Sampling distributions and the central limit theorem: standard errors, confidence intervals and the √N rule

Ten problems on the sample mean and the central limit theorem: the mean and standard error of an average and of a sum, the sum of 100 dice with a continuity correction, a poll's margin, the sample size for a target error, a 95% confidence interval, the error of a Monte Carlo estimate, minibatch gradient noise, a sum of exponential waits, an A/B test, and a rare event where the normal approximation fails, with worked solutions and the mistakes of dividing by N instead of √N, standardizing a mean with σ, doubling the data to halve the error, adding standard errors, and trusting the normal when Np is small.

BEFORE YOU START

Every number estimated from data is an average of something: a conversion rate, a validation accuracy, a Monte Carlo integral, a minibatch gradient. The sample mean is itself a random variable, and two facts describe it. Its standard deviation, the standard error, is $\sigma/\sqrt{N}$, so the error shrinks with the square root of the amount of data and not in proportion to it. And whatever the distribution of a single observation, as long as its variance is finite, the central limit theorem says the average is close to normal once N is large. Together they turn a sample into a probability statement: how far the estimate is likely to be from the truth, how many samples a target accuracy needs, and whether a difference between two rates is bigger than the noise. These ten problems work through the standard error, the normal approximation for sums and proportions, confidence intervals, sample sizes, Monte Carlo and minibatch noise and a two-sample comparison, and end with a case where the approximation is poor. The five mistakes are the ones that give a confident wrong number.

- $x_1, \dots, x_N$ are independent draws from one distribution with mean μ and variance $\sigma^2 < \infty$. The **sample mean** is $\bar{x} = \frac{1}{N}\sum_n x_n$ and the **sum** is $S_N = \sum_n x_n = N\bar{x}$.
- $\mathbb{E}[\bar{x}] = \mu$ and $\mathrm{Var}(\bar{x}) = \sigma^2/N$, because the variances of independent terms add and $\mathrm{Var}(cX) = c^2\,\mathrm{Var}(X)$. The **standard error** is $\mathrm{SE} = \sigma/\sqrt{N}$. For the sum, $\mathbb{E}[S_N] = N\mu$ and $\mathrm{Var}(S_N) = N\sigma^2$.
- **Central limit theorem.** For large N, $\dfrac{\bar{x} - \mu}{\sigma/\sqrt{N}}$ is approximately Normal$(0, 1)$, whatever the distribution of each x_n. The approximation is better for symmetric distributions and worse for skewed ones.
- Φ is the standard normal CDF, with $\Phi(-z) = 1 - \Phi(z)$. Useful values: $\Phi(1.25) = 0.8944$, $\Phi(1.645) = 0.95$, $\Phi(1.96) = 0.975$, $\Phi(2) = 0.9772$.

- A **proportion** is the mean of Bernoulli(p) variables, which have variance $p(1-p)$, so the standard error of $\hat{p}$ is $\sqrt{p(1-p)/N}$. A rule of thumb for the normal approximation is $Np \geq 10$ and $N(1-p) \geq 10$.
- A **95% confidence interval** for μ is $\bar{x} \pm 1.96$ SE, with σ replaced by the sample standard deviation s when it is unknown and N is large.
- **Continuity correction.** When a sum of integers is approximated by a normal, $P(S_N \geq k)$ is approximated by the normal probability above $k - \frac{1}{2}$.

Builds on: Variance, covariance and correlation

Problems

1. Measurements have mean $\mu = 10$ and standard deviation $\sigma = 4$. For the average of $N = 25$ of them, find $\mathbb{E}[\bar{x}]$, $\mathrm{Var}(\bar{x})$ and the standard error. Find the mean and standard deviation of their sum. difficulty ·

2. A fair die is rolled 100 times. Use the central limit theorem to approximate the probability that the total is at least 370. difficulty ··

3. In a city where exactly half the voters favor a measure, a poll asks $N = 400$ random voters. Find the standard error of the sample proportion $\hat{p}$, and approximate the probability that the poll shows 55% or more in favor. difficulty ··

4. A proportion is to be estimated with a standard error of at most 0.01, whatever the true p is. How many samples are needed? How many for a standard error of 0.005? difficulty ·

