

Central Limit Theorem and Normal Distribution

Notation. $N(\mu, \sigma)$ uses **mean** μ and **standard deviation** σ as its parameters.

From individual observations to a sum

Let $X_1, \dots, X_n$ be independent and identically distributed, with finite mean $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2 \in (0, \infty)$. For the total $T_n = \sum_{i=1}^n X_i$,

$$E[T_n] = \sum_{i=1}^n E[X_i] = n\mu.$$

Linearity of expectation does not need independence. For the variance, independence gives

$$\text{Var}(T_n) = \sum_{i=1}^n \text{Var}(X_i) = n\sigma^2, \quad \text{SD}(T_n) = \sqrt{n}\sigma.$$

These moments are exact. The **Central Limit Theorem** supplies the approximately normal shape for sufficiently large n :

$$T_n \stackrel{\text{approx}}{\sim} N(n\mu, \sqrt{n}\sigma).$$

For normal individual observations, this distribution is exact for every n .

From the sum to the average

For $\bar{X}_n = T_n/n$, scale the mean by $1/n$ and the variance by $1/n^2$:

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

Thus $\bar{X}_n \stackrel{\text{approx}}{\sim} N(\mu, \sigma/\sqrt{n})$. The total's spread grows as $\sqrt{n}$; dividing by n makes the average's spread shrink as $1/\sqrt{n}$. Its standard error is $\sigma/\sqrt{n}$.

From the average to the standard normal

Subtract the average's mean, then divide by its standard deviation:

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{T_n/n - \mu}{\sigma/\sqrt{n}} = \frac{T_n - n\mu}{\sigma\sqrt{n}}.$$

$$E[Z_n] = \frac{\mu - \mu}{\sigma/\sqrt{n}} = 0, \quad \text{Var}(Z_n) = \frac{\text{Var}(\bar{X}_n)}{\sigma^2/n} = 1.$$

Subtracting μ moves the center to zero. Dividing by $\sigma/\sqrt{n}$ measures deviations in *sampling standard deviation units*. Shifting and rescaling preserve the approximately normal shape supplied by the CLT; mean zero and standard deviation one alone do not imply normality.

The rule $n \geq 30$ is only a heuristic: skewness and heavy tails can require larger samples.

The final Central Limit Theorem statement

For iid observations with finite mean μ and $0 < \sigma^2 < \infty$,

$$\boxed{Z_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow[n \rightarrow \infty]{d} N(0, 1).}$$

Equivalently, $\Pr(Z_n \leq z) \rightarrow \Phi(z)$ for every real z , where Φ is the standard normal CDF. The arrow denotes *convergence of sampling distributions*, not pointwise convergence of realized sample values.

The standard normal density and CDF

For $Z \sim N(0, 1)$,

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \Phi(z) = \Pr(Z \leq z) = \int_{-\infty}^z f(t) dt.$$

The density $f(z)$ is a height; the CDF $\Phi(z)$ is an area to the left. In particular, $f(0) \approx 0.3989$, $\Phi(0) = 0.5$, and $\Pr(Z = 0) = 0$. A single point has zero probability under a continuous distribution.

Normal probabilities and percentiles

For a normal random variable W with mean m and standard deviation $s > 0$, standardize with $Z = (W - m)/s$. For $a < b$,

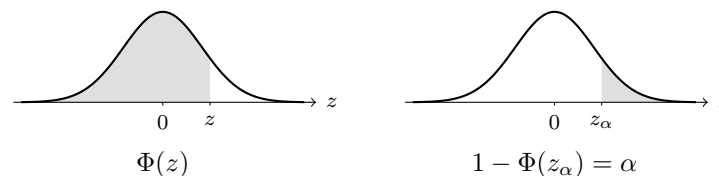
$$\begin{aligned} \Pr(W \leq a) &= \Phi\left(\frac{a - m}{s}\right), \\ \Pr(W > a) &= 1 - \Phi\left(\frac{a - m}{s}\right), \\ \Pr(a < W < b) &= \Phi\left(\frac{b - m}{s}\right) - \Phi\left(\frac{a - m}{s}\right). \end{aligned}$$

Outside an interval, add the two tail areas: $\Pr(Z < a \text{ or } Z > b) = \Phi(a) + 1 - \Phi(b)$. Also, $\Phi(-z) = 1 - \Phi(z)$. For a CLT approximation, use the same formulas with $\approx$ in place of $=$.

