1. Random sample

The random variables X_1, X_2, ..., X_N are a random sample of size n if

- $X_i \cap X_j = \emptyset, \forall i \neq j$.
- Every X_i has same probability distribution.

2. Statistic

- A statistic is any function of the observations in a random sample.
- Statistic is a random variable.
- Examples
  - Sample mean $\overline{X}$
  - Sample variance S^2
  - ...

3. Sampling distribution

The probability of a statistic is called a sampling distribution.

4. Point estimation

- A point estimate of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\hat{\Theta}$.
- The statistic $\hat{\Theta}$ is called the point estimator.
- Some reasonable point estimate
  - $\hat{u} = \overline{x}$
  - $\widehat{\sigma^2} = s^2$
  - $\hat{p} = \frac{x}{n}$
  - $\widehat{\mu_1} - \widehat{\mu_2} = \overline{x_1} - \overline{x_2}$
  - …

Example: There are two ponds containing lots of fish, a random sample of 20 fish were selected from each pond and record their weight. The results are as follows:

$$S_1 : \begin{pmatrix} 1.2 & 3.0 & 2.3 & 1.0 \\ 1.9 & 2.1 & 1.4 & 2.2 \\ 0.7 & 1.3 & 0.5 & 0.8 \\ 2.3 & 3.3 & 4.1 & 3.5 \\ 2.7 & 1.3 & 3.0 & 1.4 \end{pmatrix}$$

$$S_2 : \begin{pmatrix} 1.0 & 2.3 & 1.3 & 1.5 \\ 0.3 & 1.6 & 2.3 & 2.6 \\ 3.3 & 4.2 & 0.8 & 2.8 \\ 3.7 & 0.5 & 4.1 & 3.3 \\ 2.1 & 3.6 & 1.8 & 2.1 \end{pmatrix}$$

1) Estimated average weight, variance, standard deviation and standard rate of fish (weight> 2 kg) in each pond.
2) Compare the weight average, variance, standard deviation of the number of fish in two ponds and the standard rate of fish in two ponds.

