Specialist Mathematics Units 3 & 4

Chapter 12 — Confidence intervals and hypothesis testing for the mean

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Attribution: Classbasics Pty Ltd, vwx.au

Aboriginal and Torres Strait Islander content has been excluded as accuracy cannot be guaranteed, and it is better to be respectful than cover the entire curriculum inaccurately.

Significant effort has been made to ensure this content is legal. Please send any areas of concern to admin@vwx.au with appropriate document/legal references so errors can be verified and changes be made.

Contents

Theory

A population has a mean μ that is usually **unknown**. To estimate it we take a random sample $X_1, X_2, \ldots, X_n$ and compute the **sample mean** $\bar{X}$. From the previous chapter, $\bar{X}$ is an **unbiased** estimator of μ — its own mean is $E(\bar{X}) = \mu$ — with standard deviation (the **standard error**)

$$sd(\bar{X}) = \frac{\sigma}{\sqrt{n}},$$

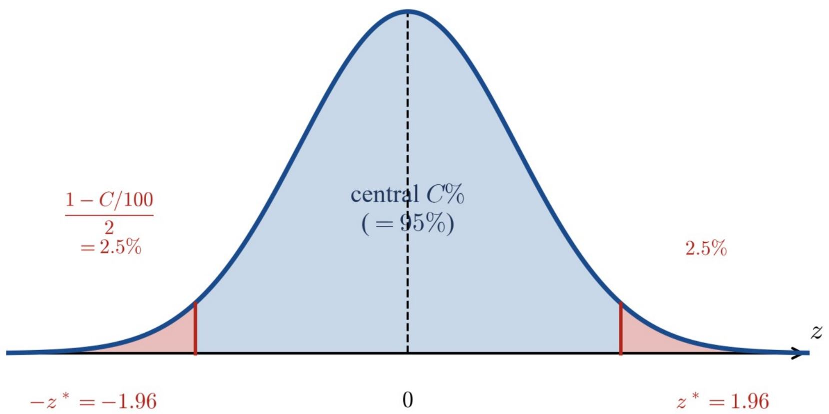
where σ is the population standard deviation. A single observed value $\bar{x}$ is a **point estimate** of μ: one number, almost certainly not exactly equal to μ. This section instead reports an **interval** of plausible values together with a stated level of reliability — a **confidence interval**.

The idea.

If the population is normal, or the sample is large, then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ (exactly so for a normal population, approximately so for a large sample), so standardising gives $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$. For a chosen **confidence level** $C\%$, let z^* be the value that places the central $C\%$ of the standard normal between $-z^*$ and z^*:

$$P(-z^* < Z < z^*) = \frac{C}{100}, z^* = \Phi^{-1}\left(\frac{1 + C/100}{2}\right),$$

where Φ^{-1} is the inverse standard-normal function. Each tail then carries area $\frac{1 - C/100}{2}$.



The most common levels give the multipliers below; these three values should be known.

Confidence level $C\%$	Tail area each side	$z^* = \Phi^{-1}\left(\frac{1+C/100}{2}\right)$
90%	0.05	1.645
95%	0.025	1.96
99%	0.005	2.576

Building the interval.

Substituting the standardised sample mean into $-z^* < Z < z^*$,

$$-z^* < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z^*,$$

and rearranging to isolate μ (multiply by the positive $\sigma/\sqrt{n}$, then subtract $\bar{X}$ and multiply by -1, which reverses the inequalities),

$$\bar{X} - z^*\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z^*\frac{\sigma}{\sqrt{n}}.$$

So the **random** interval $\bar{X} \pm z^*\frac{\sigma}{\sqrt{n}}$ captures the fixed unknown μ with probability $\frac{C}{100}$. After the sample is taken, $\bar{X}$ takes the observed value $\bar{x}$, and the $C\%$ **confidence interval** for μ is

$$\bar{x} \pm z^*\frac{\sigma}{\sqrt{n}} \text{ that is } \left(\bar{x} - z^*\frac{\sigma}{\sqrt{n}}, \bar{x} + z^*\frac{\sigma}{\sqrt{n}}\right).$$

Margin of error.

The half-width of the interval is the **margin of error**

$$E = z^*\frac{\sigma}{\sqrt{n}},$$

so the interval is $\bar{x} \pm E$ and its full width is $2E$. The margin measures the precision of the estimate: a smaller E means a tighter interval.

Interpreting a confidence interval.

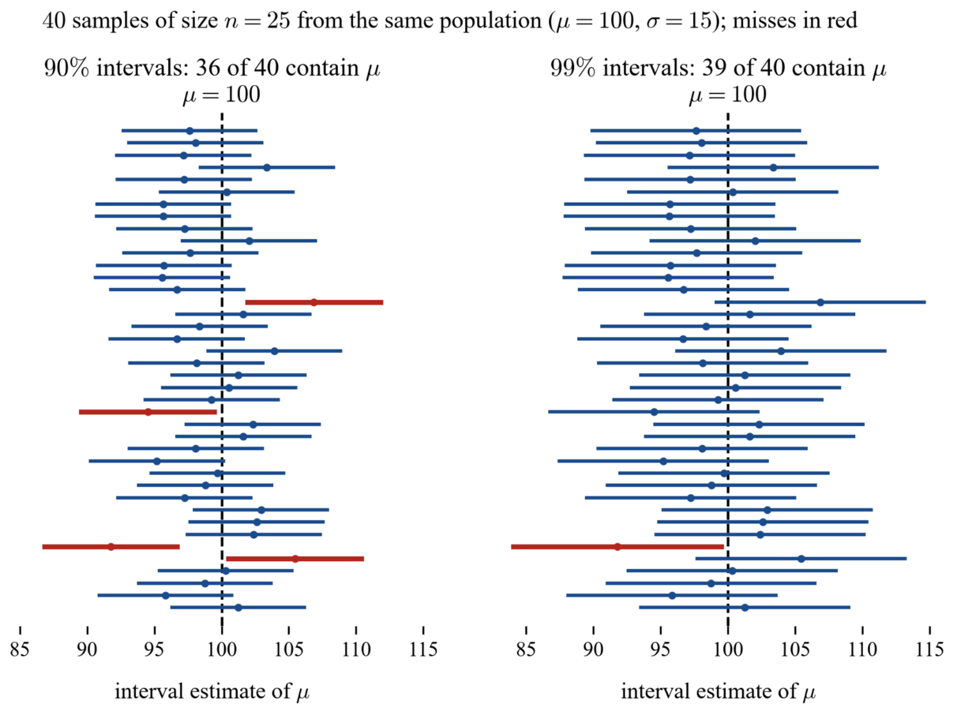
The correct reading is: *we are $C\%$ confident that the interval contains μ.* The subtle point is **which quantity is random**. Here μ is a fixed (though unknown) constant; it is the **interval** that is random, because it is built from $\bar{X}$ and shifts from sample to sample. If we repeated the whole procedure many times, about $C\%$ of the intervals produced would contain μ and about $(100-C)\%$ would miss it.

For a **computed** interval such as $(20.04, 23.96)$, it is therefore **incorrect** to write $P(20.04 < \mu < 23.96) = 0.95$: once the numbers are fixed, μ either lies in the interval or it does not — there is no longer any randomness to assign a probability to. The 95% refers to the reliability of the *method*, not to a probability about this one interval.

Seeing it by simulation.

Because "about $C\%$ of the intervals contain μ" is a statement about repeated sampling, it can be checked directly by **simulation**: choose a population with a known μ, draw many samples of the same size, build an interval from each, and count how many capture μ. Two features stand out. First, the intervals **vary from sample to sample** — they are centred on different values of $\bar{x}$, so some sit well left of μ and some well right. Second, the proportion that capture μ settles near the confidence level chosen.

The simulation below draws 40 samples of size $n = 25$ from a normal population with $\mu = 100$ and $\sigma = 15$ (so the standard error is 3), then builds an interval from **each of the same 40 samples** at two confidence levels.



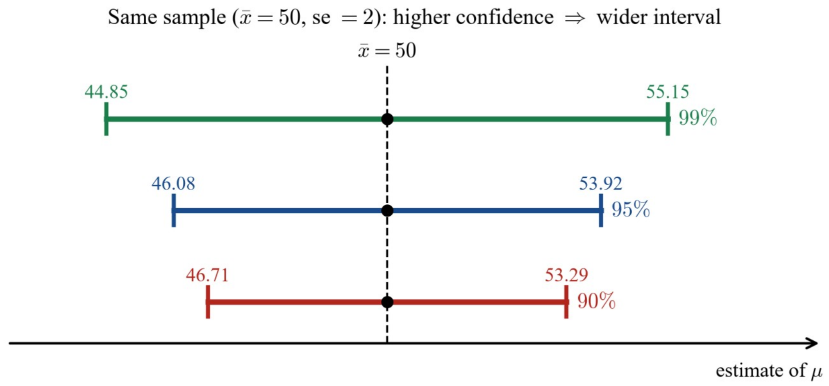
At 90 % confidence, 36 of the 40 intervals contain μ (about 90 %, as expected) and 4 miss. Raising the level to 99 % widens every interval about the same $\bar{x}$, and now 39 of the 40 contain μ: the chance that an interval captures μ **depends on the confidence level chosen**, and higher confidence buys that extra capture rate with a wider, less precise interval. In a longer run the capture proportions approach 90 % and 99 % exactly.

Effect of n and of the confidence level.

The width $2E = 2z^* \dfrac{\sigma}{\sqrt{n}}$ shows two levers.

- **Sample size n.** Because $E \propto \dfrac{1}{\sqrt{n}}$, a larger sample narrows the interval, but only slowly: to **halve** the margin the sample size must be **quadrupled**.
- **Confidence level C %.** A higher confidence needs a larger z^* (compare 1.645, 1.96, 2.576), so — for the same data — a 99 % interval is **wider** than a 95 % interval, which is wider than a 90 % interval. Greater confidence is bought with less precision.

The diagram shows the same sample ($\bar{x} = 50$, standard error 2) turned into 90 %, 95 % and 99 % intervals: all are centred on $\bar{x} = 50$, and each higher level is wider.



Choosing the sample size for a required margin.

To guarantee a margin of error at most a target value E, set

$$z^* \frac{\sigma}{\sqrt{n}} \le E \quad \sqrt{n} \ge \frac{z^* \sigma}{E} \quad n \ge \left(\frac{z^* \sigma}{E}\right)^2.$$

Since n must be a whole number and a larger n only makes the interval tighter, the smallest suitable sample size is this bound **rounded up**. (The sample size is chosen *before* the data are collected, so this planning calculation needs a value of σ — known, or estimated from past data or a pilot sample.)

Conditions, and the case of unknown σ.

The interval $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ needs the population standard deviation σ, and needs $\bar{X}$ to be normal. The second requirement holds **exactly** when the population is normal, and **approximately** when the sample is **large**, whatever the shape of the population (the central limit theorem) — so the interval may be used in either case.

In practice σ is usually **unknown**. When the sample is large — $n \ge 30$ in many practical contexts — the **sample standard deviation** s is a reliable estimate of σ, and replacing σ by s gives the **approximate** $C\%$ **confidence interval** for μ

$$\bar{x} \pm z^* \frac{s}{\sqrt{n}} \text{ that is} \left(\bar{x} - z^* \frac{s}{\sqrt{n}}, \bar{x} + z^* \frac{s}{\sqrt{n}}\right).$$

This is the form printed on the VCAA formula sheet. Nothing else changes: the multipliers z^*, the margin of error $E = z^* \frac{s}{\sqrt{n}}$, the interpretation and both levers above all carry across — only the standard error changes from $\frac{\sigma}{\sqrt{n}}$ to $\frac{s}{\sqrt{n}}$. Because s varies from sample to sample, the stated confidence level is only approximate, which is why a large n is required.

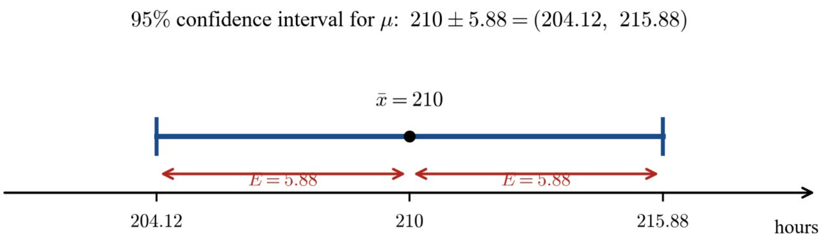
So read the data carefully: use σ when a **population** standard deviation is given, and s when a **sample** standard deviation is given.

Worked examples

Example 1 — scaffolded

A manufacturer tests rechargeable batteries whose lifetime (in hours) is normally distributed with **known** population standard deviation $\sigma = 15$ hours. A random sample of $n = 25$ batteries has mean lifetime $\bar{x} = 210$ hours. Construct a 95% confidence interval for the mean lifetime μ.

Working	Comment
Given: $\sigma = 15$ (known), $n = 25$, $\bar{x} = 210$.	The population sd is known, so a z-interval applies.
Standard error $\frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{25}} = \frac{15}{5} = 3$.	$sd(\bar{X}) = \sigma / \sqrt{n}$.
95% confidence $\Rightarrow z^* = 1.96$.	$z^* = \Phi^{-1}(0.975)$.
Margin of error $E = z^* \frac{\sigma}{\sqrt{n}} = 1.96 \times 3 = 5.88$.	$E = z^* \times$ standard error.
CI: $\bar{x} \pm E = 210 \pm 5.88 = (204.12, 215.88)$.	From $\bar{x} - E$ to $\bar{x} + E$.



We are $95\,\%$ confident that the mean battery lifetime μ lies between 204.12 and 215.88 hours.

Example 2 — scaffolded

A machine fills cereal boxes. The population standard deviation of the net mass (in grams) is **not** known. A random sample of $n = 64$ boxes has mean net mass $\bar{x} = 502$ g and sample standard deviation $s = 8$ g. Find the margin of error and an approximate $99\,\%$ confidence interval for the mean net mass μ.

Working	Comment
Given: $n = 64$, $\bar{x} = 502$, $s = 8$, with σ unknown.	$n = 64 \geq 30$ is large, so s may replace σ and $\bar{X}$ is approximately normal.
Standard error $\frac{s}{\sqrt{n}} = \frac{8}{\sqrt{64}} = \frac{8}{8} = 1$.	Estimated $sd(\bar{X})$.
$99\,\%$ confidence $\Rightarrow z^* = 2.576$.	$z^* = \Phi^{-1}(0.995)$.
$E = z^* \frac{s}{\sqrt{n}} = 2.576 \times 1 = 2.576 \approx 2.58$.	Margin of error.
CI: $502 \pm 2.576 = (499.42, 504.58)$.	Endpoints rounded to 2 dp.

A $95\,\%$ interval would instead use $z^* = 1.96$, giving the smaller margin $E = 1.96$ g and the **narrower** interval $502 \pm 1.96 = (500.04, 503.96)$ g — less confidence, but a tighter estimate.

Example 3 — unscaffolded

The weekly self-study time (in hours) of students at a college is modelled as normal with known standard deviation $\sigma = 6$ hours. A random sample of $n = 36$ students has mean $\bar{x} = 22$ hours. (a) Construct a $95\,\%$ confidence interval for the mean weekly study time μ, and interpret it. (b) A fresh sample of $n = 144$ students (same σ) is proposed. How would the margin of error change?

For part (a), the standard error is $\frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{36}} = \frac{6}{6} = 1$, and $95\,\%$ confidence gives $z^* = 1.96$, so the margin of error is $E = 1.96 \times 1 = 1.96$. The interval is

$$\bar{x} \pm E = 22 \pm 1.96 = (20.04, 23.96).$$

We are $95\,\%$ confident that the mean weekly study time μ lies between 20.04 and 23.96 hours. The interval is random — a different sample would produce a different interval — while μ is a fixed unknown; about $95\,\%$ of intervals built this way contain μ. It would be wrong to say $P(20.04 < \mu < 23.96) = 0.95$, since this particular interval either contains μ or it does not.

For part (b), quadrupling the sample to $n = 144$ gives standard error $\frac{6}{\sqrt{144}} = \frac{6}{12} = 0.5$, so $E = 1.96 \times 0.5 = 0.98$ and the interval becomes $22 \pm 0.98 = (21.02, 22.98)$.

$\therefore$ multiplying the sample size by 4 **halves** the margin of error (from 1.96 to 0.98) and so halves the width of the interval, because $E \propto \frac{1}{\sqrt{n}}$.

Example 4 — unscaffolded

The length (in mm) of a machined component is normally distributed with known standard deviation $\sigma = 4$ mm. How large must a random sample be so that a 95% confidence interval for the mean length has a margin of error of at most 0.5 mm?

The requirement is $E = z^* \frac{\sigma}{\sqrt{n}} \le 0.5$ with $z^* = 1.96$. Rearranging for n,

$$\sqrt{n} \ge \frac{z^* \sigma}{E} = \frac{1.96 \times 4}{0.5} = \frac{7.84}{0.5} = 15.68, n \ge 15.68^2 = 245.8624.$$

Because n must be a whole number and a larger sample only tightens the interval, round up to $n = 246$. (Checking: $n = 246$ gives $E = \frac{1.96 \times 4}{\sqrt{246}} \approx 0.4999 \le 0.5$, while $n = 245$ gives $E \approx 0.5009 > 0.5$.)

$\therefore$ at least 246 components must be sampled.

Practice questions

Scaffolded

Q1. Exam marks are normally distributed with known standard deviation $\sigma = 20$. A random sample of $n = 100$ scripts has mean $\acute{x} = 512$. Construct a 95% confidence interval for the mean mark μ. (a) Find the standard error $\frac{\sigma}{\sqrt{n}}$. (b) State z^* for 95% confidence. (c) Find the margin of error E. (d) Write down the confidence interval $\acute{x} \pm E$.

Q2. The population standard deviation of a quantity is unknown. A random sample of $n = 36$ measurements has mean $\acute{x} = 68$ and sample standard deviation $s = 12$. (a) Find the standard error $\frac{s}{\sqrt{n}}$, and state why s may be used here. (b) State z^* for 90% confidence. (c) Find the margin of error. (d) Write down the approximate 90% confidence interval for μ.

Q3. A normal population has known standard deviation $\sigma = 10$. A random sample of $n = 25$ has mean $\acute{x} = 150$. (a) Find the standard error. (b) State z^* for 99% confidence. (c) Find the margin of error. (d) Write down the 99% confidence interval for μ.

Q4. A normal population has known standard deviation $\sigma = 6$. A 95% confidence interval for μ is required with a margin of error of at most 1. (a) Write the inequality $z^* \frac{\sigma}{\sqrt{n}} \le 1$ using $z^* = 1.96$. (b) Solve it for $\sqrt{n}$. (c) Solve for n. (d) State the smallest suitable sample size.

Q5. A normal population has known standard deviation $\sigma = 30$. A random sample of $n = 36$ has mean $\acute{x} = 145$. (a) Construct a 95% confidence interval for μ. (b) A report states "there is a 95% probability that μ lies between the two endpoints." Explain why this wording is not quite correct, and give a better statement. (c) A simulation draws 200 independent samples of size $n = 36$ from this population and builds a 95% confidence interval from each. About how many of the 200 intervals would be expected to contain μ? Would that number be larger or smaller if 99% intervals were built instead?

Q6. A normal population has known standard deviation $\sigma = 24$. Samples have mean $\acute{x} = 200$. (a) Construct a 95% confidence interval for μ from a sample of size $n = 36$. (b) Construct a 95% confidence interval from a sample of size $n = 144$. (c) Compare the widths of the two intervals and explain the difference.

Unscaffolded

Q7. A normal population has known standard deviation $\sigma = 25$. A random sample of $n = 100$ has mean $\bar{x} = 340$. Construct a 95% confidence interval for μ.