For $0 < \alpha < 1$, the upper-tail critical value z_α is defined by

$$\Phi(z_\alpha) = 1 - \alpha, \quad \Pr(Z > z_\alpha) = \alpha.$$

Thus the cutoff with upper-tail probability α is $x = m + sz_\alpha$. For a population measurement, use $s = \sigma$; for a sample mean, use $s = \sigma/\sqrt{n}$.



Other Homework 6 tools

Linear transformations. If $Y = aX + b$, then $E[Y] = a\mu + b$ and $SD(Y) = |a|\sigma$. If X is normal and $a \neq 0$, then $Y \sim N(a\mu + b, |a|\sigma)$.

Normal interquartile range. $Q_{0.75} = \mu + 0.67449\sigma$ and $Q_{0.25} = \mu - 0.67449\sigma$, so $IQR \approx 1.349\sigma$.

Independent sums and differences. For independent U, V ,

$$E[U \pm V] = E[U] \pm E[V], \quad \text{Var}(U \pm V) = \text{Var}(U) + \text{Var}(V).$$

Independent normal variables have exactly normal sums and differences. Standard deviations do not add directly.

Binomial approximation. If $B \sim \text{Bin}(n, p)$ and $q = 1 - p$, use mean np and standard deviation $\sqrt{npq}$. The lecture rule $np > 5$ and $nq > 5$ is an adequacy guideline. Apply a half-unit continuity correction; for example,

$$\Pr(B \leq b) \approx \Phi\left(\frac{b + 0.5 - np}{\sqrt{npq}}\right) \quad \text{for integer } b.$$

Group Problems

Each group should show the random variable, model, probability statement, and reasoning. Each member should be ready to explain the solution.

All numbers are illustrative. For these exercises, assume the CLT approximation is adequate. Sample size alone does not guarantee its accuracy.

z	0	1	1.2816	1.6449	2
$\Phi(z)$	0.5000	0.8413	0.9000	0.9500	0.9772

Table values are rounded. Use $\Phi(-z) = 1 - \Phi(z)$, $z_{0.10} \approx 1.2816$, and $z_{0.05} \approx 1.6449$.

Problem 1 Cell fluorescence

Fluorescence measurements for individual cells in a stable population are strongly right-skewed, with mean 40 units and standard deviation 20 units. A laboratory reports the average fluorescence of 100 independently sampled cells. Let $\bar{X}$ be this sample average and $T = 100\bar{X}$ be the total fluorescence.

- Find the mean, standard deviation, and approximate distribution of both T and $\bar{X}$. Which statements are exact, and which are approximate?
- Approximately what proportion of repeated samples have average fluorescence between 38 and 44 units? Calculate this probability in **two ways**: using the average $\bar{X}$ and using the total T . For each route, state the event, standardize both endpoints, and use the table. Explain why the answers agree. Can this calculation determine the proportion of *individual cells* between 38 and 44 units?
- Find the approximate 90th percentile of $\bar{X}$, starting with a probability equation. Interpret it in terms of repeated samples of 100 cells.

Problem 2 Comparing two fertilizers

A greenhouse compares two-week biomass gains under two fertilizers. For fertilizer A, the population mean is 14 g and the standard deviation is 8 g. For fertilizer B, the population mean is 12 g and the standard deviation is 6 g. Both distributions may be skewed.

Initially, 100 seedlings are independently sampled for each fertilizer. Observations within each group are identically distributed, and the two groups are independent. Define $D = \bar{X}_A - \bar{X}_B$.