	$\overline{x}$	s^2	s	$P(X > 2)$
S_1	2	0.99	0.995	0.5
S_2	2.26	1.3224	1.15	0.6

5. Central Limit Theorem

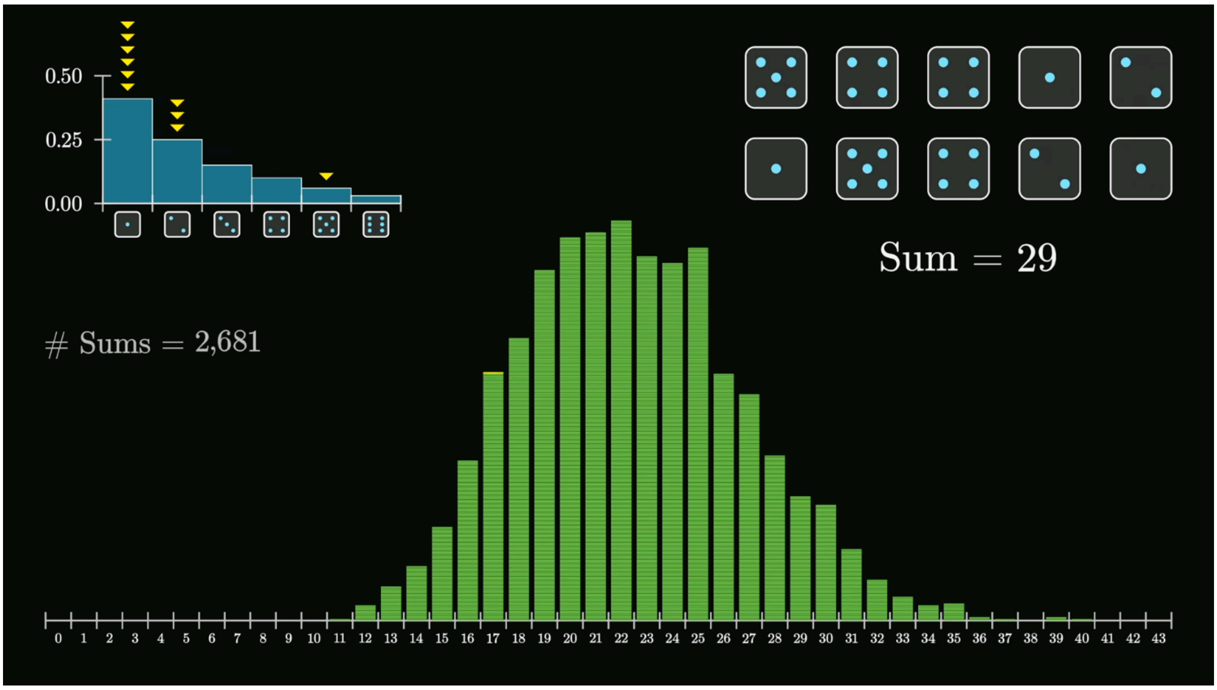

Figure 1: Probability distribution of $X_1 + \ldots + X_n$

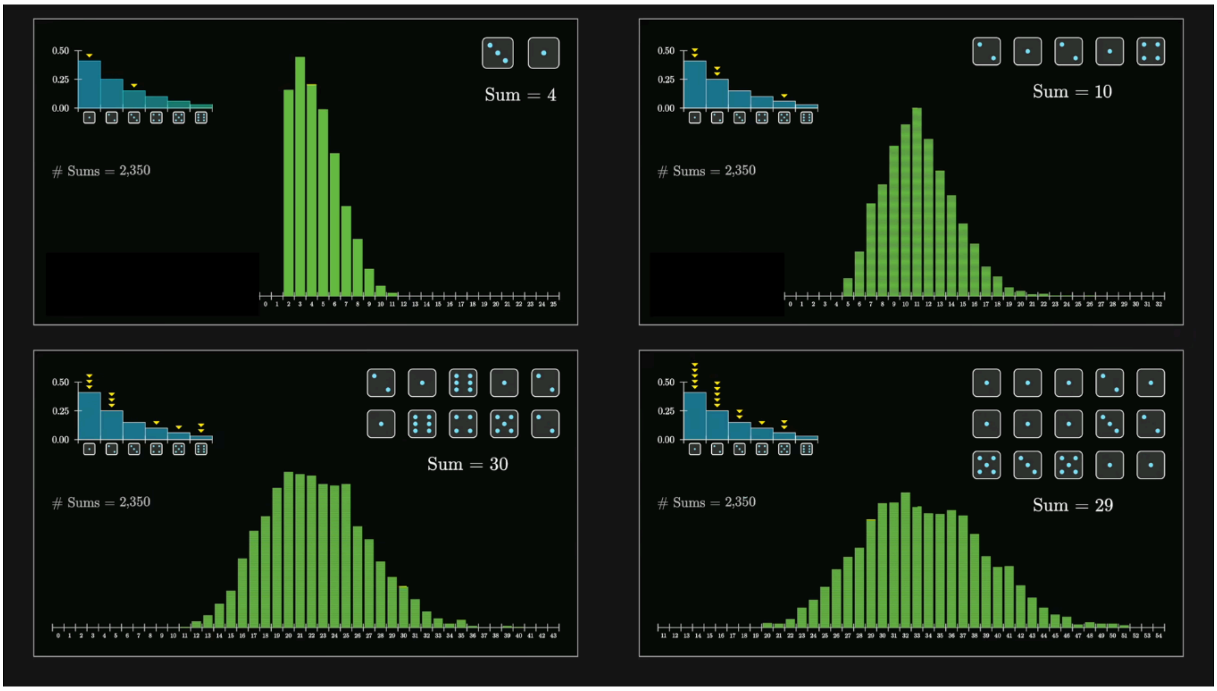

Figure 2: Probability distribution of $X_1 + \ldots + X_n$ with various n

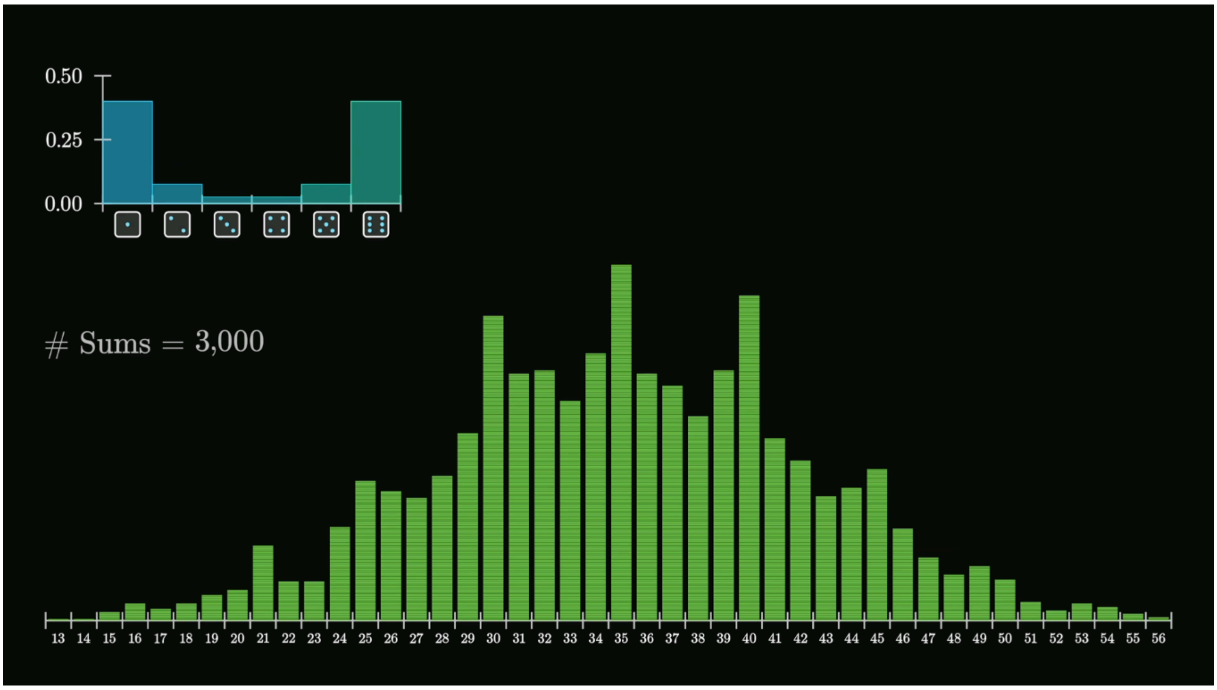

Figure 3: Probability distribution of $X_1 + ... + X_n$ with biased distribution n

Theorem 5.1: Lindeberg–Lévy CLT

$$\left.\begin{array}{l} X_1,...,X_n \text{ is random sample of size } n \text{ with } \mu,\sigma^2 \\ 0 < \operatorname{Var}[X_i] < \infty \ , \ \ \forall i \end{array}\right\}$$

$$\Rightarrow \lim_{n\to\infty} \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \lim_{n\to\infty} Z \sim \mathcal{N}(0,1)$$

Proof:

$$\begin{aligned} \mu_{\overline{X}} &= E\left[\overline{X}\right] \\ &= E\left[\frac{X_1 + ... + X_n}{n}\right] \\ &= \frac{1}{n} \cdot (E[X_1] + ..E[X_n]) \\ &= \mu \end{aligned}$$

$$\begin{aligned} \sigma_{\overline{X}} &= \sqrt{\operatorname{Var}\left[\overline{X}\right]} \\ &= \sqrt{\operatorname{Var}\left[\frac{X_1 + ... + X_n}{n}\right]} \\ &= \sqrt{\frac{1}{n^2} \cdot (V[X_1] + ..V[X_n])} \\ &= \frac{\sigma}{\sqrt{n}} \end{aligned}$$

□

Example: An electronics company manufactures resistor that have a mean resistance of 100 ohms and a standard deviation of 10 ohms. The distribution of resistance if normal. Find the probability that a random sample of n = 25 resistors will have an average resistance less than 95 ohms.

$$\mu_{\overline{X}} = \mu = 100$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

$$\Rightarrow P\left(\overline{X} < 95\right) \approx 0.0062$$

Theorem 5.2: Two variable CTL

Given two independent population with mean μ_1 and μ_2 and variances σ_1^2 and σ_2^2. $\overline{X_1}$ and $\overline{X_2}$ are the sample means of two independent random samples of sizes n_1 and n_2