Q8. A normal population has known standard deviation $\sigma = 6$. A random sample of $n = 36$ has mean $\bar{x} = 48$. Construct a 99% confidence interval for μ.

Q9. A normal population has known standard deviation $\sigma = 14$. A random sample of $n = 49$ has mean $\bar{x} = 205$. Construct a 90% confidence interval for μ.

Q10. The population standard deviation of a quantity is unknown. A random sample of $n = 81$ items has mean $\bar{x} = 60$ and sample standard deviation $s = 18$. Find the margin of error and an approximate 95% confidence interval for μ.

Q11. A normal population has known standard deviation $\sigma = 5$. Find the smallest sample size n for which a 99% confidence interval for μ has a margin of error of at most 1.

Q12. A normal population has known standard deviation $\sigma = 8$. Find the smallest sample size n for which a 90% confidence interval for μ has a margin of error of at most 2.

Q13. From a sample, a 95% confidence interval for a population mean μ is computed as $(12.3, 15.7)$. (a) A student writes $P(12.3 < \mu < 15.7) = 0.95$. Explain why this statement is not correct, and give the correct interpretation of the interval. (b) In a simulation, 500 samples of the same size are drawn from the same population and a 95% confidence interval is built from each. About how many of the 500 intervals would be expected to **miss** μ? How would that number change if 90% intervals were built instead?

Q14. A normal population has known standard deviation $\sigma = 16$. A random sample of $n = 64$ has mean $\bar{x} = 250$. (a) Construct a 95% confidence interval for μ. (b) Construct a 99% confidence interval for μ. (c) Which interval is wider, and why?

Q15. A 95% confidence interval for a population mean μ, computed from a random sample drawn from a population with known standard deviation $\sigma = 16$, is reported as $(46.08, 53.92)$. (a) Find the sample mean $\bar{x}$. (b) Find the margin of error E. (c) Find the sample size n that was used.

Q16. A 95% confidence interval for a population mean, from a random sample of $n = 40$, has a margin of error of 5. The population standard deviation and the confidence level are unchanged. (a) By what factor must the sample size increase to halve the margin of error? (b) What sample size would give a margin of error of 2.5?

Q17. The distribution of waiting times (in minutes) at a clinic is unknown, as is its population standard deviation. A random sample of $n = 50$ patients has mean waiting time $\bar{x} = 27.4$ minutes and sample standard deviation $s = 4.5$ minutes. Construct an approximate 95% confidence interval for the mean waiting time μ.

Q18. Bags of sugar have net mass (in grams) that is normally distributed with known standard deviation $\sigma = 12$ g. A random sample of $n = 64$ bags has mean net mass $\bar{x} = 155$ g. (a) Construct a 95% confidence interval for the mean net mass μ. (b) Interpret this interval in context. (c) Find the smallest sample size for which a 95% confidence interval would have a margin of error of at most 2.5 g.

Answers

Q1. se $=2$; $z^*=1.96$; $E=3.92$; 95% CI $=(508.08, 515.92)$. **Q2.** se $=2$ ($n=36\geq 30$, so s estimates σ); $z^*=1.645$; $E=3.29$; approximate 90% CI $=(64.71, 71.29)$. **Q3.** se $=2$; $z^*=2.576$; $E=5.15$; 99% CI $=(144.85, 155.15)$. **Q4.** $\frac{1.96\times 6}{\sqrt{n}}\leq 1$; $\sqrt{n}\geq 11.76$; $n\geq 138.30$; smallest $n=139$. **Q5.** (a) $(135.2, 154.8)$; (b) μ is fixed and the computed interval is fixed, so no probability applies to it; correctly, "we are 95% confident the interval contains μ"; (c) about 190 of the 200; larger at 99% (about 198), since wider intervals capture μ more often. **Q6.** (a) $(192.16, 207.84)$, width 15.68; (b) $(196.08, 203.92)$, width 7.84; (c) the $n=144$ interval is half as wide, since $E\propto 1/\sqrt{n}$ and n was quadrupled. **Q7.** $E=4.9$; 95% CI $=(335.1, 344.9)$. **Q8.** $E=2.58$; 99% CI $=(45.42, 50.58)$. **Q9.** $E=3.29$; 90% CI $=(201.71, 208.29)$. **Q10.** $E=3.92$; approximate 95% CI $=(56.08, 63.92)$. **Q11.** $n=166$. **Q12.** $n=44$. **Q13.** (a) Wrong because μ is a fixed constant and the interval $(12.3, 15.7)$ is already computed, so it either contains μ or not; correctly, "we are 95% confident that $(12.3, 15.7)$ contains μ," meaning 95% of intervals formed this way contain μ; (b) about 25 of the 500 would miss μ; at 90% about 50 would miss — twice as many, because the intervals are narrower. **Q14.** (a) $(246.08, 253.92)$; (b) $(244.85, 255.15)$; (c) the 99% interval is wider because $z^*=2.576>1.96$. **Q15.** (a) $\acute{x}=50$; (b) $E=3.92$; (c) $n=64$. **Q16.** (a) a factor of 4; (b) $n=160$. **Q17.** $E=1.25$; approximate 95% CI $=(26.15, 28.65)$. **Q18.** (a) $(152.06, 157.94)$; (b) we are 95% confident the mean net mass μ is between 152.06 g and 157.94 g; (c) $n=89$.

Fully-worked solutions

Q1. The standard error is $\frac{\sigma}{\sqrt{n}}=\frac{20}{\sqrt{100}}=\frac{20}{10}=2$. For 95% confidence, $z^*=1.96$, so the margin of error is $E=1.96\times 2=3.92$. The interval is $\acute{x}\pm E=512\pm 3.92=(508.08, 515.92)$.

Q2. Since σ is unknown and the sample is large ($n=36\geq 30$), the sample standard deviation $s=12$ is used in its place, and $\acute{X}$ is approximately normal. The standard error is $\frac{s}{\sqrt{n}}=\frac{12}{\sqrt{36}}=\frac{12}{6}=2$. For 90% confidence, $z^*=1.645$, so $E=1.645\times 2=3.29$. The approximate interval is $68\pm 3.29=(64.71, 71.29)$.

Q3. The standard error is $\frac{10}{\sqrt{25}}=\frac{10}{5}=2$. For 99% confidence, $z^*=2.576$, so $E=2.576\times 2=5.152\approx 5.15$. The interval is $150\pm 5.152=(144.85, 155.15)$.

Q4. Requiring the margin at most 1: $\frac{1.96\times 6}{\sqrt{n}}\leq 1$, so $\sqrt{n}\geq 1.96\times 6=11.76$, hence $n\geq 11.76^2=138.2976$. Rounding up, the smallest sample size is $n=139$. (Then $E=\frac{1.96\times 6}{\sqrt{139}}\approx 0.997\leq 1$, while $n=138$ gives $E\approx 1.001>1$.)

Q5. (a) The standard error is $\frac{30}{\sqrt{36}}=\frac{30}{6}=5$, and $z^*=1.96$, so $E=1.96\times 5=9.8$ and the interval is $145\pm 9.8=(135.2, 154.8)$. (b) The wording treats μ as if it were random. In fact μ is a fixed unknown constant and the interval $(135.2, 154.8)$ has already been computed, so it either contains μ or it does not — no probability attaches to this single interval. A correct statement is: "we are 95% confident that the interval $(135.2, 154.8)$ contains μ," meaning that 95% of intervals constructed by this method contain μ. (c) Before its sample is drawn, each interval contains μ with probability 0.95, so of 200 independent intervals about $0.95\times 200=190$ would be expected to contain μ (about 10 missing it). At 99% confidence every interval is wider, so more of them capture μ: the expected number would be larger, about $0.99\times 200=198$.

Q6. (a) For $n=36$, the standard error is $\frac{24}{\sqrt{36}}=\frac{24}{6}=4$, so $E=1.96\times4=7.84$ and the interval is $200\pm7.84=(192.16,207.84)$, of width 15.68. (b) For $n=144$, the standard error is $\frac{24}{\sqrt{144}}=\frac{24}{12}=2$, so $E=1.96\times2=3.92$ and the interval is $200\pm3.92=(196.08,203.92)$, of width 7.84. (c) The second interval is exactly half as wide. Multiplying the sample size by 4 divides the standard error, and hence the margin of error and the width, by $\sqrt{4}=2$, since $E\propto\frac{1}{\sqrt{n}}$.

Q7. The standard error is $\frac{25}{\sqrt{100}}=\frac{25}{10}=2.5$, and $z^*=1.96$, so $E=1.96\times2.5=4.9$. The 95% interval is $340\pm4.9=(335.1,344.9)$.

Q8. The standard error is $\frac{6}{\sqrt{36}}=\frac{6}{6}=1$, and for 99% confidence $z^*=2.576$, so $E=2.576\times1=2.576\approx2.58$. The interval is $48\pm2.576=(45.42,50.58)$.

Q9. The standard error is $\frac{14}{\sqrt{49}}=\frac{14}{7}=2$, and for 90% confidence $z^*=1.645$, so $E=1.645\times2=3.29$. The interval is $205\pm3.29=(201.71,208.29)$.

Q10. With σ unknown and $n=81\geq30$ large, the sample standard deviation $s=18$ replaces σ, so the standard error is $\frac{s}{\sqrt{n}}=\frac{18}{\sqrt{81}}=\frac{18}{9}=2$, and $z^*=1.96$, so the margin of error is $E=1.96\times2=3.92$. The approximate 95% interval is $60\pm3.92=(56.08,63.92)$.

Q11. Requiring the margin at most 1 with $z^*=2.576$: $\frac{2.576\times5}{\sqrt{n}}\leq1$, so $\sqrt{n}\geq2.576\times5=12.88$, hence $n\geq12.88^2=165.8944$. Rounding up, $n=166$. (Then $E\approx0.9997\leq1$, while $n=165$ gives $E\approx1.003>1$.)

Q12. Requiring the margin at most 2 with $z^*=1.645$: $\frac{1.645\times8}{\sqrt{n}}\leq2$, so $\sqrt{n}\geq\frac{1.645\times8}{2}=6.58$, hence $n\geq6.58^2=43.2964$. Rounding up, $n=44$. (Then $E\approx1.984\leq2$, while $n=43$ gives $E\approx2.007>2$.)

Q13. (a) The statement $P(12.3<\mu<15.7)=0.95$ is not correct because μ is a fixed (unknown) constant, and once the interval $(12.3,15.7)$ has been computed from the data it is also fixed. A fixed interval either does or does not contain μ, so no probability of 0.95 can be assigned to this one interval. The correct interpretation is: “we are 95% confident that the interval $(12.3,15.7)$ contains μ” — that is, if many samples were taken and an interval constructed from each in the same way, about 95% of those intervals would contain μ. (b) Each 95% interval fails to contain μ with probability $1-0.95=0.05$, so of 500 such intervals about $0.05\times500=25$ would be expected to miss μ. At 90% confidence each interval is narrower and misses with probability 0.10, so about $0.10\times500=50$ would miss — twice as many. The chance that an interval contains μ depends on the confidence level chosen.

Q14. The standard error is $\frac{16}{\sqrt{64}}=\frac{16}{8}=2$ in both parts. (a) For 95%, $z^*=1.96$, so $E=1.96\times2=3.92$ and the interval is $250\pm3.92=(246.08,253.92)$. (b) For 99%, $z^*=2.576$, so $E=2.576\times2=5.152\approx5.15$ and the interval is $250\pm5.152=(244.85,255.15)$. (c) The 99% interval is wider, because the same data are multiplied by the larger critical value $z^*=2.576$ instead of 1.96; greater confidence requires a wider interval.

Q15. (a) The sample mean is the midpoint of the interval, $\acute{x}=\frac{46.08+53.92}{2}=\frac{100}{2}=50$. (b) The margin of error is the half-width, $E=\frac{53.92-46.08}{2}=\frac{7.84}{2}=3.92$. (c) From $E=z^*\frac{\sigma}{\sqrt{n}}$ with $z^*=1.96$ and $\sigma=16$,

$$\sqrt{n} = \frac{z^* \sigma}{E} = \frac{1.96 \times 16}{3.92} = \frac{31.36}{3.92} = 8, n = 8^2 = 64.$$

Q16. (a) Since $E = z^* \frac{\sigma}{\sqrt{n}} \propto \frac{1}{\sqrt{n}}$, halving E requires $\sqrt{n}$ to double, and so n to increase by a factor of $2^2 = 4$. (b) A margin of 2.5 is half of 5, so the sample size must be 4 times as large: $n = 4 \times 40 = 160$.

Q17. The sample is large ($n = 50 \geq 30$), so $\acute{X}$ is approximately normal by the central limit theorem even though the population shape is not given, and the sample standard deviation $s = 4.5$ may be used in place of the unknown σ. The standard error is $\frac{s}{\sqrt{n}} = \frac{4.5}{\sqrt{50}} \approx 0.6364$, and $z^* = 1.96$, so the margin of error is $E = 1.96 \times 0.6364 \approx 1.25$. The approximate 95% interval is $27.4 \pm 1.25 = (26.15, 28.65)$ minutes.

Q18. (a) The standard error is $\frac{12}{\sqrt{64}} = \frac{12}{8} = 1.5$, and $z^* = 1.96$, so $E = 1.96 \times 1.5 = 2.94$ and the interval is $155 \pm 2.94 = (152.06, 157.94)$. (b) We are 95% confident that the mean net mass μ of all such bags lies between 152.06 g and 157.94 g; the confidence refers to the reliability of the method, not to a probability about this single interval. (c) Requiring the margin at most 2.5: $\frac{1.96 \times 12}{\sqrt{n}} \leq 2.5$, so $\sqrt{n} \geq \frac{1.96 \times 12}{2.5} = 9.408$, hence $n \geq 9.408^2 = 88.510464$. Rounding up, the smallest sample size is $n = 89$.

Chapter 12 Confidence intervals and hypothesis testing for the mean — 12B Hypothesis testing for the mean

Theory

A **confidence interval**, developed in an earlier section, uses a sample to *estimate* an unknown population mean μ. A **hypothesis test** (also called a **significance test**) answers a different kind of question: it *judges a specific claim* about μ. Someone asserts that the mean is a particular value μ_0; a random sample gives a sample mean $\bar{x}$ that differs from μ_0; and we must decide whether that difference is large enough to be real, or small enough to be explained by ordinary sampling variation.

Throughout this section we assume the sample mean is normally distributed: either the population is **normal with known standard deviation** σ, or the sample is **large**, so that the central limit theorem makes $\bar{X}$ approximately normal. Either way, if the true mean is μ_0 then the sample mean is a random variable with

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right), sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

The quantity $\frac{\sigma}{\sqrt{n}}$ is the **standard error** of the sample mean — the typical distance a sample mean falls from the true mean.

Null and alternative hypotheses.

Every test is a contest between two competing statements about μ.

- The **null hypothesis** $H_0 : \mu = \mu_0$ is the claim on trial — the "no change" or "status quo" value μ_0. All calculations are made *assuming H_0 is true*.
- The **alternative hypothesis** H_1 is the departure from μ_0 that we are testing for.

In this introduction the alternative is always **one-sided** (a single tail): the context tells us in advance which direction matters, so H_1 is either

$$H_1 : \mu > \mu_0 \text{ or } H_1 : \mu < \mu_0.$$

For instance, "has the machine started to *over*-fill?" points to $H_1 : \mu > \mu_0$, while "is the battery lasting *less* than claimed?" points to $H_1 : \mu < \mu_0$. Tests whose alternative allows a departure in **either** direction (two-tail tests) are treated in a later section.

The test statistic.

To measure how far the observed $\bar{x}$ falls from μ_0, we convert it to a standard score using the sampling distribution above. Assuming H_0, the **test statistic** is

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}.$$

This z records **how many standard errors** the observed sample mean lies from the claimed mean. A value close to 0 means $\bar{x}$ is about where H_0 predicts; a large $|z|$ means $\bar{x}$ is far out in the tail, which is hard to explain if H_0 really holds. Because z is a standardised normal score, we compare it against the standard normal distribution $Z \sim N(0, 1)$.

Level of significance.

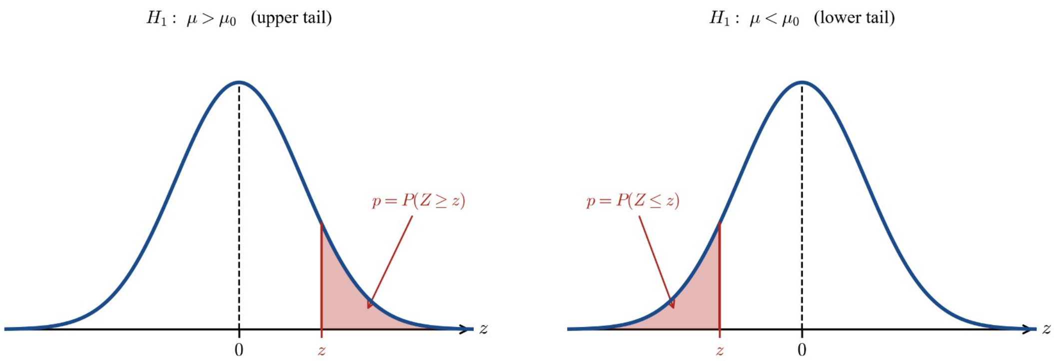
Before looking at the data we fix a **level of significance** α — the threshold that decides how unlikely an outcome must be before we treat it as evidence against H_0. Common choices are $\alpha = 0.05$ (5%) and $\alpha = 0.01$ (1%). Formally α

is the probability of rejecting H_0 when it is in fact true; that risk, and the errors of a test, are examined in a later section. Here α serves as the fixed cut-off for our decision.

The p-value.

The heart of the test is the **p-value**: assuming H_0 is true, it is the probability of obtaining a test statistic **at least as extreme** as the one observed, in the direction of H_1. For a one-sided test it is a single tail area under the standard normal curve, cut off at the observed z:

alternative H_1	direction	p-value
$\mu > \mu_0$	upper tail	$P(Z \geq z)$
$\mu < \mu_0$	lower tail	$P(Z \leq z)$



A **small** p-value means that, if H_0 were true, data as extreme as ours would be very unlikely — so the data cast doubt on H_0. A **large** p-value means our sample is quite consistent with H_0. The p-value is therefore a measure of how surprising the evidence is under the null hypothesis: the smaller it is, the stronger the evidence against H_0.

The decision rule.

Comparing the p-value with the significance level gives a clear-cut decision:

$$\text{reject } H_0 \text{ if } p\text{-value} < \alpha \text{; otherwise do not reject } H_0.$$

Rejecting H_0 means the sample provides significant evidence for the alternative H_1. Failing to reject H_0 does **not** prove that $\mu = \mu_0$; it means only that the sample does not provide enough evidence to overturn the claim at the chosen level α.

Stating the conclusion in context.

A test is not finished until the decision is translated back into the language of the problem. “Reject H_0” becomes, for example, “there is significant evidence at the 5% level that the mean fill has increased”; “do not reject H_0” becomes “there is insufficient evidence at the 5% level that the mean fill has increased”. Because the p-value is compared with α, the same data can lead to different conclusions at different significance levels — a borderline result may be significant at 5% but not at 1%.

Summary — the steps of a one-tail test for a mean.