5. A model's latency is measured on $N = 64$ requests, with sample mean $\bar{x} = 52.3$ ms and sample standard deviation $s = 8$ ms. Find a 95% confidence interval for the true mean latency, and the width of the interval with $N = 256$ requests and the same s. difficulty ··

6. To estimate $I = \int_0^1 x^2\,dx$ by Monte Carlo, draw $U_1, \dots, U_N \sim \mathrm{Uniform}(0,1)$ and average U_n^2. Find the mean and variance of one term, the standard error with $N = 10\,000$, and the N needed for a standard error of 0.001. difficulty ··

7. The gradient of a loss with respect to one weight varies across training examples with mean g and standard deviation 3. A minibatch gradient is the average over B examples drawn independently. Find the standard deviation of the minibatch gradient for $B = 32$, and the batch size that halves it. difficulty ··

8. A job runs 50 independent stages, each taking an Exponential time with rate 2 per minute. Approximate the probability that the whole job takes more than 30 minutes, and say why the exact answer is larger. difficulty ···

9. An A/B test shows each version of a page to 2000 visitors. Version A converts 10% of them and version B 12%. Find the standard error of the difference $\hat{p}_B - \hat{p}_A$, the z-statistic, and whether the difference is significant at the two-sided 5% level. difficulty ···

10. A defect occurs on 1% of parts, independently. In a batch of $N = 100$, find the exact probability of no defects, the Poisson approximation, and the normal approximation with a continuity correction. Which is poor, and why? difficulty ···

Answers

1. $\mathbb{E}[\bar{x}] = 10$, $\mathrm{Var}(\bar{x}) = 0.64$, $\mathrm{SE} = 0.8$; the sum has mean 250 and standard deviation 20
2. $P(S_{100} \geq 370) \approx 0.127$
3. $\mathrm{SE} = 0.025$; $P(\hat{p} \geq 0.55) \approx 0.023$, or 0.026 with the continuity correction
4. $N = 2500$ for $\mathrm{SE} \leq 0.01$, and $N = 10000$ for $\mathrm{SE} \leq 0.005$
5. $(50.34, 54.26)$ ms, width 3.92 ms; with 256 requests the width is 1.96 ms
6. $\mathbb{E}[U^2] = \frac{1}{3}$, $\mathrm{Var}(U^2) = \frac{4}{45}$, $\mathrm{SE} \approx 0.0030$ at $N = 10\,000$; $N \approx 88\,900$ for $\mathrm{SE} = 0.001$
7. $\mathrm{sd}(\hat{g}) \approx 0.530$ at $B = 32$; $B = 128$ halves it to ≈ 0.265
8. $P(T > 30) \approx 0.079$ by the central limit theorem; the exact value is 0.084
9. $\mathrm{SE} \approx 0.0099$, $z \approx 2.02$: significant at 5%, since $|z| > 1.96$ (two-sided $p \approx 0.043$)
10. Exact 0.3660, Poisson 0.3679, normal 0.3077; the normal is poor because $Np = 1$

Worked solutions

Problem 1

> Measurements have mean $\mu = 10$ and standard deviation $\sigma = 4$. For the average of $N = 25$ of them, find $\mathbb{E}[\bar{x}]$, $\mathrm{Var}(\bar{x})$ and the standard error. Find the mean and standard deviation of their sum.

$\mathbb{E}[\bar{x}] = \frac{1}{25}\sum_n \mathbb{E}[x_n] = \frac{25 \cdot 10}{25} = 10$	Linearity of expectation; independence is not needed for this part.
$\mathrm{Var}(\bar{x}) = \frac{1}{25^2}\sum_n \mathrm{Var}(x_n) = \frac{25 \cdot 16}{625} = \frac{16}{25} = 0.64$	Independence makes the variances add, and the factor $\frac{1}{25}$ comes out squared.
$\mathrm{SE} = \sqrt{0.64} = 0.8 = \frac{4}{\sqrt{25}}$	The standard deviation of the average is $\sigma/\sqrt{N}$.
$\mathbb{E}[S_{25}] = 250$ and $\mathrm{Var}(S_{25}) = 25 \cdot 16 = 400$	The sum is $25\bar{x}$, so its variance is $25^2 \cdot 0.64$.
$\mathbb{E}[\bar{x}] = 10$, $\mathrm{Var}(\bar{x}) = 0.64$, $\mathrm{SE} = 0.8$; **the sum has mean** 250 **and standard deviation** 20	The sum spreads out like $\sqrt{N}\sigma$ while the average tightens like $\sigma/\sqrt{N}$. Dividing by N instead of $\sqrt{N}$ gives 0.16 (Mistake 1).

