i=1}^{n} X_i$ be the sum of the X_i.

The CLT says that for large n the normalized sum

$$\frac{1}{\sqrt{n}\sigma}(S_n - n\mu)$$

has approximately a standard normal distribution.

From that one can conclude that

$$S_n \sim \mathcal{N}(n\mu, n\sigma^2) \text{ approximately,}$$

where the approximation is best around the mean $n\mu$.

Probabilities of the values of S_n can then be approximated by probabilities of a normally distributed RV.

Example 64: An insurance company has 25,000 policy holders.

Considering the yearly claim of a policy holder as a RV, the company has observed that

- the mean of the claims is $\mu = € 320$
- the standard deviation is $\sigma = € 540$

What is the probability that the total yearly claim is $> € 8.3$ Mio ?

Example 64: An insurance company has 25,000 policy holders. Considering the yearly claim of a policy holder as a RV, the company has observed that

- the mean of the claims is $\mu = €\, 320$
- the standard deviation is $\sigma = €\, 540$

What is the probability that the total yearly claim is $> €\, 8.3$ Mio?

Let C_i be the yearly claim of policy holder i, and

$S_n = \sum_{i=1}^{n} C_i$ be the yearly sum of claims, $n = 25{,}000$.

$\overline{C}_n = \frac{1}{n} S_n$ be the average of the claims.

We want to know $P[S_n > s]$, where $s = 8.3$ Mio.

From the CLT, we conclude that

$$S_n \sim N(n \cdot \mu, n\sigma^2) \text{ approx.}$$

Hence

$$P[S_n > s] = P\left[\frac{S_n - n\mu}{\sqrt{n}\,\sigma} > \frac{s - n\mu}{\sqrt{n}\,\sigma}\right]$$

$$\approx P\left[Z > \frac{s - n\mu}{\sqrt{n}\,\sigma}\right] = 1 - \Phi\left(\frac{s - n\mu}{\sqrt{n}\,\sigma}\right)$$

Now:

$$n\mu = 25{,}000 \times 320$$
$$= 8 \times 10^6$$

$$\sqrt{n}\,\sigma = \sqrt{25{,}000} \times 540$$
$$= \sqrt{2.5} \times 5.4 \times 10^2 \times 10^2$$

$$s - n\mu = 8.3 \times 10^6 - 8 \times 10^6$$
$$= 3 \times 10^5$$

$$\frac{s - n\mu}{\sqrt{n}\,\sigma} = \frac{3}{\sqrt{2.5} \times 5.4} \frac{10^5}{10^4}$$
$$= 0.351 \times 10 = 3.51$$

Thus $P[S_n > s] = 1 - \Phi(3.51) = 1 - 0.9998 = 0.0002$

Normal and Binomial Distribution

Corollary: Let X_i be independent Bernoulli(p) RVs. Then

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{n} \cdot \sqrt{p \cdot (1-p)}} \longrightarrow N(0,1)$$

in distribution.

Rules of Thumb: A Binomial(n,p) distribution is close to

- $N(np, np(1-p))$ if both $np > 5$, and $n(1-p) > 5$
- Poisson(np) if $np < 5$ or $n(1-p) < 5$, and $n > 20$

Example 65: An airplane fits 150 passengers.

On a busy route, only 30% of the people that buy a ticket take the plane.
If the airline sells 450 tickets per flight, what is the probability that the plane is overbooked?

Example 65: An airplane fits 150 passengers.

On a busy route, only 30 % of the people that buy a ticket take the plane.

If the airline sells 450 tickets per flight, what is the probability that the plane is overbooked?

The number of passenger P taking the plane is a binomial RV with mean $n \cdot p$ and variance $n \cdot p(1-p)$ where

$$n = 450 \quad , \quad p = 0.3 \; .$$

Let $s = 150$ be the number of seats available.

The plane is overbooked if

$$P > 150 \, .$$

We can approximate $\mathcal{P}$ by a RV $\mathfrak{X} \sim \mathcal{N}(np, np(1-p))$. Then

$$P[\mathcal{P} > S] = P[\mathfrak{X} > S + 0.5]$$

adjustment when translating a discrete *) into a continuous problem

$$= P\left[\frac{\mathfrak{X} - np}{\sqrt{n}\sqrt{p(1-p)}} > \frac{S+0.5 - np}{\sqrt{n}\sqrt{p(1-p)}} \right] = 1 - \Phi\left(\frac{S+0.5-np}{\sqrt{n}\sqrt{p(1-p)}} \right)$$

$$= 1 - \Phi(1.59) = 1 - 0.944 = 0.056 = 5.6\%$$

*) called continuity correction