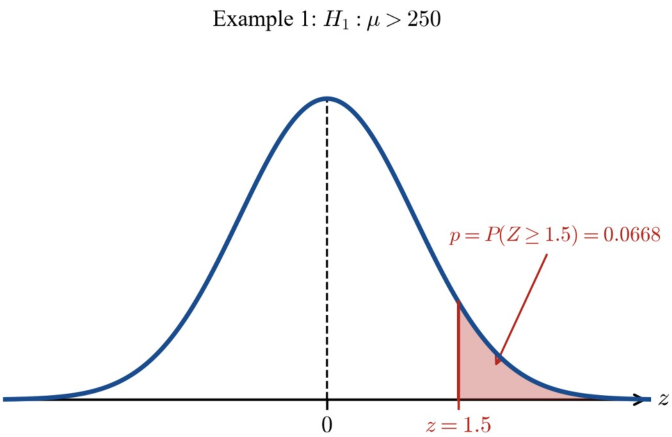
1. State the hypotheses $H_0 : \mu = \mu_0$ and the one-sided H_1 ($\mu > \mu_0$ or $\mu < \mu_0$), and the significance level α.
2. Assuming H_0, model the sample mean as $\acute{X} \sim N\left(\mu_0, \sigma^2 / n\right)$ and compute the standard error $\dfrac{\sigma}{\sqrt{n}}$.
3. Evaluate the test statistic $z = \dfrac{\acute{x} - \mu_0}{\sigma / \sqrt{n}}$.
4. Find the p-value as the tail area beyond z in the direction of H_1.
5. Compare with α: reject H_0 if the p-value $< \alpha$, then state the conclusion in context.

Worked examples

Example 1 — scaffolded

A drinks machine dispenses volumes that are normally distributed with known standard deviation $\sigma=4$ mL, and is set to pour a mean of 250 mL. A technician suspects the machine has begun to **over-fill**. A random sample of $n=16$ cups has mean $\acute{x}=251.5$ mL. Test at the $\alpha=0.05$ level whether the mean volume has increased.

Working	Comment
$H_0: \mu=250$ versus $H_1: \mu>250$.	"Over-fill" is one-sided, so an upper-tail alternative.
Standard error $\frac{\sigma}{\sqrt{n}}=\frac{4}{\sqrt{16}}=\frac{4}{4}=1$.	Assuming H_0, $\acute{X}\sim N(250,1^2)$.
$z=\frac{\acute{x}-\mu_0}{\sigma/\sqrt{n}}=\frac{251.5-250}{1}=1.5$.	$\acute{x}$ lies 1.5 standard errors above μ_0.
p-value $=P(Z\geq 1.5)=1-\Phi(1.5)=1-0.9332=0.066$.	Upper-tail area beyond the observed z.
Since $0.0668>0.05=\alpha$, do not reject H_0.	The p-value is not below α.
There is insufficient evidence at the 5% level that the machine over-fills.	Conclusion stated in context.

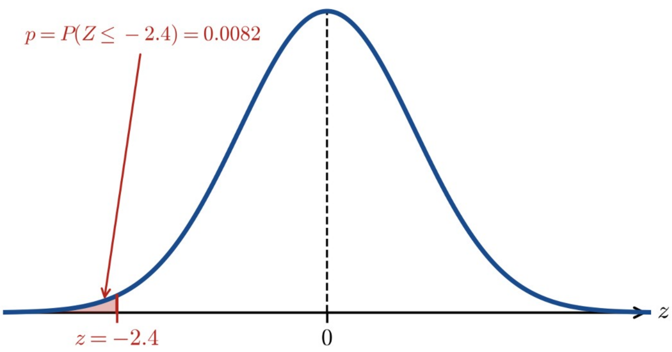


Example 2 — scaffolded

A manufacturer claims that its batteries have a mean lifetime of 500 hours; lifetimes are normally distributed with known standard deviation $\sigma=30$ hours. A consumer group suspects the true mean is **lower** than claimed. A random sample of $n=36$ batteries has mean lifetime $\acute{x}=488$ hours. Test the claim at the $\alpha=0.05$ level.

Working	Comment
$H_0: \mu=500$ versus $H_1: \mu<500$.	"Lower than claimed" gives a lower-tail alternative.
Standard error $\frac{\sigma}{\sqrt{n}}=\frac{30}{\sqrt{36}}=\frac{30}{6}=5$.	Assuming H_0, $\acute{X}\sim N(500,5^2)$.
$z=\frac{488-500}{5}=\frac{-12}{5}=-2.4$.	$\acute{x}$ lies 2.4 standard errors below μ_0.
p-value $=P(Z\leq -2.4)=\Phi(-2.4)=0.0082$.	Lower-tail area below the observed z.
Since $0.0082<0.05=\alpha$, reject H_0.	The p-value is below α.
There is significant evidence at the 5% level that the mean lifetime is less than 500 hours.	Conclusion stated in context.

Example 2: $H_1 : \mu < 500$



Example 3 — unscaffolded

A crop is grown on plots whose yields are normally distributed with known standard deviation $\sigma = 0.8$ kg, and the long-run mean yield is 12.5 kg per plot. An agronomist believes a new fertiliser **increases** the mean yield. On $n = 25$ plots treated with the new fertiliser the sample mean is $\bar{x} = 12.9$ kg. Test this belief at the $\alpha = 0.01$ level.

Because the alternative is that the mean has increased, set $H_0 : \mu = 12.5$ against $H_1 : \mu > 12.5$. Assuming H_0, the sample mean is $\bar{X} \sim N(12.5, \sigma^2 / n)$ with standard error

$$\frac{\sigma}{\sqrt{n}} = \frac{0.8}{\sqrt{25}} = \frac{0.8}{5} = 0.16.$$

The test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{12.9 - 12.5}{0.16} = \frac{0.4}{0.16} = 2.5,$$

so the observed mean is 2.5 standard errors above 12.5. As H_1 is an upper-tail alternative, the p-value is

$$P(Z \ge 2.5) = 1 - \Phi(2.5) = 1 - 0.9938 = 0.0062.$$

Since $0.0062 < 0.01 = \alpha$, we reject H_0.

$\therefore$ there is significant evidence at the 1% level that the new fertiliser increases the mean yield.

Example 4 — unscaffolded

Calls at a help desk have handling times that are normally distributed with known standard deviation $\sigma = 1.2$ minutes, and the current mean handling time is 8 minutes. After a staff-training program, a manager believes the mean has **decreased**. A random sample of $n = 9$ calls has mean handling time $\bar{x} = 7.4$ minutes. Test the manager's belief at the $\alpha = 0.05$ level.

The alternative is that the mean has fallen, so $H_0 : \mu = 8$ against $H_1 : \mu < 8$. Assuming H_0, $\bar{X} \sim N(8, \sigma^2 / n)$ with standard error

$$\frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{9}} = \frac{1.2}{3} = 0.4.$$

The test statistic is

$$z = \frac{7.4 - 8}{0.4} = \frac{-0.6}{0.4} = -1.5,$$

so $\bar{x}$ is 1.5 standard errors below 8. As H_1 is a lower-tail alternative, the p-value is

$$P(Z \le -1.5) = \Phi(-1.5) = 0.0668.$$

Since $0.0668 > 0.05 = \alpha$, we do not reject H_0.

$\therefore$ there is insufficient evidence at the 5% level that the training has reduced the mean handling time. (A p-value of 0.0668 shows the data lean slightly towards H_1, but not strongly enough to reject H_0 at the 5% level.)

Practice questions

Scaffolded

Q1. Cereal boxes have filled masses that are normally distributed with known standard deviation $\sigma = 5$ g and a target mean of 375 g. A quality inspector suspects the mean has **increased**. A sample of $n = 25$ boxes has mean $\acute{x} = 377$ g. Test at $\alpha = 0.05$. (a) State H_0 and H_1. (b) Find the standard error and the test statistic z. (c) Find the p-value $P(Z \geq z)$. (d) State the decision and the conclusion in context.

Q2. A supplier claims light bulbs last a mean of 1000 hours, with lifetimes normally distributed and known standard deviation $\sigma = 40$ hours. A tester suspects the true mean is **less**. A sample of $n = 64$ bulbs has mean $\acute{x} = 990$ hours. Test at $\alpha = 0.05$. (a) State H_0 and H_1. (b) Find the standard error and the test statistic z. (c) Find the p-value $P(Z \leq z)$. (d) State the decision and the conclusion in context.

Q3. Machined components have masses that are normally distributed with known standard deviation $\sigma = 2.5$ g and target mean 50 g. An engineer tests whether the mean has **increased**. A sample of $n = 25$ components has mean $\acute{x} = 50.7$ g. Test at $\alpha = 0.05$. (a) State H_0 and H_1. (b) Find the standard error and the test statistic z. (c) Find the p-value. (d) State the decision and the conclusion in context.

Q4. A courier's delivery times are normally distributed with known standard deviation $\sigma = 6$ minutes and a stated mean of 30 minutes. A customer suspects deliveries are now **faster** (a lower mean). A sample of $n = 16$ deliveries has mean $\acute{x} = 27$ minutes. Test at the stricter level $\alpha = 0.01$. (a) State H_0 and H_1. (b) Find the standard error and the test statistic z. (c) Find the p-value. (d) State the decision and the conclusion in context.

Q5. A school's mean result on a standard test has long been 65, with results modelled by a known standard deviation $\sigma = 20$. After a new program, a sample of $n = 100$ students scores a mean of $\acute{x} = 68.4$; the principal tests whether the mean has **improved**. Test at $\alpha = 0.05$. (a) Explain why the sample mean may be treated as normal even without assuming a normal population. (b) State H_0 and H_1, and find the standard error and z. (c) Find the p-value. (d) State the decision and the conclusion in context.

Q6. Bolts have diameters that are normally distributed with known standard deviation $\sigma = 0.5$ mm and target mean 20 mm. An inspector suspects the mean diameter has **fallen**. A sample of $n = 9$ bolts has mean $\acute{x} = 19.7$ mm. Test at $\alpha = 0.05$. (a) State H_0 and H_1. (b) Find the standard error and the test statistic z. (c) Find the p-value. (d) State the decision and the conclusion in context.

Unscaffolded

Q7. A chemical process yields a mean of 40 g of product per batch, with yields normally distributed and known standard deviation $\sigma = 6$ g. A new catalyst is believed to **increase** the yield. A sample of $n = 9$ batches has mean $\acute{x} = 45.4$ g. Test at $\alpha = 0.05$, stating the conclusion in context.

Q8. A year level's mean mark on a common assessment has been 50, with a known standard deviation $\sigma = 10$ and marks modelled as normal. A teacher suspects one class scored a **lower** mean. A sample of $n = 25$ students from that class has mean $\acute{x} = 46.8$. Test at $\alpha = 0.05$.

Q9. A website receives a mean of 200 visits per hour, normally distributed with known standard deviation $\sigma = 15$. After an advertising campaign, an analyst tests whether the mean has **increased**. A sample of $n = 25$ hours has mean $\acute{x} = 209$ visits. Test at the 1% level.

Q10. Commuters on a route have a mean travel time of 25 minutes, with known standard deviation $\sigma = 5$ minutes. A new express service is believed to be **faster**. A sample of $n = 100$ commuters has mean $\bar{x} = 23.6$ minutes. Test at $\alpha = 0.01$, and explain why the sample mean may be treated as normal.

Q11. A coating has a mean thickness of 10 micrometres, normally distributed with known standard deviation $\sigma = 2$ micrometres. An operator tests whether a machine now applies a **thicker** coating. A sample of $n = 16$ items has mean $\bar{x} = 10.6$ micrometres. Test at $\alpha = 0.05$.

Q12. A treatment is claimed to leave patients with a mean discomfort score of 80 on a standard scale, with scores modelled as normal with known standard deviation $\sigma = 20$. A researcher tests whether the true mean is **lower**. A sample of $n = 16$ patients has mean $\bar{x} = 69.5$. (a) Carry out the test at $\alpha = 0.05$. (b) Repeat the decision at $\alpha = 0.01$. (c) Comment on how the significance level affected the conclusion.

Q13. Orders at a warehouse take a mean processing time of 50 seconds, normally distributed with known standard deviation $\sigma = 9$ seconds. A supervisor tests whether a change has **increased** the mean time. A sample of $n = 36$ orders has mean $\bar{x} = 52.4$ seconds. Test at $\alpha = 0.05$.

Q14. A production line averages 15 minor defects per batch, with counts modelled as normal with known standard deviation $\sigma = 3$. After a redesign, an engineer tests whether the mean has been **reduced**. A sample of $n = 36$ batches has mean $\bar{x} = 13.9$. Test at $\alpha = 0.05$.

Q15. A data centre uses a mean of 500 kWh of energy per day, with a known standard deviation $\sigma = 25$ kWh. After new equipment is installed, a manager tests whether daily usage has **increased**. A sample of $n = 100$ days has mean $\bar{x} = 506$ kWh. Test at $\alpha = 0.05$, and state the result you rely on to treat $\bar{X}$ as normal.

Q16. A bottling line is set to a mean fill of 20 litres, with fills normally distributed and known standard deviation $\sigma = 1.5$ litres. An auditor tests whether the line is **underfilling** (a lower mean). A sample of $n = 25$ fills has mean $\bar{x} = 19.61$ litres. Test at $\alpha = 0.05$.

Q17. In a right-tailed test of $H_0: \mu = 25$ against $H_1: \mu > 25$, a sample gives a test statistic $z = 2.17$, and hence a p-value of 0.0150. (a) State and interpret the decision at $\alpha = 0.05$. (b) Would the decision be the same at $\alpha = 0.01$? Explain, referring to the decision rule.

Q18. An internet provider's baseline mean download speed is 120 Mbps, with speeds normally distributed and known standard deviation $\sigma = 16$ Mbps. The provider claims a network upgrade has **increased** the mean speed. A sample of $n = 64$ connections has mean $\bar{x} = 124.6$ Mbps. Test the claim at $\alpha = 0.05$. (a) State H_0 and H_1. (b) Find the standard error and the test statistic z. (c) Find the p-value. (d) State the decision and the conclusion in context.

Answers

Q1. $z=2$, p-value $=0.0228<0.05$; reject H_0 — significant evidence the mean mass has increased. **Q2.** $z=-2$, p-value $=0.0228<0.05$; reject H_0 — significant evidence the mean lifetime is below 1000 h. **Q3.** $z=1.4$, p-value $=0.0808>0.05$; do not reject H_0 — insufficient evidence the mean mass has increased. **Q4.** $z=-2$, p-value $=0.0228>0.01$; do not reject H_0 — insufficient evidence at the 1% level that deliveries are faster. **Q5.** $z=1.7$, p-value $=0.0446<0.05$; reject H_0 — significant evidence the mean result has improved. **Q6.** $z=-1.8$, p-value $=0.0359<0.05$; reject H_0 — significant evidence the mean diameter has fallen. **Q7.** $z=2.7$, p-value $=0.0035<0.05$; reject H_0 — significant evidence the catalyst increases the yield. **Q8.** $z=-1.6$, p-value $=0.0548>0.05$; do not reject H_0 — insufficient evidence the class mean is lower. **Q9.** $z=3$, p-value $=0.0013<0.01$; reject H_0 — significant evidence visits have increased. **Q10.** $z=-2.8$, p-value $=0.0026<0.01$; reject H_0 — significant evidence the express service is faster. **Q11.** $z=1.2$, p-value $=0.1151>0.05$; do not reject H_0 — insufficient evidence the coating is thicker. **Q12.** $z=-2.1$, p-value $=0.0179$. (a) $0.0179<0.05$: reject H_0. (b) $0.0179>0.01$: do not reject H_0. (c) The same data are significant at 5% but not at 1%; a smaller α demands stronger evidence. **Q13.** $z=1.6$, p-value $=0.0548>0.05$; do not reject H_0 — insufficient evidence the mean time increased. **Q14.** $z=-2.2$, p-value $=0.0139<0.05$; reject H_0 — significant evidence the mean number of defects has fallen. **Q15.** $z=2.4$, p-value $=0.0082<0.05$; reject H_0 — significant evidence daily usage has increased (central limit theorem, $n=100$). **Q16.** $z=-1.3$, p-value $=0.0968>0.05$; do not reject H_0 — insufficient evidence of underfilling. **Q17.** (a) $0.0150<0.05$: reject H_0; significant evidence at the 5% level that $\mu>25$. (b) No: $0.0150>0.01$, so at $\alpha=0.01$ do not reject H_0 — the evidence is not strong enough at the 1% level. **Q18.** $z=2.3$, p-value $=0.0107<0.05$; reject H_0 — significant evidence the mean download speed has increased.

Fully-worked solutions

Q1. Test $H_0: \mu=375$ against $H_1: \mu>375$ (an increase). The standard error is $\frac{\sigma}{\sqrt{n}}=\frac{5}{\sqrt{25}}=\frac{5}{5}=1$, so assuming H_0, $\bar{X} \sim N(375, 1^2)$. The test statistic is

$$z=\frac{377-375}{1}=2.$$

As H_1 is upper-tailed, the p-value is $P(Z\geq 2)=1-\Phi(2)=1-0.9772=0.0228$. Since $0.0228<0.05=\alpha$, we reject H_0: there is significant evidence at the 5% level that the mean mass has increased above 375 g.

Q2. Test $H_0: \mu=1000$ against $H_1: \mu<1000$ (lower than claimed). The standard error is $\frac{40}{\sqrt{64}}=\frac{40}{8}=5$, so $\bar{X} \sim N(1000, 5^2)$ under H_0. Then

$$z=\frac{990-1000}{5}=\frac{-10}{5}=-2.$$

As H_1 is lower-tailed, the p-value is $P(Z\leq -2)=\Phi(-2)=0.0228$. Since $0.0228<0.05=\alpha$, we reject H_0: there is significant evidence at the 5% level that the mean lifetime is below 1000 hours.

Q3. Test $H_0: \mu=50$ against $H_1: \mu>50$. The standard error is $\frac{2.5}{\sqrt{25}}=\frac{2.5}{5}=0.5$, so $\bar{X} \sim N(50, 0.5^2)$ under H_0. Then

$$z=\frac{50.7-50}{0.5}=\frac{0.7}{0.5}=1.4,$$

and the upper-tail p-value is $P(Z\geq 1.4)=1-\Phi(1.4)=1-0.9192=0.0808$. Since $0.0808>0.05=\alpha$, we do not reject H_0: there is insufficient evidence at the 5% level that the mean mass has increased.

Q4. Test $H_0: \mu = 30$ against $H_1: \mu < 30$ (faster). The standard error is $\frac{6}{\sqrt{16}} = \frac{6}{4} = 1.5$, so $\acute{X} \sim N(30, 1.5^2)$ under H_0. Then

$$z = \frac{27 - 30}{1.5} = \frac{-3}{1.5} = -2,$$

and the lower-tail p-value is $P(Z \le -2) = \Phi(-2) = 0.0228$. Comparing with the stricter level $\alpha = 0.01$: since $0.0228 > 0.01$, we do not reject H_0. There is insufficient evidence at the 1% level that deliveries are faster. (Note that at the 5% level this same result *would* be significant — the conclusion depends on the level chosen.)

Q5. (a) The sample size $n = 100$ is large, so by the central limit theorem the sample mean is approximately normal whatever the population shape. (b) Test $H_0: \mu = 65$ against $H_1: \mu > 65$ (improvement). The standard error is $\frac{20}{\sqrt{100}} = \frac{20}{10} = 2$, so $\acute{X} \approx N(65, 2^2)$, and

$$z = \frac{68.4 - 65}{2} = \frac{3.4}{2} = 1.7.$$

(c) The upper-tail p-value is $P(Z \ge 1.7) = 1 - \Phi(1.7) = 1 - 0.9554 = 0.0446$. (d) Since $0.0446 < 0.05 = \alpha$, we reject H_0: there is significant evidence at the 5% level that the mean result has improved.