Problem 2

> A fair die is rolled 100 times. Use the central limit theorem to approximate the probability that the total is at least 370.

One roll: $\mu = \frac{1+2+\cdots+6}{6} = 3.5$ and $\mathbb{E}[x^2] = \frac{91}{6}$, so $\sigma^2 = \frac{91}{6} - 3.5^2 = \frac{35}{12}$	$1 + 4 + 9 + 16 + 25 + 36 = 91$.
$\mathbb{E}[S_{100}] = 350$ and $\mathrm{Var}(S_{100}) = 100 \cdot \frac{35}{12} \approx 291.67$, so $\mathrm{sd}(S_{100}) \approx 17.078$	Means and variances of independent rolls add.
$P(S_{100} \geq 370) \approx P\left(Z \geq \frac{369.5 - 350}{17.078}\right) = P(Z \geq 1.1418)$	The total is an integer, so $S \geq 370$ is the same event as $S > 369.5$; the continuity correction uses the midpoint.

$1 - \Phi(1.1418) \approx 0.1268$	From a table or `scipy.stats.norm.sf`.
$P(S_{100} \geq 370) \approx 0.127$	The exact probability, by convolving the die's distribution with itself 100 times, is 0.1269. Without the continuity correction, $z = 20/17.078 = 1.171$ and the estimate is 0.1208, noticeably worse.

Problem 3

> In a city where exactly half the voters favor a measure, a poll asks $N = 400$ random voters. Find the standard error of the sample proportion $\hat{p}$, and approximate the probability that the poll shows 55% or more in favor.

$\mathrm{SE}(\hat{p}) = \sqrt{\dfrac{0.5 \cdot 0.5}{400}} = \dfrac{0.5}{20} = 0.025$	A proportion is a mean of Bernoulli(0.5) variables, each with variance 0.25.
$z = \dfrac{0.55 - 0.5}{0.025} = 2$	Standardize with the standard error, not with the standard deviation of one vote.
$P(\hat{p} \geq 0.55) \approx 1 - \Phi(2) \approx 0.0228$	The normal approximation is fine here: $Np = N(1-p) = 200$.
With the continuity correction, $\hat{p} \geq 0.55$ is $k \geq 220$ yes-votes, and $P(k \geq 219.5) \approx 1 - \Phi(1.95) \approx 0.0256$	k has mean 200 and standard deviation $\sqrt{400 \cdot 0.25} = 10$.
$\mathrm{SE} = 0.025$; $P(\hat{p} \geq 0.55) \approx 0.023$, **or** 0.026 **with the continuity correction**	The exact binomial probability is 0.0255. The usual "margin of error" of such a poll is $1.96 \times 0.025 \approx 4.9$ percentage points.

Problem 4

> A proportion is to be estimated with a standard error of at most 0.01, whatever the true p is. How many samples are needed? How many for a standard error of 0.005?

$p(1-p) \leq \frac{1}{4}$, with equality at $p = \frac{1}{2}$	The worst case: $p(1-p)$ is a downward parabola with its peak at $\frac{1}{2}$.

$\sqrt{\dfrac{0.25}{N}} \le 0.01 \iff N \ge \dfrac{0.25}{0.0001} = 2500$	Square both sides and solve for N.
$\sqrt{\dfrac{0.25}{N}} \le 0.005 \iff N \ge \dfrac{0.25}{0.000025} = 10000$	The same calculation with the smaller target.
Halving the standard error multiplies N by 4	$\text{SE} \propto 1/\sqrt{N}$, so $N \propto 1/\text{SE}^2$.
$N = 2500$ **for** $\text{SE} \le 0.01$, **and** $N = 10000$ **for** $\text{SE} \le 0.005$	Each extra decimal digit of accuracy costs 100 times the data. Doubling to 5000 gives $\text{SE} = 0.00707$, not 0.005 (Mistake 3).

