Chapter 21

Simulation and the Central Limit Theorem

Lecture 21 | Devore 9e: 5.4–5.5

Last time: samples, statistics, and the closed-form derivation of the sampling distribution of $\bar{X}$ for two exponentials. That derivation was already unpleasant. Today: what to do when it is impossible.

21.1 Numerical simulation

Closed form works when both the underlying distribution and the statistic are simple. It fails as soon as

- n is large — imagine the region $\{x_1 + \cdots + x_{100} \le 100t\}$;
- the underlying distribution is complicated;
- the statistic is awkward, e.g. $|X_1| + |X_2| + \cdots + |X_n|$ or the rank of X_1 among $X_1, \ldots, X_n$.

Simulation handles all of them. The setup:

- a statistic $T = T(X_1, \ldots, X_n)$ — e.g. $\bar{X}$ or S;
- a sample distribution, which we can *draw from* even if we cannot integrate against it (a black box);
- the goal: $F_T(t) = \mathbb{P}\{T \le t\}$, or the pdf of T.

The algorithm. Repeat k times:

$$\begin{aligned} t_1 &= T(x_1, \ldots, x_n) \quad \text{from a freshly generated sample} \\ t_2 &= T(x_1, \ldots, x_n) \quad \text{from another fresh sample} \\ &\vdots \\ t_k &= T(x_1, \ldots, x_n) \end{aligned}$$

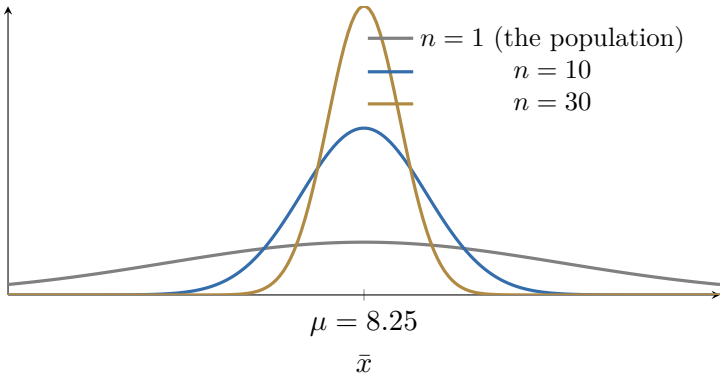
Then histogram $t_1, \ldots, t_k$. That histogram *is* an estimate of the pdf of T, and the proportion of t_i's below t estimates $F_T(t)$.

Example 21.1. Take $X_i \sim \text{Normal}(\mu, \sigma^2)$ with $\mu = 8.25$, $\sigma = 0.75$, sample size $n = 10$, and $k = 500$ repetitions. Each repetition gives one number $\bar{x}_j = (x_1 + \cdots + x_{10})/10$; we get 500 of them — 8.0, 8.1, 7.9, …, 8.6, 7.5 — and histogram them.

Remark 21.2 (Two knobs, two different effects)**.**

larger k	more *accuracy*: the histogram converges to the true pdf of T, whatever it is.
larger n	the pdf of T *itself changes*: the peak stays at μ but the spread shrinks.

Confusing these two is the classic error. Increasing k tells you more about the same object; increasing n changes the object.



21.2 Mean and variance of the sample mean

Before the theorem, two computations that need no assumptions beyond i.i.d.

Theorem 21.3. *If $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2, then*

$$\mathbb{E}[\bar{X}] = \mu, \qquad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}, \qquad \text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Proof. By linearity of expectation (which needs no independence),

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]}{n} = \frac{n\mu}{n} = \mu.$$

For the variance we *do* need independence, which makes variances add:

$$\text{Var}(\bar{X}) = \frac{1}{n^2}\,\text{Var}(X_1 + \cdots + X_n) = \frac{\text{Var}(X_1) + \cdots + \text{Var}(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$ □

Remark 21.4 (The $\sqrt{n}$ law)**.** The precision of an average improves like $\sqrt{n}$, not like n. To halve your error bar you must *quadruple* the sample size. This single fact governs the cost of every survey and every experiment.

21.3 The central limit theorem

Simulation shows something striking: whatever you start with — Weibull, uniform, something lumpy and skewed — the histogram of $\bar{X}$ comes out bell-shaped once n is large enough.

Theorem 21.5 (Central limit theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. with mean μ and variance $\sigma^2 < \infty$. Then for large n,*

$$\bar{X} \overset{approx}{\sim} \text{Normal}\left(\mu,\ \frac{\sigma^2}{n}\right), \qquad \textit{equivalently} \qquad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{approx}{\sim} \text{Normal}(0,1).$$

The mean and variance are exactly what we computed above; the content of the theorem is that the *shape* is normal, regardless of the underlying distribution.

Remark 21.6 (Rule of thumb)**.** $n > 30$ is the usual working threshold. It is only a rule of thumb: a nearly symmetric underlying distribution needs far fewer, and a heavily skewed one (or one with a heavy tail) needs far more. Note the hypothesis $\sigma^2 < \infty$ — for the heavy-tailed pmf of Lecture 9, where even the mean is infinite, the theorem simply does not apply.

Example 21.7 (Why this matters)**.** $X_1, \ldots, X_{1000}$ are i.i.d. from some complicated Weibull. What is $\mathbb{P}\{\bar{X} \le t\}$? Deriving it exactly is hopeless. By the CLT,

$$\mathbb{P}\{\bar{X} \le t\} \approx \Phi\left(\frac{t-\mu}{\sigma/\sqrt{1000}}\right),$$

which is a table lookup.

Example 21.8 (Normal approximation to the binomial, explained)**.** $X \sim \text{Binomial}(n,p)$ is the sum of n i.i.d. Bernoulli variables, each with mean p and variance $p(1-p)$. Applying the CLT to that sum recovers exactly the approximation quoted in Lecture 14:

$$X \overset{\text{approx}}{\sim} \text{Normal}\big(np,\ np(1-p)\big).$$

21.4 The exactly-normal case

When the sample is drawn from a normal population, the approximations become identities and the sample variance gets a distribution too.

Theorem 21.9. *Let $X_1, \ldots, X_n$ be i.i.d.* Normal(μ, σ^2)*. Then*

1. $\bar{X} \sim \text{Normal}\left(\mu, \dfrac{\sigma^2}{n}\right)$ *exactly, for every n;*
2. $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$*;*
3. $\bar{X}$ *and* S^2 *are independent.*

Remark 21.10. Part (2) is where the $n-1$ finally earns its keep, and where the name "degrees of freedom" comes from: the n deviations $X_i - \bar{X}$ satisfy one linear constraint, $\sum_i (X_i - \bar{X}) = 0$, so only $n-1$ of them are free. Part (3) — that the sample mean and the sample variance are independent — is special to the normal distribution, and it is what makes the t-statistic work in a later course.