Q6. Test $H_0: \mu = 20$ against $H_1: \mu < 20$. The standard error is $\frac{0.5}{\sqrt{9}} = \frac{0.5}{3} = 0.1\overline{6}$, so $\acute{X} \sim N(20, (\sigma / \sqrt{n})^2)$ under H_0. Then

$$z = \frac{19.7 - 20}{0.5/3} = \frac{-0.3}{0.5/3} = \frac{-0.3 \times 3}{0.5} = -1.8,$$

and the lower-tail p-value is $P(Z \le -1.8) = \Phi(-1.8) = 0.0359$. Since $0.0359 < 0.05 = \alpha$, we reject H_0: there is significant evidence at the 5% level that the mean diameter has fallen.

Q7. Test $H_0: \mu = 40$ against $H_1: \mu > 40$. The standard error is $\frac{6}{\sqrt{9}} = \frac{6}{3} = 2$, so $\acute{X} \sim N(40, 2^2)$ under H_0. Then

$$z = \frac{45.4 - 40}{2} = \frac{5.4}{2} = 2.7,$$

and the upper-tail p-value is $P(Z \ge 2.7) = 1 - \Phi(2.7) = 1 - 0.9965 = 0.0035$. Since $0.0035 < 0.05 = \alpha$, we reject H_0: there is significant evidence at the 5% level that the new catalyst increases the yield.

Q8. Test $H_0: \mu = 50$ against $H_1: \mu < 50$. The standard error is $\frac{10}{\sqrt{25}} = \frac{10}{5} = 2$, so $\acute{X} \sim N(50, 2^2)$ under H_0. Then

$$z = \frac{46.8 - 50}{2} = \frac{-3.2}{2} = -1.6,$$

and the lower-tail p-value is $P(Z \le -1.6) = \Phi(-1.6) = 0.0548$. Since $0.0548 > 0.05 = \alpha$, we do not reject H_0: there is insufficient evidence at the 5% level that the class scored a lower mean.

Q9. Test $H_0: \mu = 200$ against $H_1: \mu > 200$. The standard error is $\frac{15}{\sqrt{25}} = \frac{15}{5} = 3$, so $\acute{X} \sim N(200, 3^2)$ under H_0. Then

$$z = \frac{209 - 200}{3} = \frac{9}{3} = 3,$$

and the upper-tail p-value is $P(Z \geq 3) = 1 - \Phi(3) = 1 - 0.9987 = 0.0013$. Since $0.0013 < 0.01 = \alpha$, we reject H_0: there is significant evidence at the 1% level that the mean number of visits has increased.

Q10. Because $n = 100$ is large, the central limit theorem makes $\acute{X}$ approximately normal regardless of the population shape. Test $H_0 : \mu = 25$ against $H_1 : \mu < 25$ (faster). The standard error is $\frac{5}{\sqrt{100}} = \frac{5}{10} = 0.5$, so $\acute{X} \approx N(25, 0.5^2)$, and

$$z = \frac{23.6 - 25}{0.5} = \frac{-1.4}{0.5} = -2.8.$$

The lower-tail p-value is $P(Z \leq -2.8) = \Phi(-2.8) = 0.0026$. Since $0.0026 < 0.01 = \alpha$, we reject H_0: there is significant evidence at the 1% level that the express service is faster.

Q11. Test $H_0 : \mu = 10$ against $H_1 : \mu > 10$. The standard error is $\frac{2}{\sqrt{16}} = \frac{2}{4} = 0.5$, so $\acute{X} \sim N(10, 0.5^2)$ under H_0. Then

$$z = \frac{10.6 - 10}{0.5} = \frac{0.6}{0.5} = 1.2,$$

and the upper-tail p-value is $P(Z \geq 1.2) = 1 - \Phi(1.2) = 1 - 0.8849 = 0.1151$. Since $0.1151 > 0.05 = \alpha$, we do not reject H_0: there is insufficient evidence at the 5% level that the coating is thicker.

Q12. Test $H_0 : \mu = 80$ against $H_1 : \mu < 80$. The standard error is $\frac{20}{\sqrt{16}} = \frac{20}{4} = 5$, so $\acute{X} \sim N(80, 5^2)$ under H_0. Then

$$z = \frac{69.5 - 80}{5} = \frac{-10.5}{5} = -2.1,$$

and the lower-tail p-value is $P(Z \leq -2.1) = \Phi(-2.1) = 0.0179$. (a) At $\alpha = 0.05$: since $0.0179 < 0.05$, reject H_0 — significant evidence the mean score is lower. (b) At $\alpha = 0.01$: since $0.0179 > 0.01$, do not reject H_0. (c) The same data give a significant result at the 5% level but not at the 1% level: choosing a smaller significance level demands stronger evidence (a smaller p-value) before H_0 is rejected.

Q13. Test $H_0 : \mu = 50$ against $H_1 : \mu > 50$. The standard error is $\frac{9}{\sqrt{36}} = \frac{9}{6} = 1.5$, so $\acute{X} \sim N(50, 1.5^2)$ under H_0. Then

$$z = \frac{52.4 - 50}{1.5} = \frac{2.4}{1.5} = 1.6,$$

and the upper-tail p-value is $P(Z \geq 1.6) = 1 - \Phi(1.6) = 1 - 0.9452 = 0.0548$. Since $0.0548 > 0.05 = \alpha$, we do not reject H_0: there is insufficient evidence at the 5% level that the mean processing time has increased.

Q14. Test $H_0 : \mu = 15$ against $H_1 : \mu < 15$ (a reduction). The standard error is $\frac{3}{\sqrt{36}} = \frac{3}{6} = 0.5$, so $\acute{X} \sim N(15, 0.5^2)$ under H_0. Then

$$z = \frac{13.9 - 15}{0.5} = \frac{-1.1}{0.5} = -2.2,$$

and the lower-tail p-value is $P(Z \leq -2.2) = \Phi(-2.2) = 0.0139$. Since $0.0139 < 0.05 = \alpha$, we reject H_0: there is significant evidence at the 5% level that the redesign has reduced the mean number of defects.

Q15. Because $n = 100$ is large, by the central limit theorem $\acute{X}$ is approximately normal. Test $H_0 : \mu = 500$ against $H_1 : \mu > 500$. The standard error is $\frac{25}{\sqrt{100}} = \frac{25}{10} = 2.5$, so $\acute{X} \approx N(500, 2.5^2)$, and

$$z = \frac{506 - 500}{2.5} = \frac{6}{2.5} = 2.4.$$

The upper-tail p-value is $P(Z \geq 2.4) = 1 - \Phi(2.4) = 1 - 0.9918 = 0.0082$. Since $0.0082 < 0.05 = \alpha$, we reject H_0: there is significant evidence at the 5% level that daily energy usage has increased.

Q16. Test $H_0 : \mu = 20$ against $H_1 : \mu < 20$ (underfilling). The standard error is $\frac{1.5}{\sqrt{25}} = \frac{1.5}{5} = 0.3$, so $\acute{X} \sim N(20, 0.3^2)$ under H_0. Then

$$z = \frac{19.61 - 20}{0.3} = \frac{-0.39}{0.3} = -1.3,$$

and the lower-tail p-value is $P(Z \leq -1.3) = \Phi(-1.3) = 0.0968$. Since $0.0968 > 0.05 = \alpha$, we do not reject H_0: there is insufficient evidence at the 5% level that the line is underfilling.

Q17. (a) The decision rule rejects H_0 when the p-value is below α. Here the p-value is 0.0150 and $0.0150 < 0.05$, so we reject H_0: there is significant evidence at the 5% level that $\mu > 25$. In words, a sample mean this far above 25 would occur only about 1.5% of the time if the true mean were 25, which is unlikely enough at the 5% level to conclude the mean exceeds 25. (b) No. At $\alpha = 0.01$ the p-value $0.0150 > 0.01$, so we do not reject H_0: the evidence, while suggestive, is not strong enough at the stricter 1% level. The conclusion of a test can change with the significance level.

Q18. (a) Test $H_0 : \mu = 120$ against $H_1 : \mu > 120$ (the upgrade increases the mean). (b) The standard error is $\frac{16}{\sqrt{64}} = \frac{16}{8} = 2$, so $\acute{X} \sim N(120, 2^2)$ under H_0, and

$$z = \frac{124.6 - 120}{2} = \frac{4.6}{2} = 2.3.$$

(c) The upper-tail p-value is $P(Z \geq 2.3) = 1 - \Phi(2.3) = 1 - 0.9893 = 0.0107$. (d) Since $0.0107 < 0.05 = \alpha$, we reject H_0: there is significant evidence at the 5% level that the network upgrade has increased the mean download speed.

Chapter 12 Confidence intervals and hypothesis testing for the mean — 12C One-tail and two-tail tests

Theory

A **hypothesis test** for a population mean μ begins with a **null hypothesis** $H_0: \mu = \mu_0$, which fixes an assumed value μ_0. Assuming H_0 is true and that the population standard deviation σ is known, a random sample of size n with sample mean $\acute{x}$ produces the **test statistic**

$$z = \frac{\acute{x} - \mu_0}{\sigma / \sqrt{n}},$$

which measures how many standard errors the observed mean lies from μ_0. (The test statistic and the idea of a p-value were introduced in the previous section.) A large value of $|z|$ says the data sit far from μ_0 — evidence against H_0.

As in the previous section, this z may be referred to the standard normal distribution only when the sample mean is normally distributed. That holds in two situations: either the population itself is **normally distributed** with known σ, in which case $\acute{X} \sim N\left(\mu_0, \sigma^2 / n\right)$ exactly whatever the sample size; or the sample is **large** ($n \geq 30$), in which case the central limit theorem makes $\acute{X}$ approximately normal whatever the population shape. One of these two conditions is satisfied by every test in this section, and by every test in the questions that follow.

This section is about the other half of setting up a test: the **alternative hypothesis** H_1, which decides *what kind of departure from* μ_0 *counts as evidence*, and how that choice changes the p-value, the rejection region and the final decision.

Choosing the alternative hypothesis from the context.

The alternative hypothesis states what we suspect is really going on. Its form is read from the wording of the problem — in particular, whether we are looking for a departure in **one** direction only, or in **either** direction.

The context suggests that	Alternative H_1	Type of test
the mean has **increased** / is **greater than** / **more than** μ_0	$\mu > \mu_0$	one-tail (upper)
the mean has **decreased** / is **less than** / **smaller than** μ_0	$\mu < \mu_0$	one-tail (lower)
the mean has **changed** / **differs from** / is **not equal to** μ_0	$\mu \neq \mu_0$	two-tail

A **one-tail test** ($H_1: \mu > \mu_0$ or $H_1: \mu < \mu_0$) is used when only a departure in a stated direction is of interest — for example a claim that a redesign has *increased* battery life, where a decrease would not support the claim. A **two-tail test** ($H_1: \mu \neq \mu_0$) is used when a departure in *either* direction matters — for example checking whether a filling machine has *drifted off* its target, too high or too low. The null hypothesis is always the equality $H_0: \mu = \mu_0$; only H_1 carries the direction.

The p-value for each type of test.

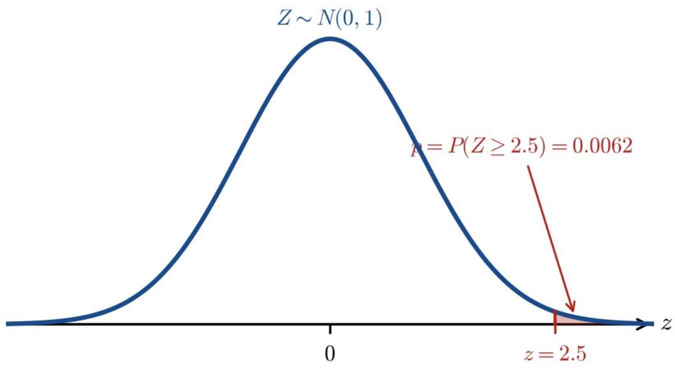
The **p-value** is the probability, computed assuming H_0 is true (so $Z \sim N(0, 1)$), of a test statistic **at least as extreme as** the one observed, *in the direction(s) named by* H_1. This is where the tail count enters.

- **Upper one-tail** $H_1: \mu > \mu_0$: extreme means large positive z, so
$$p = P(Z \geq z) = 1 - \Phi(z).$$
- **Lower one-tail** $H_1: \mu < \mu_0$: extreme means large negative z, so

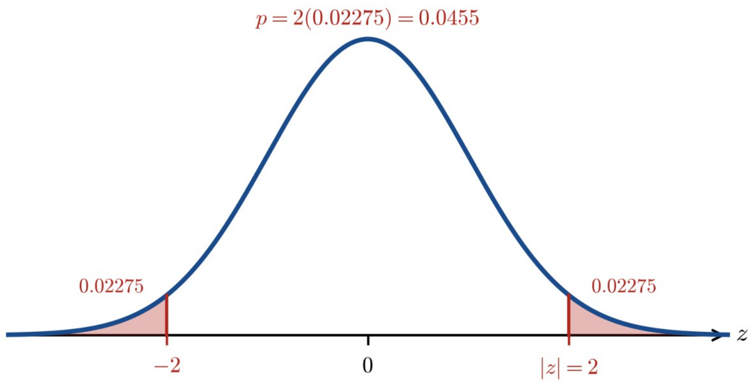
$$p = P(Z \le z) = \Phi(z).$$

- **Two-tail** $H_1 : \mu \ne \mu_0$: extreme means large $|z|$ in **either** direction, so we take the tail beyond $|z|$ and **double** it,
$$p = 2P(Z \ge |z|) = 2(1 - \Phi(|z|)).$$

The rule is the same in every case: **reject** H_0 **when** $p < \alpha$, where α is the chosen **level of significance** (commonly 0.05, 0.01 or 0.10); otherwise do not reject H_0. The diagram below shows a one-tail p-value — the single upper tail beyond $z = 2.5$, area 0.0062.



For a two-tail test the same value of $|z|$ gives **twice** the area, because a result equally far from μ_0 on the other side is also counted as extreme. The picture shows $z = 2$: each tail has area 0.02275, so the two-tail p-value is $2 \times 0.02275 = 0.0455$.



The critical-value approach.

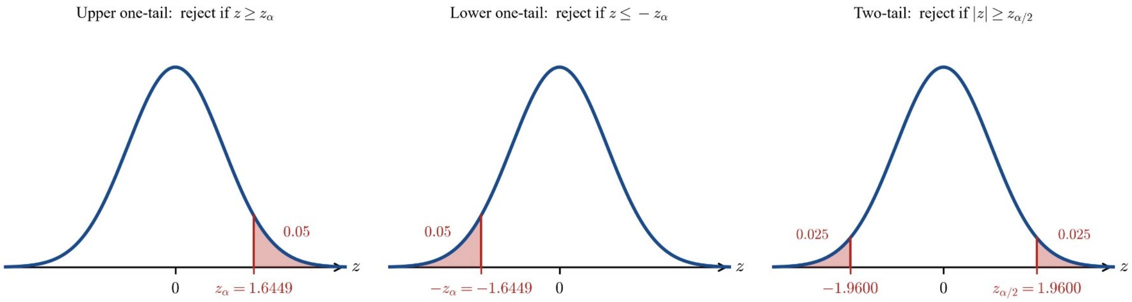
Instead of computing a p-value, we may fix in advance a **rejection region** on the z-scale that carries probability exactly α, and reject H_0 when the test statistic lands in it. The boundary is a **critical value** z^*:

- **Upper one-tail:** reject if $z \ge z_\alpha$, where $z_\alpha = \Phi^{-1}(1-\alpha)$.
- **Lower one-tail:** reject if $z \le -z_\alpha$, with the same $z_\alpha = \Phi^{-1}(1-\alpha)$.
- **Two-tail:** reject if $|z| \ge z_{\alpha/2}$, where $z_{\alpha/2} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$.

For a two-tail test the significance α is **split** into two tails of $\alpha/2$ each, so the relevant quantile is $z_{\alpha/2}$, not z_α. The common critical values are:

α	one-tail $z_\alpha = \Phi^{-1}(1-\alpha)$	two-tail $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$
0.10	1.2816	1.6449
0.05	1.6449	1.9600
0.01	2.3263	2.5758

The three rejection regions, all at $\alpha = 0.05$, are shaded below: the upper one-tail beyond 1.6449, the lower one-tail below -1.6449, and the two-tail split beyond ± 1.9600.



Critical value(s) of $\acute{x}$.

Because $z = \dfrac{\acute{x} - \mu_0}{\sigma / \sqrt{n}}$ is an increasing function of $\acute{x}$, each critical value of z corresponds to a **critical value of the sample mean** $\acute{x}$. Rearranging the boundary $z = z^*$ gives

$$\acute{x}_c = \mu_0 \pm z^* \frac{\sigma}{\sqrt{n}}.$$

For an upper one-tail test we reject if $\acute{x} \geq \mu_0 + z_\alpha \dfrac{\sigma}{\sqrt{n}}$; for a lower one-tail test if $\acute{x} \leq \mu_0 - z_\alpha \dfrac{\sigma}{\sqrt{n}}$; and for a two-tail test if $\acute{x}$ falls **outside** the two boundaries $\mu_0 \pm z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$. This gives the same decision as the z-value or the p-value — it just measures the boundary in the original units. The p-value method and the critical-value method are equivalent: $p < \alpha$ **exactly when z lies in the rejection region.**

A two-tail test is more conservative.

For the same significance α, the two-tail boundary is **further out** than the one-tail boundary: at $\alpha = 0.05$ we need $|z| \geq 1.9600$ rather than $z \geq 1.6449$. Equivalently, the two-tail p-value is **double** the one-tail value, so it crosses α less readily. A test statistic such as $z = 1.8$ therefore **rejects** H_0 as a one-tail test at 5% ($1.8 > 1.6449$) but **fails to reject** as a two-tail test ($1.8 < 1.9600$). Splitting the significance between two tails makes the two-tail test harder to reject — it is the **more conservative** choice for a given α, the price of allowing departures in either direction. This is why the direction of H_1 must be justified from the context *before* the data are seen, never chosen afterwards to manufacture a significant result.

Deciding and interpreting.

Every test ends with a decision expressed in context. If $p < \alpha$ (equivalently z is in the rejection region), we **reject** H_0 and conclude there is evidence, at the α level, for the alternative — "the mean has increased", "the mean differs from μ_0", and so on. If not, we **do not reject** H_0: the data are consistent with $\mu = \mu_0$. Failing to reject H_0 is **not** proof that H_0 is true; it means the sample did not provide enough evidence against it.

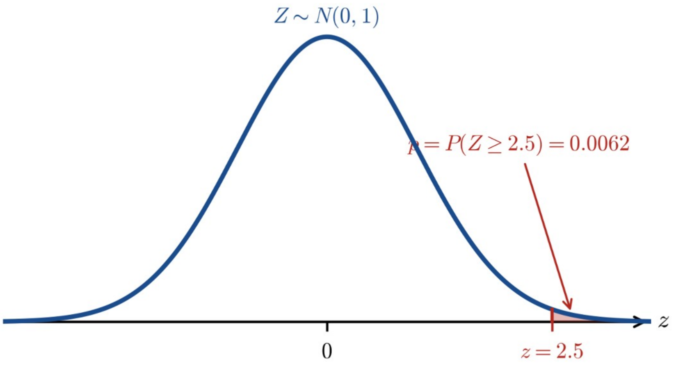
Worked examples

Example 1 — scaffolded

A manufacturer's standard rechargeable batteries last a mean of 8 hours per charge, with a known standard deviation of 1.2 hours. A redesigned battery is claimed to last **longer**. A random sample of $n = 36$ redesigned batteries has mean life $\acute{x} = 8.5$ hours. Test at the 5% level whether the redesign has increased the mean battery life. (a) State H_0 and H_1, and say why the test is one-tailed. (b) Compute the test statistic z. (c) Find the p-value. (d) State the decision and interpret it.