Problem 5

> A model's latency is measured on $N = 64$ requests, with sample mean $\bar{x} = 52.3$ ms and sample standard deviation $s = 8$ ms. Find a 95% confidence interval for the true mean latency, and the width of the interval with $N = 256$ requests and the same s.

$\text{SE} \approx \dfrac{s}{\sqrt{N}} = \dfrac{8}{8} = 1$ ms	σ is unknown, so the sample standard deviation stands in for it; with $N = 64$ that is a good approximation.
$\bar{x} \pm 1.96\ \text{SE} = 52.3 \pm 1.96$	1.96 leaves 2.5% in each tail: $\Phi(1.96) = 0.975$.
The interval is $(50.34, 54.26)$ ms	Procedures like this one cover the true mean in 95% of samples.
With $N = 256$, $\text{SE} = \frac{8}{16} = 0.5$ and the half-width is 0.98 ms	Four times the requests halves the interval.
$(50.34, 54.26)$ **ms, width** 3.92 **ms; with** 256 **requests the width is** 1.96 **ms**	The 95% describes the procedure, not this particular interval: before the data are drawn, $P(\lvert\bar{x} - \mu\rvert \le 1.96\ \text{SE}) \approx 0.95$. Using 1.645, the one-sided 95% point, gives a 90% interval.

Problem 6

To estimate $I = \int_0^1 x^2\,dx$ by Monte Carlo, draw $U_1, \dots, U_N \sim \text{Uniform}(0,1)$ and average U_n^2. Find the mean and variance of one term, the standard error with $N = 10\,000$, and the N needed for a standard error of 0.001.

$\mathbb{E}[U^2] = \int_0^1 u^2\,du = \dfrac{1}{3}$	The estimator is unbiased: its mean is the integral.
$\mathbb{E}[U^4] = \dfrac{1}{5}$, so $\text{Var}(U^2) = \dfrac{1}{5} - \dfrac{1}{9} = \dfrac{4}{45} \approx 0.0889$	The variance shortcut, with U^2 as the variable.
$\text{SE} = \sqrt{\dfrac{4/45}{10000}} \approx 0.00298$	Standard deviation of one term, 0.2981, over $\sqrt{10000} = 100$.
$\dfrac{4/45}{N} \le 10^{-6} \iff N \ge \dfrac{4}{45} \times 10^6 \approx 88\,889$	Square the target.
$\mathbb{E}[U^2] = \frac{1}{3}$, $\text{Var}(U^2) = \frac{4}{45}$, $\text{SE} \approx 0.0030$ **at** $N = 10\,000$; $N \approx 88\,900$ **for** $\text{SE} = 0.001$	The error of Monte Carlo shrinks like $1/\sqrt{N}$ in any dimension, which is why it is used for high-dimensional integrals where grids are hopeless, and why variance reduction pays.

Problem 7

The gradient of a loss with respect to one weight varies across training examples with mean g and standard deviation 3. A minibatch gradient is the average over B examples drawn independently. Find the standard deviation of the minibatch gradient for $B = 32$, and the batch size that halves it.

The minibatch gradient is $\hat{g} = \frac{1}{B}\sum_{b=1}^{B} g_b$, with $\mathbb{E}[\hat{g}] = g$	An average of unbiased per-example gradients is unbiased.
$\text{sd}(\hat{g}) = \dfrac{3}{\sqrt{32}} \approx 0.530$	The standard error of a mean.

