Sampling Mean, Variance, and the Law of Large Numbers (rewritten theory)

Let $X_1, \ldots, X_n$ be i.i.d. with population mean $\mu = E[X_i]$ and variance $\sigma^2 = Var(X_i)$ (finite).
Sample statistics:

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

("corrected"/Bessel's) and the uncorrected version

$$U = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

1) Properties of the sampling mean $\bar{X}$

- **Unbiasedness**: $\mathbb{E}[\bar{X}] = \mu$.

- **Variance / standard error**:, $\mathbb{V}ar(\bar{X}) = \sigma^2/n;\ \mathrm{SE}(\bar{X}) = \sigma/\sqrt{n}$. With unknown σ, replace by $S/\sqrt{n}$.

- **Distribution**
  - If the parent is **normal**, then $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.
  - **CLT (general parents)**:. $\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ Hence for large n the mean is approximately normal.
  - If parent is normal and σ unknown, the **t-statistic**

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

- **Consistency (LLN)**: $\bar{X} \xrightarrow{p} \mu$ as $n \to \infty$ (and almost surely under mild conditions).
- **Independence (normal case only)**: For normal parents, $\bar{X}$ and S^2 are independent—useful for inference.

2) Properties of the sampling variance

- **Unbiasedness (corrected)**:. $\mathbb{E}[S^2] = \sigma^2$.
- **Bias (uncorrected)**: $\mathbb{E}[U] = \frac{n-1}{n}\sigma^2$ (biased low).

- **Sampling distribution (normal parent)**:

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

Consequently,

$$\mathbb{E}[S^2] = \sigma^2, \qquad \mathbb{V}ar(S^2) = \frac{2\sigma^4}{n-1}.$$

For U, $\mathbb{V}ar(U) = \left(\frac{n-1}{n}\right)^2 \mathbb{V}ar(S^2).$

- **General-parent variability**: letting $k=\mu_4/\sigma^4$ be kurtosis,

$$\mathbb{V}ar(S^2) = \frac{\sigma^4}{n}\left(k - \frac{n-3}{n-1}\right)$$

(reduces to $2\sigma^4/(n-1)$ when k=3, i.e., normal). Heavy tails (k>3) inflate the dispersion of S^2.

- **Consistency**: $S^2 \xrightarrow{p} \sigma^2.$

3) Law of Large Numbers (LLN)

- **Weak LLN**:. $\bar{X} \xrightarrow{p} \mu$ when $\mathbb{E}|X| < \infty$.
- **Strong LLN**: $\bar{X} \xrightarrow{a.s.} \mu$ under similar conditions.

Intuition: averages stabilize as nn grows; variability shrinks at rate $1/\sqrt{n}$ (via SE).

4) Illustrative applications in cybersecurity

- **Traffic baselining & anomaly detection**
Use rolling $\bar{X}$ (e.g., bytes/flow, inter-arrival time, failed-login count). By LLN the baseline stabilizes; trigger alerts when

$$\bar{x}_{\text{window}} \notin \bar{x}_{\text{baseline}} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}.$$

- **User behavior analytics**
Per-user averages of logins, session duration, API calls. Significant deviations from the learned mean/variance flag account takeovers or insider threats.

- **Detector performance estimation**
  Estimate false-positive rate p from repeated trials; $\hat{p} \to p$ by LLN, and $\mathrm{SE}(\hat{p}) = \sqrt{p(1-p)/n}$ guides sample size and confidence intervals.
- **Randomness tests for crypto/keys**
  For RNG or key-material checks, means/variances of bit streams should match theory (e.g., mean ≈0.5); chi-square and variance-based tests rely on the sampling distributions above.
- **Capacity planning / DDoS triage**
  Stable long-run means of requests/sec help distinguish genuine step changes from short spikes; heavy-tailed metrics warn that S^2 is more volatile (use robust estimators).

5) Takeaways

- is $\bar{X}$ is unbiased, approximately/ exactly normal (CLT/normal parent), and concentrates at rate $1/\sqrt{n}$.
- S^2 is unbiased with Bessel's correction; under normality its scaled form is $(X_{n-1})^2$ with variance $2\sigma^4/(n-1)$; heavy tails increase its variability.
- LLN underpins practical baselining and thresholding in security analytics, while awareness of distributional assumptions (normal vs heavy-tailed) prevents overconfident decisions.