Working	Comment
The claim is that the mean has **increased**, so $H_0: \mu = 8$ and $H_1: \mu > 8$.	A direction ("longer") is named, so this is an upper one-tail test.
$\frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{36}} = \frac{1.2}{6} = 0.2$.	Standard error — the standard deviation of $\bar{X}$.
$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{8.5 - 8}{0.2} = 2.5$.	Test statistic assuming H_0 is true.
$p = P(Z \geq 2.5) = 1 - \Phi(2.5) = 1 - 0.9938 = 0.0062$.	Upper one-tail: the single tail beyond z.
$0.0062 < 0.05$, so **reject** H_0.	Reject H_0 when $p < \alpha$.

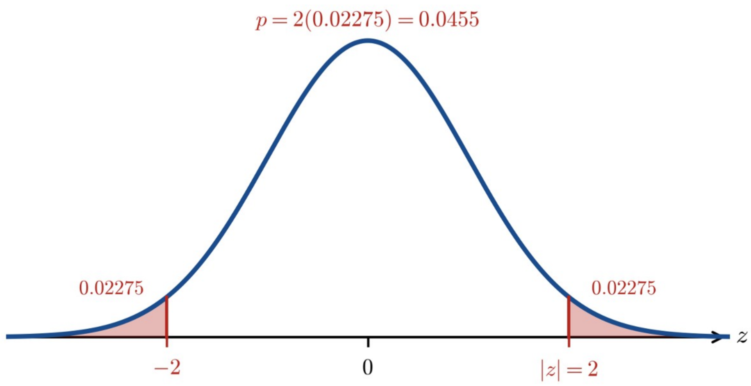
At the 5% level there is evidence that the redesign has increased the mean battery life.



Example 2 — scaffolded

A bottling line is set to fill bottles to a mean of 330 mL, with known standard deviation 6 mL. Quality control wishes to know whether the mean fill has **drifted from** 330 mL in **either** direction. A random sample of $n = 36$ bottles has mean $\bar{x} = 332$ mL. Test at the 5% level whether the mean differs from 330 mL. (a) State H_0 and H_1, and say why the test is two-tailed. (b) Compute z. (c) Find the p-value. (d) State the decision.

Working	Comment
"Differs from" names no direction, so $H_0: \mu = 330$ and $H_1: \mu \neq 330$.	Either an increase or a decrease matters: a two-tail test.
$\frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{36}} = 1$, so $z = \frac{332 - 330}{1} = 2$.	Standard error 1; test statistic $z = 2$.
One tail beyond $\lvert z \rvert = 2$ is $P(Z \geq 2) = 1 - \Phi(2) = 1 - 0.97725 = 0.02275$.	Area in a single tail.
Two-tail $p = 2 \times 0.02275 = 0.0455$.	Double the tail: the alternative counts both directions.
$0.0455 < 0.05$, so **reject** H_0.	Evidence the mean fill has changed from 330 mL.



Example 3 — unscaffolded

A standard treatment gives a mean recovery time of 30 days, with known standard deviation 6 days. A new therapy is trialled on a random sample of $n=36$ patients, whose mean recovery time is $\bar{x}=28$ days. Using the **critical-value method** at the $5\,\%$ level, test whether the new therapy **reduces** the mean recovery time, and state the critical value of $\bar{x}$.

Researchers are looking for a *reduction*, so the test is a lower one-tail test with $H_0: \mu=30$ and $H_1: \mu<30$. The standard error is $\frac{\sigma}{\sqrt{n}}=\frac{6}{\sqrt{36}}=1$, so the test statistic is

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}=\frac{28-30}{1}=-2.$$

At $\alpha=0.05$ the one-tail critical value is $z_\alpha=\Phi^{-1}(1-0.05)=\Phi^{-1}(0.95)=1.6449$, and for a lower-tail test we reject H_0 when $z\le -z_\alpha=-1.6449$. The equivalent critical value of the sample mean is

$$\bar{x}_c=\mu_0-z_\alpha\frac{\sigma}{\sqrt{n}}=30-1.6449\times 1=28.36\,\text{days},$$

so we reject H_0 when $\bar{x}\le 28.36$. Here $z=-2\le -1.6449$ (equivalently $\bar{x}=28\le 28.36$), so the test statistic lies in the rejection region. As a check, the p-value is $P(Z\le -2)=\Phi(-2)=0.0228<0.05$, giving the same decision.

$\therefore$ at the $5\,\%$ level there is evidence that the new therapy reduces the mean recovery time below 30 days.

Example 4 — unscaffolded

A cereal machine fills boxes to masses that are normally distributed with known standard deviation 20 g, and is set to a mean fill of 500 g. A random sample of $n=16$ boxes has mean $\bar{x}=509$ g. Using the **critical-value method** at the $5\,\%$ level, test whether the mean fill **differs** from 500 g, giving the critical values of $\bar{x}$; comment on how the decision would change under a one-tail test.

“Differs” names no direction, so this is a two-tail test with $H_0: \mu=500$ and $H_1: \mu\ne 500$. The standard error is $\frac{\sigma}{\sqrt{n}}=\frac{20}{\sqrt{16}}=5$, so

$$z=\frac{509-500}{5}=1.8.$$

At $\alpha=0.05$ the two-tail critical value is $z_{\alpha/2}=\Phi^{-1}(1-0.025)=\Phi^{-1}(0.975)=1.9600$, and we reject H_0 when $|z|\ge 1.9600$. The equivalent critical values of $\bar{x}$ are

$$\bar{x}_c=\mu_0\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}=500\pm 1.9600\times 5=500\pm 9.80,$$

that is 490.20 g and 509.80 g; we reject only if $\acute{x} \leq 490.20$ or $\acute{x} \geq 509.80$. Here $|z| = 1.8 < 1.9600$ (equivalently $\acute{x} = 509$ lies inside $[490.20, 509.80]$), so we **do not** reject H_0. The two-tail p-value confirms this: $p = 2(1 - \Phi(1.8)) = 0.0719 > 0.05$.

Note the effect of the tail count: had the very same data been analysed as an upper one-tail test ($H_1 : \mu > 500$), the boundary would be only $z_\alpha = 1.6449$, and since $1.8 > 1.6449$ we would have rejected H_0. Splitting α between two tails makes the two-tail test more conservative for the same significance level.

∴ at the 5% level there is insufficient evidence that the mean fill differs from 500 g.

Practice questions

Scaffolded

Q1. A coffee shop's regular espresso shots are pulled to a mean of 200 mL with known standard deviation 4 mL. A barista suspects a new grinder has **increased** the mean volume. A random sample of $n = 64$ shots has mean $\acute{x} = 200.8$ mL. Test at the 5% level. (a) State H_0 and H_1, and the type of test. (b) Find the standard error and the test statistic z. (c) Find the p-value. (d) State and interpret the decision.

Q2. A component is manufactured to lengths that are normally distributed with known standard deviation 5 mm and a mean of 50 mm. An engineer checks whether the mean length now **differs** from 50 mm. A random sample of $n = 25$ components has mean $\acute{x} = 52.2$ mm. Test at the 5% level. (a) State H_0 and H_1, and why the test is two-tailed. (b) Find z. (c) Find the two-tail p-value (the single tail, doubled). (d) State the decision.

Q3. A supplier claims machine parts have a mean mass of 100 g, with masses normally distributed and known standard deviation 8 g. A buyer suspects the true mean is **less** than 100 g and takes a random sample of $n = 16$ parts with mean $\acute{x} = 96.5$ g. Use the **critical-value method** at the 5% level. (a) State H_0 and H_1. (b) Find z. (c) State the critical value $-z_\alpha$ and the rejection rule. (d) State the decision.

Q4. A drink is filled to volumes that are normally distributed with known standard deviation 40 mL and a mean of 750 mL. Quality control tests whether the mean has **changed**. A random sample of $n = 25$ bottles has mean $\acute{x} = 768$ mL. Use the **critical-value method** at the 5% level. (a) State H_0 and H_1. (b) Find z. (c) State the two-tail critical value $z_{\alpha/2}$ and the critical values of $\acute{x}$. (d) State the decision.

Q5. A study reports a mean reaction time of 60 ms with known standard deviation 9 ms. A researcher believes a stimulant **raises** the mean reaction time and samples $n = 36$ trials with mean $\acute{x} = 64.2$ ms. Test at the 1% level. (a) Explain why $H_1 : \mu > 60$ (an upper one-tail test). (b) Find z. (c) Find the p-value. (d) State the decision at $\alpha = 0.01$.

Q6. A packing line targets a mean net weight of 15 kg, with net weights normally distributed and known standard deviation 2 kg. A checker tests whether the mean now **differs** from 15 kg. A random sample of $n = 16$ packs has mean $\acute{x} = 15.85$ kg. Test at the 5% level. (a) State H_0 and H_1 (two-tail). (b) Find z and the two-tail p-value. (c) State the decision. (d) State what the decision would have been for a one-tail test $H_1 : \mu > 15$ at $\alpha = 0.05$, and comment on which test is more conservative.

Unscaffolded

Q7. A production process fills sacks to a mean of 500 kg with known standard deviation 25 kg. A supervisor suspects the mean has **increased** and samples $n = 100$ sacks with mean $\acute{x} = 505$ kg. Test at the 5% level using a p-value, and state the decision.

Q8. A bus company advertises a mean journey time of 45 minutes with known standard deviation 6 minutes. A commuter group believes the true mean is **less** than 45 minutes and records a random sample of $n = 36$ journeys with mean $\acute{x} = 42.6$ minutes. Test at the 1% level using a p-value.

Q9. A snack bar is labelled with a mean energy of 250 kJ, known standard deviation 12 kJ. A consumer group tests whether the mean **differs** from 250 kJ, sampling $n=36$ bars with mean $\acute{x}=253$ kJ. Test at the 5% level using a p-value.

Q10. A tyre is rated for a mean tread life of 1000 km per mm of tread, with tread lives normally distributed and known standard deviation 50 km per mm. A laboratory suspects a new compound **increases** this and samples $n=25$ tyres with mean $\acute{x}=1018$ km per mm. Using the **critical-value method** at the 5% level, state z_α, the critical value of $\acute{x}$, and the decision.

Q11. A machine cuts rods to a mean length of 20 cm, known standard deviation 3 cm. An operator tests whether the mean has **drifted** from 20 cm. A random sample of $n=36$ rods has mean $\acute{x}=21.2$ cm. Using the **critical-value method** at the 5% level, state $z_{\alpha/2}$, the critical values of $\acute{x}$, and the decision.

Q12. A café claims a mean wait of 5 minutes, with waiting times normally distributed and known standard deviation 1.2 minutes. A customer believes the true mean is **less** and samples $n=16$ visits with mean $\acute{x}=4.55$ minutes. Using the **critical-value method** at the 10% level, state the critical value of $\acute{x}$ and the decision.

Q13. A supplement is stated to deliver a mean of 80 mg of an active ingredient, with the amount delivered normally distributed and known standard deviation 10 mg. An auditor tests whether the mean **differs** from 80 mg, sampling $n=25$ capsules with mean $\acute{x}=85.6$ mg. Test at the 1% level using a p-value.

Q14. A charity's donation records show a historic mean gift of $\$350$, with gifts modelled as normally distributed with known standard deviation $\$15$. After a campaign, an analyst asks whether the mean gift has **changed**, and samples $n=25$ gifts with mean $\acute{x}=\$355.70$. (a) State the appropriate H_1 and the type of test. (b) Test at the 5% level using a p-value and state the decision.

Q15. A sensor is calibrated to a mean reading of 120 units with known standard deviation 16 units. A random sample of $n=64$ readings has mean $\acute{x}=123.7$ units, giving $z=1.85$. (a) Test $H_1: \mu>120$ (upper one-tail) at $\alpha=0.05$ using a p-value. (b) Test $H_1: \mu\neq 120$ (two-tail) at $\alpha=0.05$ using a p-value. (c) Explain, using your two p-values, why the two-tail test is more conservative.

Q16. A dispenser is set to a mean of 2.5 mL with known standard deviation 0.4 mL. A technician tests whether the mean **differs** from 2.5 mL, sampling $n=64$ doses with mean $\acute{x}=2.61$ mL. Using the **critical-value method** at the 5% level, state the critical values of $\acute{x}$ and the decision.

Q17. An exam has a historic mean score of 72, with scores modelled as normally distributed with known standard deviation 10. A teacher believes a new program has **raised** the mean and samples $n=25$ students with mean $\acute{x}=77.2$. Test at the 5% level using a p-value.

Q18. A factory bags cement to a mean of 600 g with known standard deviation 30 g. A random sample of $n=36$ bags has mean $\acute{x}=612$ g, and management asks whether the mean fill has **changed**. (a) State H_0 and H_1. (b) Find z and the two-tail p-value; state the decision at $\alpha=0.05$. (c) Using the critical-value method at $\alpha=0.05$, state $z_{\alpha/2}$ and the critical values of $\acute{x}$, and confirm the decision.

Answers

Q1. $H_0: \mu = 200$, $H_1: \mu > 200$ (upper one-tail); $\sigma / \sqrt{n} = 0.5$, $z = 1.6$; $p = 0.0548$; $0.0548 > 0.05$, do not reject H_0 — insufficient evidence of an increase. **Q2.** $H_0: \mu = 50$, $H_1: \mu \neq 50$ (two-tail); $z = 2.2$; $p = 2(0.0139) = 0.0278$; $0.0278 < 0.05$, reject H_0 — the mean length differs from 50 mm. **Q3.** $H_0: \mu = 100$, $H_1: \mu < 100$; $z = -1.75$; critical value $-z_\alpha = -1.6449$, reject if $z \leq -1.6449$; $-1.75 \leq -1.6449$, reject H_0. **Q4.** $H_0: \mu = 750$, $H_1: \mu \neq 750$; $z = 2.25$; $z_{\alpha/2} = 1.9600$, critical values of $\acute{x}$ are 734.32 and 765.68 mL; $\acute{x} = 768 > 765.68$, reject H_0. **Q5.** $H_1: \mu > 60$; $z = 2.8$; $p = 0.0026$; $0.0026 < 0.01$, reject H_0 — evidence the mean reaction time is raised. **Q6.** $H_0: \mu = 15$, $H_1: \mu \neq 15$; $z = 1.7$, two-tail $p = 0.0891$; $0.0891 > 0.05$, do not reject H_0; a one-tail test would give $p = 0.0446 < 0.05$ and reject — the two-tail test is more conservative. **Q7.** $H_1: \mu > 500$; $z = 2$; $p = 0.0228 < 0.05$, reject H_0 — evidence the mean has increased. **Q8.** $H_1: \mu < 45$; $z = -2.4$; $p = \Phi(-2.4) = 0.0082 < 0.01$, reject H_0. **Q9.** $H_1: \mu \neq 250$; $z = 1.5$; two-tail $p = 2(0.0668) = 0.1336 > 0.05$, do not reject H_0. **Q10.** $z_\alpha = 1.6449$; $z = 1.8$; critical value of $\acute{x}$ is 1016.45 km per mm; $1018 > 1016.45$ ($z = 1.8 \geq 1.6449$), reject H_0. **Q11.** $z_{\alpha/2} = 1.9600$; $z = 2.4$; critical values of $\acute{x}$ are 19.02 and 20.98 cm; $21.2 > 20.98$, reject H_0. **Q12.** $z = -1.5$; critical value of $\acute{x}$ is 4.62 minutes ($-z_\alpha = -1.2816$); $4.55 \leq 4.62$, reject H_0. **Q13.** $H_1: \mu \neq 80$; $z = 2.8$; two-tail $p = 2(0.00256) = 0.0051 < 0.01$, reject H_0. **Q14.** (a) $H_1: \mu \neq 350$, two-tail; (b) $z = 1.9$, $p = 0.0574 > 0.05$, do not reject H_0 — insufficient evidence that the mean gift has changed. **Q15.** (a) $z = 1.85$, $p = 0.0322 < 0.05$, reject; (b) $p = 0.0643 > 0.05$, do not reject; (c) the two-tail p-value is double the one-tail value, so it exceeds α while the one-tail value does not — the two-tail test is more conservative. **Q16.** $z_{\alpha/2} = 1.9600$; critical values of $\acute{x}$ are 2.40 and 2.60 mL; $\acute{x} = 2.61 > 2.60$ ($z = 2.2$), reject H_0. **Q17.** $H_1: \mu > 72$; $z = 2.6$; $p = 0.0047 < 0.05$, reject H_0. **Q18.** (a) $H_0: \mu = 600$, $H_1: \mu \neq 600$; (b) $z = 2.4$, $p = 0.0164 < 0.05$, reject H_0; (c) $z_{\alpha/2} = 1.9600$, critical values of $\acute{x}$ are 590.20 and 609.80 g, and $612 > 609.80$ confirms rejection.

Fully-worked solutions

Q1. The barista suspects an **increase**, so $H_0: \mu = 200$ and $H_1: \mu > 200$ (upper one-tail). The standard error is $\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{64}} = \frac{4}{8} = 0.5$, so $z = \frac{200.8 - 200}{0.5} = 1.6$. The p-value is $P(Z \geq 1.6) = 1 - \Phi(1.6) = 1 - 0.9452 = 0.0548$. Since $0.0548 > 0.05$, we do **not** reject H_0: there is insufficient evidence at the 5% level that the new grinder has increased the mean volume.

Q2. “Differs from” gives $H_0: \mu = 50$, $H_1: \mu \neq 50$ (two-tail). The standard error is $\frac{5}{\sqrt{25}} = 1$, so $z = \frac{52.2 - 50}{1} = 2.2$. The single tail beyond $|z| = 2.2$ is $P(Z \geq 2.2) = 1 - \Phi(2.2) = 1 - 0.9861 = 0.0139$, so the two-tail p-value is $2 \times 0.0139 = 0.0278$. Since $0.0278 < 0.05$, reject H_0: the mean length differs from 50 mm.

Q3. The buyer suspects the mean is **less**, so $H_0: \mu = 100$, $H_1: \mu < 100$ (lower one-tail). The standard error is $\frac{8}{\sqrt{16}} = 2$, so $z = \frac{96.5 - 100}{2} = -1.75$. At $\alpha = 0.05$ the critical value is $z_\alpha = \Phi^{-1}(0.95) = 1.6449$, and we reject H_0 when $z \leq -1.6449$. Since $z = -1.75 \leq -1.6449$, the statistic lies in the rejection region, so we reject H_0: there is evidence at the 5% level that the true mean is less than 100 g.

Q4. “Changed” gives $H_0: \mu = 750$, $H_1: \mu \neq 750$ (two-tail). The standard error is $\frac{40}{\sqrt{25}} = 8$, so $z = \frac{768 - 750}{8} = 2.25$. At $\alpha = 0.05$ the two-tail critical value is $z_{\alpha/2} = \Phi^{-1}(0.975) = 1.9600$, giving critical values of $\acute{x}$

$$\acute{x}_c = 750 \pm 1.9600 \times 8 = 750 \pm 15.68,$$

that is 734.32 mL and 765.68 mL. Since $\acute{x} = 768 > 765.68$ (equivalently $|z| = 2.25 \geq 1.9600$), we reject H_0: the mean has changed from 750 mL.

Q5. A stimulant is expected to **raise** the mean, so the alternative is directional, $H_1: \mu > 60$ (upper one-tail), with $H_0: \mu = 60$. The standard error is $\frac{9}{\sqrt{36}} = 1.5$, so $z = \frac{64.2 - 60}{1.5} = 2.8$. The p-value is $P(Z \geq 2.8) = 1 - \Phi(2.8) = 1 - 0.9974 = 0.0026$. Since $0.0026 < 0.01$, reject H_0 at the 1% level: there is strong evidence the mean reaction time is raised.