$$\Rightarrow \lim_{n \to \infty} \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \lim_{n \to \infty} Z \sim \mathcal{N}(0,1)$$

Proof:

$$\begin{aligned} \mu_{\overline{X_1} - \overline{X_2}} &= E\left[\overline{X_1} - \overline{X_2}\right] \\ &= E\left[\overline{X_1}\right] - E\left[\overline{X_2}\right] \\ &= \mu_1 - \mu_2 \end{aligned}$$

$$\begin{aligned} \sigma_{\overline{X_1} - \overline{X_2}} &= \sqrt{\operatorname{Var}\left[\overline{X_1} - \overline{X_2}\right]} \\ &= \sqrt{\operatorname{Var}\left[\overline{X_1}\right] + \operatorname{Var}\left[\overline{X_2}\right]} \\ &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \end{aligned}$$

□

Example: The effective life of a component used in a jet-turbine aircraft engine is a random variable with mean 5000 hours and standard deviation 40 hours. The distribution of effective life is fairly close to a normal distribution. The engine manufacturer introduces an improvement into the manufacturing process for this component that increases the mean life to 5050 hours and decreases the standard deviation to 30 hours. Suppose that a random sample of $n_1 = 16$ components is selected from the "old" process and a random sample of $n_2 = 15$ components is selected from the "new" process. What is the probability that the difference in the two sample means $\overline{X_1} - \overline{X_2}$ is at least 25 hours? Assume that the old and new processes can be regarded as independent populations.

$$\mu_{\overline{X_1} - \overline{X_2}} = \mu_1 - \mu_2 = 5000 - 5050 = -50$$

$$\sigma_{\overline{X_1} - \overline{X_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{40^2}{16} + \frac{30^2}{15}} = 4\sqrt{10} \approx 12.65$$

$$P\left(\overline{X_1} - \overline{X_2} > 25\right) = 1.5214 \cdot 10^{-9}$$

6. References

- Central limit theorem
- But what is the Central Limit Theorem? - 3Blue1Brown