- Find the mean, standard deviation, and approximate normal distribution of D . Show the variance calculation.
- Despite fertilizer A's greater population mean, what is the approximate probability that its observed sample mean is below fertilizer B's? Express the event in terms of D first.
- Now use n seedlings in *each* group. Find the smallest integer n for which the normal approximation gives at least a 95% probability that A's observed mean exceeds B's by more than 1 g. Show how n enters the standard deviation and probability.

Worked Solutions

Normal notation is $N(\text{mean}, SD)$. Normal probabilities below use the approximations assumed in the problems.

Problem 1 Cell fluorescence

(a) Independence gives the exact moments

$$\begin{aligned} E[T] &= 100(40) = 4000, & \text{Var}(T) &= 100(20^2) = 40000, & SD(T) &= 200, \\ E[\bar{X}] &= 4000/100 = 40, & \text{Var}(\bar{X}) &= 40000/100^2 = 4, & SD(\bar{X}) &= 2. \end{aligned}$$

The CLT gives $T \overset{\text{approx}}{\sim} N(4000, 200)$ and $\bar{X} \overset{\text{approx}}{\sim} N(40, 2)$. Their normal shapes are approximate; the means and standard deviations are exact.

(b) The average and total describe the same event:

$$\begin{aligned} \Pr(38 < \bar{X} < 44) &\approx \Phi\left(\frac{44 - 40}{2}\right) - \Phi\left(\frac{38 - 40}{2}\right) = \Phi(2) - \Phi(-1), \\ \Pr(3800 < T < 4400) &\approx \Phi\left(\frac{4400 - 4000}{200}\right) - \Phi\left(\frac{3800 - 4000}{200}\right) = \Phi(2) - \Phi(-1). \end{aligned}$$

Both routes give $0.9772 - (1 - 0.8413) = \boxed{0.8185}$ using the table (unrounded CDFs give 0.8186 to four decimals). About 81.85% of repeated samples have an average in this interval. This does *not* determine the proportion of individual cells in the interval: their distribution is skewed, their SD is 20, and the first two moments alone do not specify that probability.

(c) Start with $\Pr(\bar{X} \leq c) \approx 0.90$. Then

$$\Phi\left(\frac{c - 40}{2}\right) = 0.90 \implies c \approx 40 + 2z_{0.10} \approx 40 + 2(1.2816) \approx \boxed{42.56 \text{ units}}.$$

About 90% of repeated samples of 100 cells have an average at or below this cutoff.

Problem 2 Comparing two fertilizers

(a) For $D = \bar{X}_A - \bar{X}_B$, independence makes the variances add:

$$E[D] = 14 - 12 = 2, \quad \text{Var}(D) = \frac{8^2}{100} + \frac{6^2}{100} = 1, \quad SD(D) = 1.$$

The moments are exact; the approximate distribution is $\boxed{D \overset{\text{approx}}{\sim} N(2, 1)}$ (in grams).

(b) A's observed mean is below B's exactly when $D < 0$. Thus

$$\Pr(\bar{X}_A < \bar{X}_B) = \Pr(D < 0) \approx \Phi\left(\frac{0 - 2}{1}\right) = \Phi(-2) = 1 - \Phi(2) \approx \boxed{0.0228}.$$

The ordering reverses in about 2.28% of repeated pairs of samples.

(c) With n seedlings in each group, $E[D] = 2$, $\text{Var}(D) = 100/n$, and $SD(D) = 10/\sqrt{n}$. The requested event is $D > 1$, so

$$\Pr(D > 1) \approx 1 - \Phi\left(\frac{1 - 2}{10/\sqrt{n}}\right) = \Phi\left(\frac{\sqrt{n}}{10}\right) \geq 0.95 \iff n \geq (10z_{0.05})^2 \approx 270.554.$$

Round up: $\boxed{n = 271 \text{ seedlings per group}}$, or 542 total. Unrounded normal CDF values give 0.949826 at $n = 270$ and 0.950140 at $n = 271$. This minimum is under the normal approximation, not an exact guarantee for arbitrary population distributions.