Q6. "Differs" gives $H_0: \mu = 15$, $H_1: \mu \neq 15$ (two-tail). The standard error is $\frac{2}{\sqrt{16}} = 0.5$, so $z = \frac{15.85 - 15}{0.5} = 1.7$. The single tail beyond 1.7 is $1 - \Phi(1.7) = 0.04457$, so the two-tail p-value is $2 \times 0.04457 = 0.0891$. Since $0.0891 > 0.05$, we do **not** reject H_0. For a one-tail test $H_1: \mu > 15$ the p-value would be just $0.0446 < 0.05$, which **would** reject H_0. The two-tail test, needing $|z| \geq 1.9600$ rather than $z \geq 1.6449$, is the more conservative of the two.

Q7. An **increase** is suspected, so $H_0: \mu = 500$, $H_1: \mu > 500$ (upper one-tail). The standard error is $\frac{25}{\sqrt{100}} = 2.5$, so $z = \frac{505 - 500}{2.5} = 2$. The p-value is $P(Z \geq 2) = 1 - \Phi(2) = 1 - 0.9772 = 0.0228$. Since $0.0228 < 0.05$, reject H_0: there is evidence the mean has increased.

Q8. The group believes the mean is **less**, so $H_0: \mu = 45$, $H_1: \mu < 45$ (lower one-tail). The standard error is $\frac{6}{\sqrt{36}} = 1$, so $z = \frac{42.6 - 45}{1} = -2.4$. The p-value is $P(Z \leq -2.4) = \Phi(-2.4) = 0.0082$. Since $0.0082 < 0.01$, reject H_0 at the 1% level: there is evidence the mean journey time is less than 45 minutes.

Q9. "Differs" gives $H_0: \mu = 250$, $H_1: \mu \neq 250$ (two-tail). The standard error is $\frac{12}{\sqrt{36}} = 2$, so $z = \frac{253 - 250}{2} = 1.5$. The single tail is $1 - \Phi(1.5) = 0.0668$, so the two-tail p-value is $2 \times 0.0668 = 0.1336$. Since $0.1336 > 0.05$, do **not** reject H_0: the data are consistent with a mean of 250 kJ.

Q10. An **increase** is suspected, so $H_0: \mu = 1000$, $H_1: \mu > 1000$ (upper one-tail). The standard error is $\frac{50}{\sqrt{25}} = 10$, so $z = \frac{1018 - 1000}{10} = 1.8$. At $\alpha = 0.05$ the critical value is $z_\alpha = \Phi^{-1}(0.95) = 1.6449$; the critical value of $\acute{x}$ is

$$\acute{x}_c = 1000 + 1.6449 \times 10 = 1016.45 \text{ km per mm}.$$

Since $\acute{x} = 1018 \geq 1016.45$ (equivalently $z = 1.8 \geq 1.6449$), reject H_0: there is evidence the new compound increases mean tread life.

Q11. "Drifted" gives $H_0: \mu = 20$, $H_1: \mu \neq 20$ (two-tail). The standard error is $\frac{3}{\sqrt{36}} = 0.5$, so $z = \frac{21.2 - 20}{0.5} = 2.4$. At $\alpha = 0.05$ the two-tail critical value is $z_{\alpha/2} = \Phi^{-1}(0.975) = 1.9600$; the critical values of $\acute{x}$ are

$$\acute{x}_c = 20 \pm 1.9600 \times 0.5 = 20 \pm 0.98,$$

that is 19.02 cm and 20.98 cm. Since $\acute{x} = 21.2 > 20.98$ (equivalently $|z| = 2.4 \geq 1.9600$), reject H_0: the mean length has drifted from 20 cm.

Q12. The customer believes the mean is **less**, so $H_0: \mu = 5$, $H_1: \mu < 5$ (lower one-tail). The standard error is $\frac{1.2}{\sqrt{16}} = 0.3$, so $z = \frac{4.55 - 5}{0.3} = -1.5$. At $\alpha = 0.10$ the one-tail critical value is $z_\alpha = \Phi^{-1}(0.90) = 1.2816$, so the critical value of $\acute{x}$ is

$$\acute{x}_c = 5 - 1.2816 \times 0.3 = 4.62 \text{ minutes},$$

and we reject H_0 when $\acute{x} \le 4.62$. Since $\acute{x} = 4.55 \le 4.62$ (equivalently $z = -1.5 \le -1.2816$), reject H_0: there is evidence at the 10% level that the mean wait is less than 5 minutes.

Q13. "Differs" gives $H_0: \mu = 80$, $H_1: \mu \ne 80$ (two-tail). The standard error is $\frac{10}{\sqrt{25}} = 2$, so $z = \frac{85.6 - 80}{2} = 2.8$. The single tail is $1 - \Phi(2.8) = 0.00256$, so the two-tail p-value is $2 \times 0.00256 = 0.0051$. Since $0.0051 < 0.01$, reject H_0 at the 1% level: the mean amount of active ingredient differs from 80 mg.

Q14. (a) "Changed" names no direction, so $H_1: \mu \ne 350$ and the test is two-tailed ($H_0: \mu = 350$). (b) The standard error is $\frac{15}{\sqrt{25}} = 3$, so $z = \frac{355.70 - 350}{3} = 1.9$. The single tail is $1 - \Phi(1.9) = 0.0287$, so the two-tail p-value is $2 \times 0.0287 = 0.0574$. Since $0.0574 > 0.05$, do **not** reject H_0: there is insufficient evidence at the 5% level that the mean gift has changed after the campaign.

Q15. The standard error is $\frac{16}{\sqrt{64}} = 2$, so $z = \frac{123.7 - 120}{2} = 1.85$. (a) For the one-tail test the p-value is $P(Z \ge 1.85) = 1 - \Phi(1.85) = 0.0322$; since $0.0322 < 0.05$, reject H_0. (b) For the two-tail test the p-value is $2(1 - \Phi(1.85)) = 2 \times 0.03216 = 0.0643$; since $0.0643 > 0.05$, do **not** reject H_0. (c) The same statistic $z = 1.85$ gives a two-tail p-value (0.0643) that is **double** the single-tail probability (0.0322). Doubling pushes it above $\alpha = 0.05$, so the two-tail test fails to reject where the one-tail test rejects: the two-tail test is more conservative.

Q16. "Differs" gives $H_0: \mu = 2.5$, $H_1: \mu \ne 2.5$ (two-tail). The standard error is $\frac{0.4}{\sqrt{64}} = 0.05$, so $z = \frac{2.61 - 2.5}{0.05} = 2.2$. At $\alpha = 0.05$ the two-tail critical value is $z_{\alpha/2} = \Phi^{-1}(0.975) = 1.9600$; the critical values of $\acute{x}$ are

$$\acute{x}_c = 2.5 \pm 1.9600 \times 0.05 = 2.5 \pm 0.098,$$

that is 2.40 mL and 2.60 mL. Since $\acute{x} = 2.61 > 2.60$ (equivalently $|z| = 2.2 \ge 1.9600$), reject H_0: the mean dose differs from 2.5 mL.

Q17. A **rise** is expected, so $H_0: \mu = 72$, $H_1: \mu > 72$ (upper one-tail). The standard error is $\frac{10}{\sqrt{25}} = 2$, so $z = \frac{77.2 - 72}{2} = 2.6$. The p-value is $P(Z \ge 2.6) = 1 - \Phi(2.6) = 1 - 0.9953 = 0.0047$. Since $0.0047 < 0.05$, reject H_0: there is evidence the program has raised the mean score.

Q18. (a) "Changed" gives $H_0: \mu = 600$ and $H_1: \mu \ne 600$ (two-tail). (b) The standard error is $\frac{30}{\sqrt{36}} = 5$, so $z = \frac{612 - 600}{5} = 2.4$. The single tail is $1 - \Phi(2.4) = 0.0082$, so the two-tail p-value is $2 \times 0.0082 = 0.0164$; since $0.0164 < 0.05$, reject H_0. (c) At $\alpha = 0.05$ the two-tail critical value is $z_{\alpha/2} = \Phi^{-1}(0.975) = 1.9600$, giving critical values of $\acute{x}$

$$\acute{x}_c = 600 \pm 1.9600 \times 5 = 600 \pm 9.80,$$

that is 590.20 g and 609.80 g. Since $\acute{x} = 612 > 609.80$, the sample mean lies in the rejection region, confirming the decision to reject H_0: the mean fill has changed from 600 g.

Chapter 12 Confidence intervals and hypothesis testing for the mean — 12D Two-tail tests and confidence intervals

Theory

A **two-tail test** and a **confidence interval** are two views of the same calculation. This section revisits the two-tail test for a population mean and shows that it is equivalent to asking whether a proposed value lies inside a confidence interval built from the same sample. Throughout, the population standard deviation σ is known and the sample mean is modelled as normal — exactly so for a normal population, and approximately for a large sample — so every result is expressed on the standard-normal scale $Z \sim N(0,1)$.

The two-tail test, recalled.

As developed in earlier sections, a two-tail test of

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu \neq \mu_0$$

uses a random sample of size n with mean $\bar{x}$ and the **test statistic**

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}},$$

which measures how many standard errors the sample mean sits from the hypothesised mean μ_0. At significance level α there are two equivalent decision rules:

- **Critical-value method.** Reject H_0 when $|z| \geq z_{\alpha/2}$, where the **critical value** $z_{\alpha/2}$ cuts off an area $\frac{\alpha}{2}$ in each tail of $N(0,1)$.
- **p-value method.** Compute $p = 2P(Z \geq |z|)$ and reject H_0 when $p \leq \alpha$.

The confidence interval, recalled.

From the same sample, the $(1-\alpha)$ **confidence interval** for μ is

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \text{ that is } \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right).$$

The quantity $E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ is the **margin of error** — the same critical value $z_{\alpha/2}$ and the same standard error $\frac{\sigma}{\sqrt{n}}$ that appear in the test.

The duality.

Because the test and the interval are built from the same three ingredients, they must agree. Starting from the rejection rule and rearranging,

$$|z| \geq z_{\alpha/2} \Leftrightarrow \left| \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \right| \geq z_{\alpha/2} \Leftrightarrow |\bar{x} - \mu_0| \geq z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = E.$$

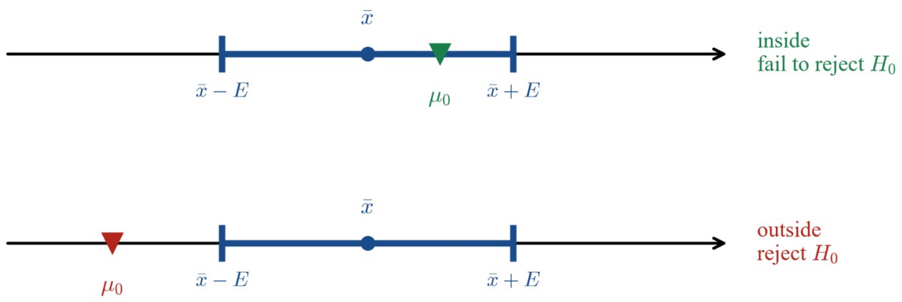
The last inequality says that μ_0 is at least a margin of error away from $\bar{x}$, i.e.

$$\mu_0 \leq \bar{x} - E \text{ or } \mu_0 \geq \bar{x} + E,$$

which is exactly the statement that μ_0 lies **outside** the interval $(\bar{x} - E, \bar{x} + E)$. Reading the chain in reverse gives the complementary statement. Hence

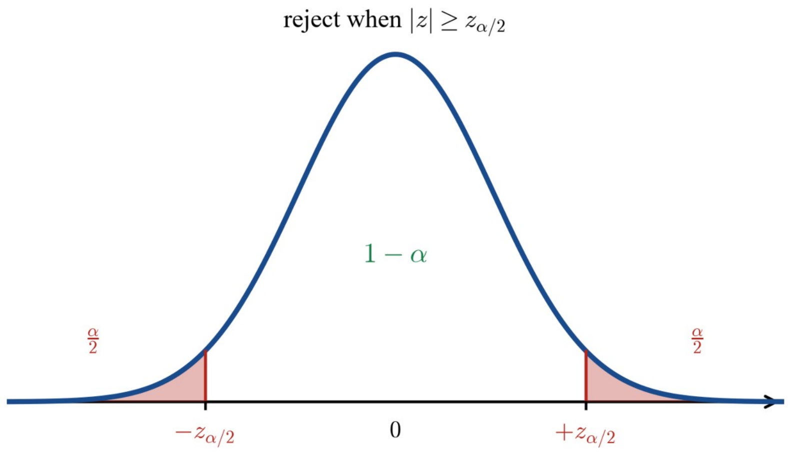
$$\text{a two-tail test of } H_0 : \mu = \mu_0 \text{ at level } \alpha \text{ rejects } H_0 \Leftrightarrow \mu_0 \text{ lies OUTSIDE the } (1-\alpha) \text{ confidence interval for } \mu.$$

Equivalently, the test **fails to reject** H_0 exactly when μ_0 lies **inside** the interval. The confidence level $1-\alpha$ and the significance level α are two names for the same choice: a 95% interval pairs with a two-tail test at $\alpha = 0.05$, a 99% interval with $\alpha = 0.01$, and so on.



The rejection region.

On the standard-normal scale the decision is a statement about z: reject H_0 when z falls in either shaded tail, $|z| \geq z_{\alpha/2}$. The two tails together carry probability α, split as $\frac{\alpha}{2}$ each; the unshaded centre carries the confidence $1-\alpha$. A value of μ_0 inside the confidence interval corresponds to a test statistic z inside this central region, and a value outside the interval to a z out in a tail.



Using a confidence interval to run a test — and the reverse.

The duality is a labour-saving device. Once the $(1-\alpha)$ interval has been constructed, **any** hypothesised mean μ_0 can be tested at level α by inspection: inside the interval means "do not reject", outside means "reject" — no test statistic need be recomputed. Conversely, if a two-tail test at level α reports that $H_0: \mu = \mu_0$ is rejected, then μ_0 must lie outside the $(1-\alpha)$ interval; if the test does not reject, μ_0 lies inside. The three methods — p-value, critical-value and confidence interval — are one inequality written three ways, so they always reach the same decision.

Confidence level $1-\alpha$	Significance α	Critical value $z_{\alpha/2}$
90 %	0.10	1.645
95 %	0.05	1.960
99 %	0.01	2.576

Effect of the confidence level, and borderline cases.

Raising the confidence level lowers α, enlarges $z_{\alpha/2}$ and so **widens** the interval; lowering it narrows the interval. Because a wider interval is more likely to contain μ_0, a higher confidence level makes rejection *harder*. This is why the same data can reject H_0 at the 5% level yet fail to reject it at the 1% level: μ_0 can lie outside the 95% interval but inside the wider 99% interval.

The decision is most delicate when μ_0 sits **near an endpoint** of the interval. Then the test statistic $|z|$ is close to the critical value $z_{\alpha/2}$ and the p-value is close to α, so a small change in the data — or in the chosen level — can flip the outcome. In such borderline cases it is good practice to quote the p-value and the interval, not merely the "reject / do not reject" label, so the reader can judge how close the call was.

Interpreting the decision.

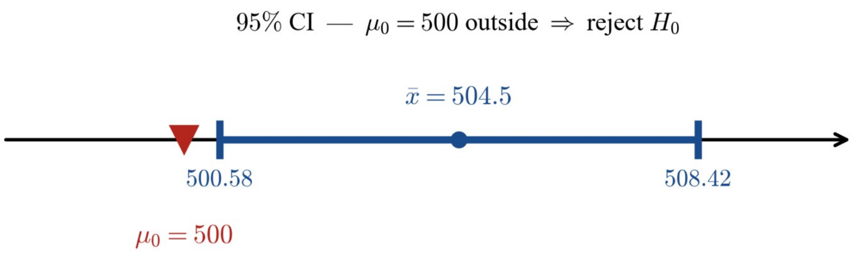
Rejecting H_0 means the sample gives significant evidence, at the stated level, that the true mean differs from μ_0 — in *either* direction, since the alternative is two-sided. Failing to reject means the evidence is not strong enough to distinguish μ from μ_0; it does **not** prove $\mu = \mu_0$. The confidence interval adds what a bare decision cannot: it displays the whole range of means consistent with the data, so its width and position show both the direction and the size of any plausible departure from μ_0.

Worked examples

Example 1 — scaffolded

A machine is set to fill bottles to a mean of 500 mL. The fill volume has known standard deviation $\sigma = 8$ mL. A random sample of $n = 16$ bottles has mean $\dot{x} = 504.5$ mL. Test $H_0 : \mu = 500$ against $H_1 : \mu \neq 500$ at the 5% level, using the confidence-interval, critical-value and p-value methods, and confirm the three agree.

Working	Comment
$H_0 : \mu = 500$, $H_1 : \mu \neq 500$; two-tail, $\alpha = 0.05$.	A two-tail test, so pair it with the 95% interval.
$\dfrac{\sigma}{\sqrt{n}} = \dfrac{8}{\sqrt{16}} = \dfrac{8}{4} = 2$.	Standard error of the sample mean.
95% CI: $504.5 \pm 1.96 \times 2 = 504.5 \pm 3.92 = (500.58, 508.42)$.	Margin of error $E = z_{0.025}\dfrac{\sigma}{\sqrt{n}} = 3.92$.
CI method: $\mu_0 = 500$ is below 500.58, so it lies *outside* the interval $\Rightarrow$ reject H_0.	Outside the interval means reject.
Critical-value method: $z = \dfrac{504.5 - 500}{2} = 2.25$; since $|z| = 2.25 \geq 1.96$, reject H_0.	The statistic lies in the rejection region.
p-value method: $p = 2P(Z \geq 2.25) = 2(1 - 0.9878) = 0.0244$; since $0.0244 \leq 0.05$, reject H_0.	$p \leq \alpha$, so reject.
All three methods reject H_0.	The interval, the critical value and the p-value agree.



There is significant evidence at the 5% level that the machine's mean fill differs from 500 mL; the 95% interval $(500.58, 508.42)$ suggests it is over-filling.

Example 2 — scaffolded

A supplier states that the mean protein content of its feed pellets is $\mu_0 = 250$ g/kg, with known standard deviation $\sigma = 12$ g/kg. A random sample of $n = 36$ batches gives $\bar{x} = 253$ g/kg. Construct a 95% confidence interval and use it to test, at the 5% level, the two claims $\mu = 250$ and $\mu = 248$.

Working	Comment
$\dfrac{\sigma}{\sqrt{n}} = \dfrac{12}{\sqrt{36}} = \dfrac{12}{6} = 2$.	Standard error.
95% CI: $253 \pm 1.96 \times 2 = 253 \pm 3.92 = (249.08, 256.92)$.	One interval will settle every 5% two-tail test.
Test $\mu = 250$: 250 lies between 249.08 and 256.92, so it is *inside* $\Rightarrow$ do not reject H_0.	Inside the interval means fail to reject.
Test $\mu = 248$: 248 is below 249.08, so it is *outside* $\Rightarrow$ reject H_0.	The same interval rejects this value.
Check ($\mu_0 = 250$): $z = \dfrac{253 - 250}{2} = 1.5$, $\lvert z \rvert = 1.5 < 1.96$, so do not reject — as the interval said.	The critical-value method confirms the interval.

One 95% interval carries out the two-tail test for every μ_0 at once: 250 is consistent with the data at the 5% level, whereas 248 is not.

Example 3 — unscaffolded

A manufacturer claims the mean reaction time of a component is $\mu_0 = 1.20$ s. The times are normally distributed with known standard deviation $\sigma = 0.15$ s. A random sample of $n = 25$ components has mean $\bar{x} = 1.11$ s. Test the claim at the 1% level, showing that the p-value, critical-value and confidence-interval methods agree, and interpret the result.