Halving it needs $\sqrt{B}$ to double, so $B = 4 \cdot 32 = 128$	$\frac{3}{\sqrt{128}} \approx 0.265$.
128 examples cost 4 times the computation of 32	One step with $B = 128$ costs as much as four steps with $B = 32$, and only halves the noise.
$\mathrm{sd}(\hat{g}) \approx 0.530$ **at** $B = 32$; $B = 128$ **halves it to** ≈ 0.265	This is the arithmetic behind the trade-off between batch size and number of steps. Sampling without replacement from a finite dataset makes the noise slightly smaller, by the factor $\sqrt{(M - B)/(M - 1)}$ for a dataset of M examples.

Problem 8

> A job runs 50 independent stages, each taking an Exponential time with rate 2 per minute. Approximate the probability that the whole job takes more than 30 minutes, and say why the exact answer is larger.

One stage: mean $\frac{1}{2}$ minute and variance $\frac{1}{2^2} = \frac{1}{4}$	For an Exponential(λ), the mean is $\frac{1}{\lambda}$ and the variance $\frac{1}{\lambda^2}$.
The total T has $\mathbb{E}[T] = 25$ and $\mathrm{Var}(T) = 12.5$, so $\mathrm{sd}(T) \approx 3.536$	Fifty independent stages.
$P(T > 30) \approx 1 - \Phi\left(\frac{30 - 25}{3.536}\right) = 1 - \Phi(1.414) \approx 0.0786$	No continuity correction: T is continuous.
The exact T has a Gamma$(50, 2)$ distribution, and $P(T > 30) = P(\mathrm{Poisson}(60) \le 49) \approx 0.0844$	More than 30 minutes means at most 49 stage completions in 30 minutes of a rate-2 process.

$P(T > 30) \approx 0.079$ **by the central limit theorem; the exact value is** 0.084	Each exponential is skewed to the right, and 50 terms are not quite enough to make the sum symmetric, so the right tail is heavier than the normal's.

Problem 9

An A/B test shows each version of a page to 2000 visitors. Version A converts 10% of them and version B 12%. Find the standard error of the difference $\hat{p}_B - \hat{p}_A$, the z-statistic, and whether the difference is significant at the two-sided 5% level.

$\mathrm{Var}(\hat{p}_A) = \dfrac{0.10 \cdot 0.90}{2000} = 4.5 \times 10^{-5}$ and $\mathrm{Var}(\hat{p}_B) = \dfrac{0.12 \cdot 0.88}{2000} = 5.28 \times 10^{-5}$	The variance of each proportion, with the observed rates in place of the unknown ones.
$\mathrm{Var}(\hat{p}_B - \hat{p}_A) = 4.5 \times 10^{-5} + 5.28 \times 10^{-5} = 9.78 \times 10^{-5}$	The two groups are independent, so the variances add, even for a difference.
$\mathrm{SE} = \sqrt{9.78 \times 10^{-5}} \approx 0.00989$	The square root comes after the adding.
$z = \dfrac{0.12 - 0.10}{0.00989} \approx 2.02$	The difference in units of its standard error.
$\mathrm{SE} \approx 0.0099$, $z \approx 2.02$: **significant at 5%, since** $\lvert z \rvert > 1.96$ **(two-sided** $p \approx 0.043$**)**	Only just. Adding the two standard errors instead of the variances gives 0.0140 and $z = 1.43$ (Mistake 4).

Problem 10

A defect occurs on 1% of parts, independently. In a batch of $N = 100$, find the exact probability of no defects, the Poisson approximation, and the normal approximation with a continuity correction. Which is poor, and why?

$P(k = 0) = 0.99^{100} \approx 0.3660$	Each of the 100 parts must be good.
Poisson with $\lambda = Np = 1$: $P(k = 0) \approx e^{-1} \approx 0.3679$	For a rare event in many trials, the binomial is close to a Poisson with the same mean.

Normal: mean 1, standard deviation $\sqrt{100 \cdot 0.01 \cdot 0.99} \approx 0.995$, and $P(k \leq 0.5) \approx \Phi\left(\frac{0.5-1}{0.995}\right) = \Phi(-0.5025) \approx 0.3077$	Continuity correction: $k = 0$ is everything below 0.5.
$Np = 1$, far below the rule of thumb of 10	The distribution of k is piled up at 0 and 1 with a long right tail, nothing like a bell curve, and the normal also puts probability on negative counts.
Exact 0.3660, **Poisson** 0.3679, **normal** 0.3077; **the normal is poor because** $Np = 1$	The central limit theorem is a statement about large N for a fixed distribution; when p shrinks as N grows the limit is the Poisson, not the normal.

Where this goes wrong

1. Dividing σ by N instead of √N

The average of N values is the sum divided by N, so its spread is divided by N too.

$\sigma = 4$ and $N = 25$, as in Problem 1	Right so far.
"The average divides everything by N, including the standard deviation."	Dividing by N divides the standard deviation of the sum by N, but the sum's standard deviation is $\sqrt{N}\sigma = 20$, not $N\sigma$.
SE $= \frac{4}{25} = 0.16$ ✗	Five times too small. $\mathrm{Var}(\bar{x}) = \sigma^2/N = 0.64$, so SE $= 0.8$. An interval built on 0.16 would claim far more precision than 25 measurements can give.

2. Standardizing a mean with σ

The z-score of a single observation is reused for an average.

Measurements have $\mu = 10$ and $\sigma = 4$, and $\bar{x}$ averages $N = 25$ of them	Right so far: Problem 1.

“$P(\bar{x} > 11)$: $z = \frac{11-10}{4} = 0.25$.”	The step that causes the mistake: 4 is the spread of one measurement; the average has spread 0.8.
$P(\bar{x} > 11) \approx 1 - \Phi(0.25) \approx 0.4013$ ✗	That is the chance that a single measurement exceeds 11. For the average, $z = \frac{1}{0.8} = 1.25$ and $P(\bar{x} > 11) \approx 1 - 0.8944 = 0.1056$.

3. Doubling the data to halve the error

The error is assumed to shrink in proportion to the amount of data.

$N = 2500$ gives SE ≤ 0.01 for a proportion	Right so far: Problem 4.
“Twice the data, half the error.”	The error is proportional to $1/\sqrt{N}$, not $1/N$.
$N = 5000$ gives SE ≤ 0.005 ✗	$\sqrt{0.25/5000} \approx 0.00707$: doubling the data cuts the error only by $\sqrt{2}$. Halving it needs four times the data, $N = 10\,000$.

4. Adding standard errors instead of variances

The standard error of a difference of two independent estimates is taken as the sum of their standard errors.

$\text{SE}(\hat{p}_A) = \sqrt{4.5 \times 10^{-5}} \approx 0.00671$ and $\text{SE}(\hat{p}_B) = \sqrt{5.28 \times 10^{-5}} \approx 0.00727$	Right so far: Problem 9.
“$\text{SE}(\hat{p}_B - \hat{p}_A) = 0.00671 + 0.00727 = 0.01397$.”	Standard deviations do not add; variances of independent quantities do.
$z = \frac{0.02}{0.01397} \approx 1.43$, so the difference is not significant ✗	The standard error is $\sqrt{4.5 \times 10^{-5} + 5.28 \times 10^{-5}} \approx 0.00989$ and $z \approx 2.02$, which is significant at 5%. Adding standard errors always overstates the noise, here enough to reverse the conclusion.

5. The normal approximation when Np is small

The central limit theorem is applied to every binomial count.

$k \sim \text{Binomial}(100, 0.01)$, with mean 1 and standard deviation 0.995	Right so far: Problem 10.
"100 trials is a large sample, so k is approximately normal."	What matters is the expected number of successes, $Np = 1$, not the number of trials.
$P(k = 0) \approx \Phi(-0.5025) \approx 0.3077$ ✗	The exact value is $0.99^{100} \approx 0.3660$, and the Poisson approximation $e^{-1} \approx 0.3679$ is far closer. The normal also gives $P(k \le -0.5) = \Phi(-1.51) \approx 0.066$ to counts below zero.

More at eulerpad.app. This is one of 82 free practice sets with worked solutions at eulerpad.app/practice, alongside guides to the topics behind them.

EulerPad, the app, reads your handwritten math on iPad, iPhone and Android and checks it line by line. It's free to start.

This set is also Chapter 15 of *Probability & Statistics for Machine Learning*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/sampling-distributions-and-the-central-limit-theorem