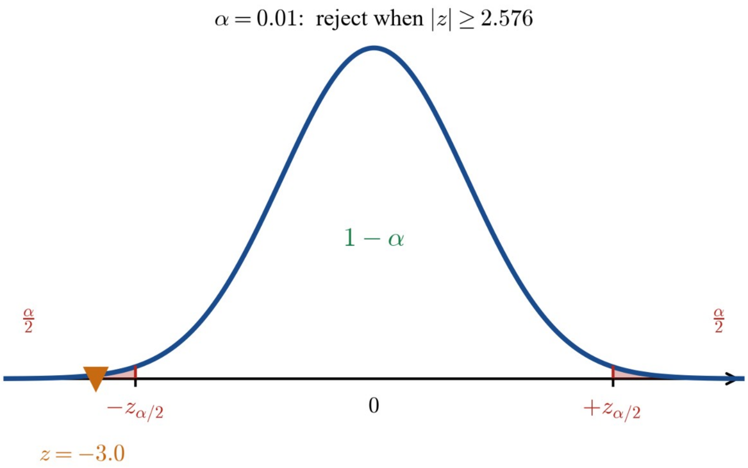
The test is two-tail, $H_0 : \mu = 1.20$ against $H_1 : \mu \neq 1.20$, at $\alpha = 0.01$. The standard error is $\dfrac{\sigma}{\sqrt{n}} = \dfrac{0.15}{\sqrt{25}} = \dfrac{0.15}{5} = 0.03$, so the test statistic is

$$z = \frac{1.11 - 1.20}{0.03} = \frac{-0.09}{0.03} = -3.0.$$

Critical-value method: for a two-tail test at $\alpha = 0.01$ the critical value is $z_{0.005} = 2.576$; since $\lvert z \rvert = 3.0 \geq 2.576$, reject H_0. **p-value method:** $p = 2P(Z \geq 3.0) = 2(1 - 0.99865) = 0.0027$, and $0.0027 \leq 0.01$, so reject H_0. **Confidence-interval method:** the 99% interval is

$$1.11 \pm 2.576 \times 0.03 = 1.11 \pm 0.0773 = (1.033, 1.187),$$

and $\mu_0 = 1.20$ lies above 1.187, so it is outside the interval — reject H_0.



∴ all three methods reject H_0 at the 1% level: there is strong evidence that the mean reaction time is not 1.20 s, and the 99% interval $(1.033, 1.187)$ places it below the claimed value.

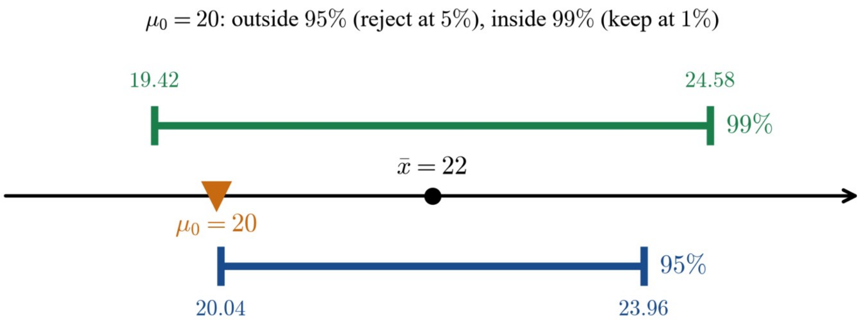
Example 4 — unscaffolded

A process has output with known standard deviation $\sigma = 5$. A random sample of $n = 25$ gives $\bar{x} = 22.0$. Test $H_0: \mu = 20$ against $H_1: \mu \neq 20$ first at the 5% level and then at the 1% level, and explain the borderline behaviour.

The standard error is $\frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{25}} = \frac{5}{5} = 1$, so the test statistic is $z = \frac{22.0 - 20}{1} = 2.0$ and the two-tail p-value is $p = 2P(Z \geq 2.0) = 2(1 - 0.97725) = 0.0455$.

At the 5% level, $z_{0.025} = 1.96$ and $|z| = 2.0 \geq 1.96$, so reject H_0; equivalently $p = 0.0455 \leq 0.05$. The 95% interval is $22.0 \pm 1.96 \times 1 = (20.04, 23.96)$, and $\mu_0 = 20$ lies just below 20.04, only 0.04 outside — reject.

At the 1% level, $z_{0.005} = 2.576$ and $|z| = 2.0 < 2.576$, so do not reject H_0; equivalently $p = 0.0455 > 0.01$. The wider 99% interval is $22.0 \pm 2.576 \times 1 = (19.42, 24.58)$, and now $\mu_0 = 20$ lies inside — do not reject.



∴ the same data reject H_0 at 5% but not at 1%. The value $\mu_0 = 20$ sits just outside the 95% interval yet inside the wider 99% interval; because the p-value 0.0455 lies between 0.01 and 0.05, this is a borderline case and the decision depends on the level chosen.

Practice questions

Scaffolded

Q1. A population has known standard deviation $\sigma = 10$. A random sample of $n = 25$ gives $\acute{x} = 46$. Test $H_0 : \mu = 50$ against $H_1 : \mu \neq 50$ at the 5% level. (a) Find the standard error $\frac{\sigma}{\sqrt{n}}$. (b) Construct the 95% confidence interval. (c) State whether $\mu_0 = 50$ is inside or outside the interval, and give the CI-method decision. (d) Find z and confirm the decision by the critical-value method.

Q2. Bags of sugar have known standard deviation $\sigma = 6$ g. A random sample of $n = 9$ bags has mean $\acute{x} = 31$ g above the nominal weight. Test $H_0 : \mu = 30$ against $H_1 : \mu \neq 30$ at the 5% level. (a) Find the standard error. (b) Construct the 95% confidence interval. (c) Is $\mu_0 = 30$ inside the interval? State the CI-method decision. (d) Find the p-value and confirm the decision.

Q3. A 95% confidence interval for a population mean μ is $(18.4, 23.6)$. Using this interval, test each hypothesis at the 5% level. (a) $H_0 : \mu = 20$. (b) $H_0 : \mu = 24$. (c) $H_0 : \mu = 18$.

Q4. A quantity has known standard deviation $\sigma = 15$. A random sample of $n = 25$ gives $\acute{x} = 112$. Test $H_0 : \mu = 100$ against $H_1 : \mu \neq 100$ at the 1% level. (a) Find the standard error and the test statistic z. (b) State the critical value $z_{0.005}$ and give the critical-value decision. (c) Construct the 99% confidence interval and give the CI-method decision.

Q5. A population has known standard deviation $\sigma = 8$. A random sample of $n = 16$ gives $\acute{x} = 53.8$. Test $H_0 : \mu = 50$ against $H_1 : \mu \neq 50$ at the 5% level. (a) Find the standard error and the test statistic z. (b) Find the p-value and give the p-value decision. (c) Construct the 95% confidence interval and confirm the decision. (d) Explain why this is a borderline case.

Q6. A two-tail test of $H_0 : \mu = 25$ at the 5% level gives $z = 2.4$, and H_0 is rejected. (a) Without computing the interval, state whether 25 lies inside or outside the 95% confidence interval, and why. (b) The sample gave $\acute{x} = 28$ with standard error 1.25. Construct the 95% interval and confirm your answer to (a).

Unscaffolded

Q7. A production line has output with known standard deviation $\sigma = 20$. A random sample of $n = 100$ items has mean $\acute{x} = 196$. Test $H_0 : \mu = 200$ against $H_1 : \mu \neq 200$ at the 5% level, using the confidence interval, and interpret the result in context.

Q8. A quantity has known standard deviation $\sigma = 5$. A random sample of $n = 100$ gives $\acute{x} = 49.2$. Test $H_0 : \mu = 50$ against $H_1 : \mu \neq 50$ at the 5% level, and confirm the CI and p-value methods agree.

Q9. A 99% confidence interval for a population mean is $(71.5, 78.5)$. Using this interval, test each hypothesis at the 1% level: (a) $\mu = 70$; (b) $\mu = 75$; (c) $\mu = 78$.

Q10. A population has known standard deviation $\sigma = 12$. A random sample of $n = 36$ gives $\acute{x} = 53.9$. Test $H_0 : \mu = 50$ against $H_1 : \mu \neq 50$ at the 5% level. Comment on how close the decision is.

Q11. The diameter of a bearing has known standard deviation $\sigma = 0.5$ mm. A random sample of $n = 25$ bearings has mean $\acute{x} = 5.28$ mm. Test $H_0 : \mu = 5.0$ against $H_1 : \mu \neq 5.0$ at the 5% level, using both the confidence-interval and p-value methods.

Q12. A population has known standard deviation $\sigma = 10$. A random sample of $n = 25$ gives $\acute{x} = 54.6$. Test $H_0 : \mu = 50$ against $H_1 : \mu \neq 50$ at the 5% level and again at the 1% level, in each case by the confidence-interval method, and account for any difference.

Q13. A quantity has known standard deviation $\sigma = 9$. A random sample of $n = 81$ gives $\acute{x} = 102.4$. Test $H_0 : \mu = 100$ against $H_1 : \mu \neq 100$ at the 5% level.

Q14. A two-tail z-test of $H_0 : \mu = 8$ at the 5% level gives a p-value of 0.03. (a) State the test decision at the 5% level. (b) Does the 95% confidence interval for μ contain 8? Explain. (c) Does the 90% confidence interval for μ contain 8 ? Explain, referring to interval width.

Q15. A 95% confidence interval for a population mean μ is $(11.8, 16.2)$. (a) State the sample mean $\bar{x}$ and the margin of error. (b) Test $H_0 : \mu = 17$ at the 5% level. (c) For the same data a 90% interval would be narrower. Explain whether the decision on $\mu = 17$ could change.

Q16. A manufacturer claims a mean battery lifetime of 100 hours; the lifetime has known standard deviation $\sigma = 25$ hours. A random sample of $n = 100$ batteries has mean $\bar{x} = 104.5$ hours. Test the claim at the 5% level and interpret the outcome in context.

Q17. A liquid is dispensed with known standard deviation $\sigma = 4$ mL. A random sample of $n = 64$ has mean $\bar{x} = 21.6$ mL. Test $H_0 : \mu = 20$ against $H_1 : \mu \neq 20$ at the 1% level, showing that the p-value, critical-value and confidence-interval methods agree.

Q18. A population has known standard deviation $\sigma = 10$. A random sample of $n = 100$ gives $\bar{x} = 52$. (a) Test $H_0 : \mu = 50$ at the 5% level by the confidence-interval method. (b) Test $H_0 : \mu = 50$ at the 1% level by the confidence-interval method. (c) Explain, in terms of interval width, how $\mu_0 = 50$ can be significant at 5% but not at 1%.

Answers

Q1. SE $=2$; 95% CI $(42.08, 49.92)$; 50 is outside $\Rightarrow$ reject H_0; $z=-2.0$, $|z|=2.0\geq 1.96$, reject. **Q2.** SE $=2$; 95% CI $(27.08, 34.92)$; 30 is inside $\Rightarrow$ do not reject H_0; $p=0.6171>0.05$, do not reject. **Q3.** (a) 20 inside $\Rightarrow$ do not reject; (b) 24 outside $\Rightarrow$ reject; (c) 18 outside $\Rightarrow$ reject. **Q4.** SE $=3$, $z=4.0$; $z_{0.005}=2.576$, $|z|\geq 2.576 \Rightarrow$ reject; 99% CI $(104.27, 119.73)$, 100 outside $\Rightarrow$ reject. **Q5.** SE $=2$, $z=1.9$; $p=0.0574>0.05 \Rightarrow$ do not reject; 95% CI $(49.88, 57.72)$, 50 just inside; borderline since $|z|=1.9$ is close to 1.96. **Q6.** (a) Outside, because a rejected two-tail test means μ_0 lies outside the matching interval; (b) 95% CI $(25.55, 30.45)$, 25 is outside — confirmed. **Q7.** SE $=2$; 95% CI $(192.08, 199.92)$; 200 outside $\Rightarrow$ reject H_0: significant evidence the mean is not 200 (it appears lower). **Q8.** SE $=0.5$; 95% CI $(48.22, 50.18)$, 50 inside $\Rightarrow$ do not reject; $z=-1.6$, $p=0.1096>0.05$, do not reject — methods agree. **Q9.** (a) 70 outside $\Rightarrow$ reject; (b) 75 inside $\Rightarrow$ do not reject; (c) 78 inside $\Rightarrow$ do not reject. **Q10.** SE $=2$, $z=1.95$; 95% CI $(49.98, 57.82)$, 50 just inside $\Rightarrow$ do not reject; $p=0.0512$, only just above 0.05 — very close. **Q11.** SE $=0.1$; 95% CI $(5.084, 5.476)$, 5.0 outside $\Rightarrow$ reject; $z=2.8$, $p=0.0051\leq 0.05$, reject. **Q12.** SE $=2$, $z=2.3$. At 5%: 95% CI $(50.68, 58.52)$, 50 outside $\Rightarrow$ reject. At 1%: 99% CI $(49.45, 59.75)$, 50 inside $\Rightarrow$ do not reject. The wider 99% interval now contains 50. **Q13.** SE $=1$, $z=2.4$; 95% CI $(100.44, 104.36)$, 100 outside $\Rightarrow$ reject H_0. **Q14.** (a) $p=0.03\leq 0.05 \Rightarrow$ reject; (b) no — a rejected 5% test means 8 is outside the 95% interval; (c) no — the 90% interval is narrower, and 8 is already outside the wider 95% interval, so it is outside the 90% one too. **Q15.** (a) $\bar{x}=14$, margin $=2.2$; (b) 17 is outside $(11.8, 16.2) \Rightarrow$ reject H_0; (c) no — a narrower interval still excludes 17, which is already outside the wider 95% interval, so the decision stands. **Q16.** SE $=2.5$; 95% CI $(99.60, 109.40)$, 100 inside $\Rightarrow$ do not reject: not enough evidence at 5% to dispute the claimed mean of 100 hours. **Q17.** SE $=0.5$, $z=3.2$; $z_{0.005}=2.576$, reject; $p=0.0014\leq 0.01$, reject; 99% CI $(20.31, 22.89)$, 20 outside, reject — all agree. **Q18.** (a) 95% CI $(50.04, 53.96)$, 50 outside $\Rightarrow$ reject; (b) 99% CI $(49.42, 54.58)$, 50 inside $\Rightarrow$ do not reject; (c) the 99% interval is wider, so it contains 50 while the narrower 95% interval does not.

Fully-worked solutions

Q1. The standard error is $\frac{\sigma}{\sqrt{n}}=\frac{10}{\sqrt{25}}=\frac{10}{5}=2$. The 95% interval is $\bar{x}\pm 1.96\frac{\sigma}{\sqrt{n}}=46\pm 1.96\times 2=46\pm 3.92=(42.08, 49.92)$. Since $\mu_0=50$ exceeds the upper endpoint 49.92, it lies outside the interval, so by the CI method reject H_0. Confirming, $z=\frac{46-50}{2}=-2.0$, and $|z|=2.0\geq 1.96$, so the critical-value method also rejects H_0.

Q2. The standard error is $\frac{6}{\sqrt{9}}=\frac{6}{3}=2$. The 95% interval is $31\pm 1.96\times 2=31\pm 3.92=(27.08, 34.92)$. Since $\mu_0=30$ lies between 27.08 and 34.92, it is inside the interval, so do not reject H_0. Confirming, $z=\frac{31-30}{2}=0.5$, so $p=2P(Z\geq 0.5)=2(1-0.69146)=0.6171>0.05$; the p-value method agrees.

Q3. The interval $(18.4, 23.6)$ tests every value at the 5% level. (a) 20 lies inside, so do not reject $H_0: \mu=20$. (b) 24 lies above 23.6, so it is outside — reject $H_0: \mu=24$. (c) 18 lies below 18.4, so it is outside — reject $H_0: \mu=18$.

Q4. The standard error is $\frac{15}{\sqrt{25}}=\frac{15}{5}=3$, so $z=\frac{112-100}{3}=\frac{12}{3}=4.0$. For a two-tail test at $\alpha=0.01$ the critical value is $z_{0.005}=2.576$; since $|z|=4.0\geq 2.576$, reject H_0. The 99% interval is $112\pm 2.576\times 3=112\pm 7.73=(104.27, 119.73)$, and $\mu_0=100$ lies below 104.27, so it is outside — reject H_0. Both methods agree.

Q5. The standard error is $\frac{8}{\sqrt{16}}=\frac{8}{4}=2$, so $z=\frac{53.8-50}{2}=\frac{3.8}{2}=1.9$. Then $p=2P(Z\geq 1.9)=2(1-0.9713)=0.0574>0.05$, so do not reject H_0. The 95% interval is $53.8\pm 1.96\times 2=53.8\pm 3.92=(49.88,57.72)$; since $\mu_0=50$ exceeds the lower endpoint 49.88 by only 0.12, it lies just inside the interval, so the CI method also fails to reject. This is a borderline case: $|z|=1.9$ is only a little below the critical value 1.96, and $p=0.0574$ is only a little above 0.05, so a slightly larger $\acute{x}$ would have reversed the decision.

Q6. (a) The two-tail test rejected $H_0: \mu=25$ at the 5% level, and by the duality a rejected level-α test means μ_0 lies outside the matching $(1-\alpha)$ interval; hence 25 lies **outside** the 95% interval. (b) The margin of error is $1.96\times 1.25=2.45$, so the 95% interval is $28\pm 2.45=(25.55,30.45)$. Since 25 lies below 25.55, it is outside the interval, confirming part (a). (As a check, $z=\frac{28-25}{1.25}=2.4$, matching the given statistic.)

Q7. The standard error is $\frac{20}{\sqrt{100}}=\frac{20}{10}=2$. The 95% interval is $196\pm 1.96\times 2=196\pm 3.92=(192.08,199.92)$. Since $\mu_0=200$ exceeds the upper endpoint 199.92, it is outside the interval, so reject H_0 at the 5% level. In context, there is significant evidence that the line's mean output is not 200; the whole interval lies below 200, so the mean appears to have fallen short.

Q8. The standard error is $\frac{5}{\sqrt{100}}=\frac{5}{10}=0.5$. The 95% interval is $49.2\pm 1.96\times 0.5=49.2\pm 0.98=(48.22,50.18)$; since $\mu_0=50$ lies inside, do not reject H_0. Confirming, $z=\frac{49.2-50}{0.5}=-1.6$, so $p=2P(Z\geq 1.6)=2(1-0.9452)=0.1096>0.05$ — the two methods agree.

Q9. The 99% interval $(71.5,78.5)$ tests each value at the 1% level. (a) 70 lies below 71.5, so it is outside — reject $H_0: \mu=70$. (b) 75 lies inside, so do not reject $H_0: \mu=75$. (c) 78 lies inside (below 78.5), so do not reject $H_0: \mu=78$.

Q10. The standard error is $\frac{12}{\sqrt{36}}=\frac{12}{6}=2$, so $z=\frac{53.9-50}{2}=1.95$. The 95% interval is $53.9\pm 1.96\times 2=53.9\pm 3.92=(49.98,57.82)$; since $\mu_0=50$ exceeds the lower endpoint 49.98 by just 0.02, it lies just inside, so do not reject H_0. The decision is extremely close: $p=2P(Z\geq 1.95)=2(1-0.9744)=0.0512$, barely above 0.05, so the test only just fails to reject.

Q11. The standard error is $\frac{0.5}{\sqrt{25}}=\frac{0.5}{5}=0.1$. The 95% interval is $5.28\pm 1.96\times 0.1=5.28\pm 0.196=(5.084,5.476)$; since $\mu_0=5.0$ lies below 5.084, it is outside — reject H_0. Confirming, $z=\frac{5.28-5.0}{0.1}=2.8$, so $p=2P(Z\geq 2.8)=2(1-0.99744)=0.0051\leq 0.05$; the p-value method agrees.

Q12. The standard error is $\frac{10}{\sqrt{25}}=\frac{10}{5}=2$, so $z=\frac{54.6-50}{2}=2.3$. **At the 5% level:** the 95% interval is $54.6\pm 1.96\times 2=54.6\pm 3.92=(50.68,58.52)$; since $\mu_0=50$ lies below 50.68, it is outside — reject H_0. **At the 1% level:** the 99% interval is $54.6\pm 2.576\times 2=54.6\pm 5.15=(49.45,59.75)$; now $\mu_0=50$ lies inside — do not reject H_0. Raising the confidence level from 95% to 99% widened the interval enough to swallow $\mu_0=50$, so the significant result at 5% becomes non-significant at 1% (consistent with $p=2P(Z\geq 2.3)=0.0214$, which lies between 0.01 and 0.05).

Q13. The standard error is $\frac{9}{\sqrt{81}}=\frac{9}{9}=1$, so $z=\frac{102.4-100}{1}=2.4$. The 95% interval is $102.4\pm 1.96\times 1=102.4\pm 1.96=(100.44,104.36)$; since $\mu_0=100$ lies below 100.44, it is outside — reject H_0. (Equivalently $|z|=2.4\geq 1.96$, or $p=2P(Z\geq 2.4)=0.0164\leq 0.05$.)

Q14. (a) Since $p=0.03 \le 0.05$, reject H_0 at the 5% level. (b) No: a rejected two-tail test at the 5% level means, by the duality, that 8 lies outside the 95% interval. (c) No: the 90% interval uses the smaller critical value $z_{0.05}=1.645<1.960$, so it is **narrower** than the 95% interval. As 8 already lies outside the wider 95% interval, it must also lie outside the narrower 90% interval; equivalently $p=0.03 \le 0.10$, so the test also rejects at the 10% level.

Q15. (a) The sample mean is the midpoint, $\acute{x}=\frac{11.8+16.2}{2}=14$, and the margin of error is half the width, $\frac{16.2-11.8}{2}=2.2$. (b) The value $\mu_0=17$ lies above the upper endpoint 16.2, so it is outside the 95% interval — reject H_0 at the 5% level. (c) No: a 90% interval (with $z_{0.05}=1.645$) is narrower and still centred at 14, so it reaches only to $14+1.645\times\frac{2.2}{1.96}\approx 15.85<17$; since 17 is already outside the wider 95% interval, it stays outside the narrower one — the rejection stands.

Q16. The standard error is $\frac{25}{\sqrt{100}}=\frac{25}{10}=2.5$. The 95% interval is $104.5 \pm 1.96\times 2.5=104.5\pm 4.90=(99.60, 109.40)$; since the claimed mean $\mu_0=100$ lies inside, do not reject H_0 at the 5% level. In context, the sample mean of 104.5 hours is higher than the claimed 100 hours, but not by enough to be significant at the 5% level: the data are consistent with a true mean of 100 hours (indeed with any mean from about 99.6 to 109.4 hours).

Q17. The standard error is $\frac{4}{\sqrt{64}}=\frac{4}{8}=0.5$, so $z=\frac{21.6-20}{0.5}=3.2$. **Critical-value:** $z_{0.005}=2.576$ and $|z|=3.2\ge 2.576$, reject. **p-value:** $p=2P(Z\ge 3.2)=2(1-0.9993)=0.0014\le 0.01$, reject. **Confidence interval:** $21.6\pm 2.576\times 0.5=21.6\pm 1.29=(20.31, 22.89)$, and $\mu_0=20$ lies below 20.31, so it is outside — reject. All three methods reject H_0 at the 1% level.

Q18. The standard error is $\frac{10}{\sqrt{100}}=\frac{10}{10}=1$, and the sample mean is $\acute{x}=52$. (a) The 95% interval is $52\pm 1.96\times 1=52\pm 1.96=(50.04, 53.96)$; since $\mu_0=50$ lies below 50.04, it is outside — reject H_0 at the 5% level. (b) The 99% interval is $52\pm 2.576\times 1=52\pm 2.576=(49.42, 54.58)$; now $\mu_0=50$ lies inside — do not reject H_0 at the 1% level. (c) The 99% interval uses the larger critical value 2.576 in place of 1.96, so it is wider than the 95% interval. That extra width is enough to include $\mu_0=50$, which the narrower 95% interval excluded; hence the mean is significantly different from 50 at the 5% level but not at the 1% level. (The test statistic $z=2.0$ gives $p=0.0455$, between 0.01 and 0.05, confirming this.)

Chapter 12 Confidence intervals and hypothesis testing for the mean — 12E Errors in hypothesis testing

Theory

A significance test for a population mean ends in one of two decisions: **reject** the null hypothesis H_0, or **do not reject** H_0. The decision is based on a single random sample, so it can be wrong. Because the sample mean $\acute{X}$ is a random variable, a sample can — by chance — point to the wrong conclusion even when the test is carried out perfectly. This section measures how often that happens, using **conditional probability**.

Recall the set-up. A population is normal with **known** standard deviation σ (or the sample is large), and we test a claim about its mean μ. The null hypothesis fixes a value $H_0 : \mu = \mu_0$; the alternative is one-tailed, $H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$, or two-tailed, $H_1 : \mu \neq \mu_0$. Under H_0 the sample mean has the sampling distribution

$$\acute{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right), \text{standard error } se = \frac{\sigma}{\sqrt{n}}.$$

We reject H_0 when the standardised test statistic $z = \frac{\acute{x} - \mu_0}{se}$ is beyond the **critical value**, equivalently when $\acute{x}$ falls in the **rejection region** and not in the **acceptance region**.

The two kinds of error.

The true state of nature (H_0 true, or H_0 false) is unknown, and our decision may or may not match it. Crossing the two possibilities gives four outcomes, two correct and two wrong.

	H_0 is true	H_0 is false ($\mu = \mu_1$)
Reject H_0	**Type I error** ($P = \alpha$)	correct decision (power $= 1 - \beta$)
Do not reject H_0	correct decision ($P = 1 - \alpha$)	**Type II error** ($P = \beta$)

A **Type I error** is rejecting H_0 when H_0 is in fact **true** — a "false alarm". Its probability is exactly the **significance level** of the test,

$$P(\text{Type I error}) = P(\text{reject } H_0 / H_0 \text{ true}) = \alpha.$$

This is a conditional probability: *given* that H_0 holds, the test statistic still lands in the rejection region a proportion α of the time. Choosing $\alpha = 0.05$ or $\alpha = 0.01$ is therefore a deliberate choice of how large a Type I error rate we are willing to tolerate — it is set **before** the data are seen, and the test controls it automatically.

A **Type II error** is failing to reject H_0 when H_0 is in fact **false** — a "missed detection". Unlike a Type I error, its probability depends on *how* false H_0 is, so we evaluate it for a **specific alternative** value $\mu = \mu_1$. Writing β for this probability,

$$P(\text{Type II error}) = P(\text{do not reject } H_0 / \mu = \mu_1) = \beta.$$

Power.

The complement of a Type II error is a correct rejection: the test detects that the mean really has shifted to μ_1. This probability is the fourth cell of the table above, the **power** of the test,

$$\text{power} = 1 - \beta = P(\text{reject } H_0 / \mu = \mu_1).$$

Like α and β it is a conditional probability, read off the sampling distribution centred at μ_1. A powerful test is one that is likely to reject H_0 when it should. Like β, the power is computed for a stated alternative μ_1: a test may have high power against a large shift and low power against a small one.

Computing β and the power.

The method has two steps: first find the **acceptance region on the $\bar{x}$ scale**, then find the probability that the sampling distribution **centred at μ_1** lands in it.

Take a right-tailed test $H_1: \mu > \mu_0$ at significance α. The critical z-value is $z^* = \Phi^{-1}(1-\alpha)$, and H_0 is rejected when $\bar{x} \geq \bar{x}_c$, where the critical sample mean is

$$\bar{x}_c = \mu_0 + z^* \frac{\sigma}{\sqrt{n}}.$$

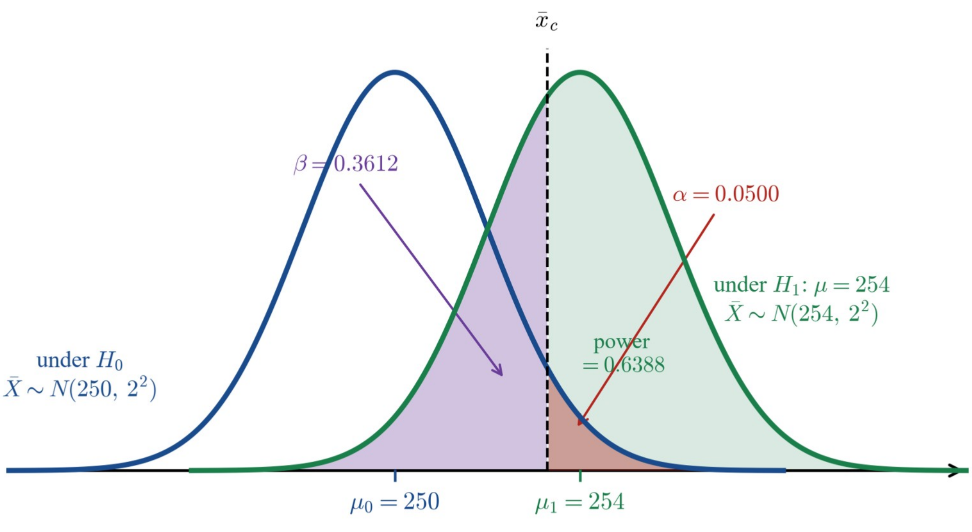
So the acceptance region is $\bar{x} < \bar{x}_c$. If the true mean is μ_1, then $\bar{X} \sim N\left(\mu_1, \sigma^2/n\right)$, and a Type II error occurs exactly when this $\bar{X}$ falls in the acceptance region:

$$\beta = P(\bar{X} < \bar{x}_c / \mu = \mu_1), \text{power} = 1 - \beta = P(\bar{X} \geq \bar{x}_c / \mu = \mu_1).$$

For a left-tailed test the rejection region is the lower tail, $\bar{x} \leq \bar{x}_c$ with $\bar{x}_c = \mu_0 - z^* \sigma / \sqrt{n}$, and $\beta = P(\bar{X} > \bar{x}_c / \mu = \mu_1)$. For a two-tailed test there are two critical values $\mu_0 \pm z^* \sigma / \sqrt{n}$ with $z^* = \Phi^{-1}(1 - \alpha/2)$, and β is the probability the μ_1-centred curve lands **between** them. In every case the boundary is fixed by H_0 and α, while β is read from the curve centred at μ_1.

Rounding. Critical values and probabilities are quoted correct to four decimal places, and the sign ≈ marks every displayed value that has been rounded. All of the arithmetic is carried out with the full-precision values stored by technology: a rounded value that is read off the page and typed back in can shift the last digit of the next answer, so keep the stored value rather than the printed one.

The picture below shows both sampling distributions for a right-tailed test with $\mu_0 = 250$, $\sigma = 8$, $n = 16$ (so $se = 2$) at $\alpha = 0.05$, against the specific alternative $\mu_1 = 254$. The critical value is $\bar{x}_c = 250 + 1.6449 \times 2 \approx 253.2897$. The red upper tail of the H_0 curve has area $\alpha = 0.05$; the purple part of the H_1 curve, lying in the acceptance region, has area $\beta \approx 0.3612$; and the remaining green part of the H_1 curve is the power, $1 - \beta \approx 0.6388$.



Numerically, since the H_1 curve is $\bar{X} \sim N(254, 2^2)$,

$$\beta = P(\bar{X} < 253.2897 / \mu = 254) \approx 0.3612, \text{power} \approx 0.6388,$$

both evaluated with technology from the sampling distribution centred at μ_1. So even though the mean really has shifted up to 254, a sample of 16 fails to detect it about 36% of the time.

What makes β small and the power large?

Three quantities control the Type II error rate. Understanding them is the heart of test design.

- **Effect size** $|\mu_1 - \mu_0|$. The further the true mean is from μ_0, the less the two curves overlap, so the μ_1-curve puts less area in the acceptance region: β falls and the power rises. A test detects a big departure more easily than a small one.
- **Sample size** n. Increasing n shrinks the standard error $\sigma / \sqrt{n}$, so **both** curves become taller and narrower and separate more cleanly (their centres μ_0 and μ_1 are a fixed distance apart, but that distance is now many standard errors). Again β falls and the power rises. Enlarging the sample is the usual way to buy more power without changing α.
- **Significance level** α. Raising α pushes the critical value $\acute{x}_c$ **towards** μ_0, enlarging the rejection region; this lowers β and raises the power — but it raises the Type I error rate by the same change in α. Lowering α does the reverse. This is the α–β **trade-off**: for a fixed sample size you cannot make both error probabilities small at once; reducing one enlarges the other. The only way to shrink both together is to collect a larger sample.

$$\alpha = P(\text{Type I error}), \beta = P(\text{Type II error}), \text{power} = 1 - \beta$$

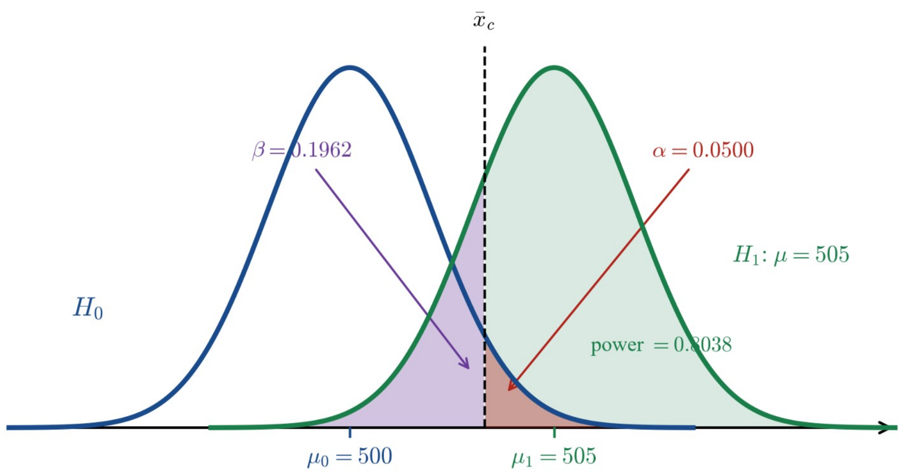
The two comparison figures accompanying Examples 3 and 4 show the effect of n and of α as the shaded areas grow and shrink.

Worked examples

Example 1 — scaffolded

A bottling line is set to a target mean fill of $\mu_0 = 500$ mL, with known standard deviation $\sigma = 6$ mL. To check for **over-filling** a quality officer tests $H_0: \mu = 500$ against $H_1: \mu > 500$ at $\alpha = 0.05$, using the mean of a random sample of $n = 9$ bottles. (a) Find the critical value $\acute{x}_c$ of the sample mean. (b) If the true mean has drifted to $\mu_1 = 505$ mL, find the probability of a Type II error. (c) State the power of the test against $\mu_1 = 505$.

Working	Comment
$se = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{9}} = \frac{6}{3} = 2.$	Standard error of $\acute{X}$; the population is normal, so $\acute{X}$ is exactly normal.
Right-tailed, $\alpha = 0.05$: $z^* = \Phi^{-1}(0.95) \approx 1.6449.$	Critical z-value for a one-tailed test.
$\acute{x}_c = \mu_0 + z^* se = 500 + 1.6449 \times 2 \approx 503.2897.$	Reject H_0 when $\acute{x} \geq 503.2897$; accept when $\acute{x} < 503.2897$.
Under $\mu_1 = 505$: $\acute{X} \sim N(505, 2^2)$, so $\beta = P(\acute{X} < 503.2897 / \mu = 505) \approx 0.1962.$	Type II error = the μ_1-curve's area in the acceptance region (technology).
Power $= 1 - \beta \approx 0.8038.$	Probability the test correctly detects the shift to 505.



Example 2 — scaffolded

A supplier states that the mean content of its jars is $\mu_0 = 100$ g. A consumer group suspects **under-filling** and tests $H_0 : \mu = 100$ against $H_1 : \mu < 100$ at $\alpha = 0.05$, with a random sample of $n = 25$ jars; the content is normal with $\sigma = 5$ g. (a) Find the critical value $\acute{x}_c$. (b) Find the probability of a Type II error if the true mean is $\mu_1 = 97$ g. (c) State the power against $\mu_1 = 97$.

Working	Comment
$se = \frac{5}{\sqrt{25}} = \frac{5}{5} = 1.$	Standard error of the sample mean.
Left-tailed, $\alpha = 0.05$: $z^* = \Phi^{-1}(0.95) \approx 1.6449.$	Same critical z; the tail is now on the **left**.
$\acute{x}_c = \mu_0 - z^* se = 100 - 1.6449 \times 1 \approx 98.3551.$	Reject H_0 when $\acute{x} \leq 98.3551$; accept when $\acute{x} > 98.3551$.
Under $\mu_1 = 97$: $\acute{X} \sim N(97, 1^2)$, so $\beta = P(\acute{X} > 98.3551 / \mu = 97) \approx 0.0877.$	Acceptance region is the **upper** side for a left-tailed test.
Power $= 1 - \beta \approx 0.9123.$	The large shift (3 g $= 3$ se below μ_0) is detected about 91% of the time.

Example 3 — unscaffolded

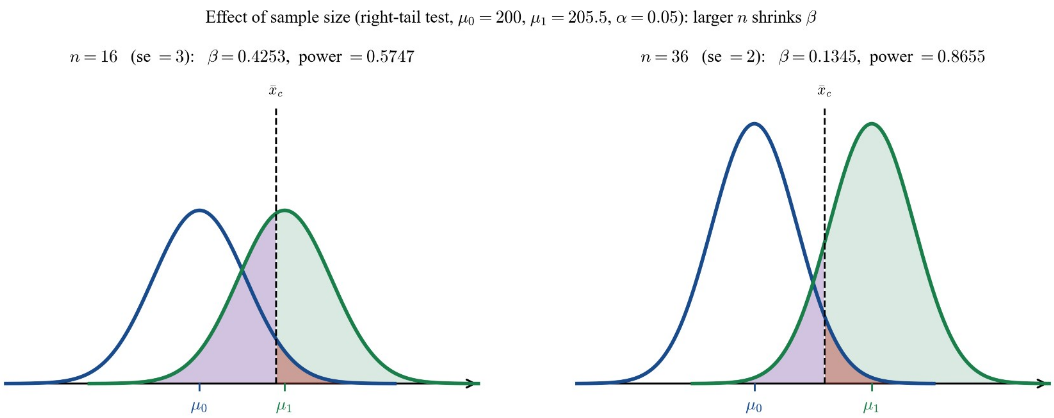
A machine cuts rods to a target mean length $\mu_0 = 200$ mm, with $\sigma = 12$ mm. A right-tailed test of $H_0 : \mu = 200$ against $H_1 : \mu > 200$ is run at $\alpha = 0.05$. The true mean has actually shifted to $\mu_1 = 205.5$ mm. Find the power of the test using a sample of size $n = 16$, and again using $n = 36$, and comment.

For $n = 16$, $se = \frac{12}{\sqrt{16}} = 3$, and the critical value is $\acute{x}_c = 200 + 1.6449 \times 3 \approx 204.9346$. Under $\mu_1 = 205.5$ the sample mean is $\acute{X} \sim N(205.5, 3^2)$, so

$$\beta = P(\acute{X} < 204.9346 / \mu = 205.5) \approx 0.4253, \text{power} \approx 0.5747.$$

For $n = 36$, $se = \frac{12}{\sqrt{36}} = 2$, and $\acute{x}_c = 200 + 1.6449 \times 2 \approx 203.2897$. Now $\acute{X} \sim N(205.5, 2^2)$ and

$$\beta = P(\acute{X} < 203.2897 / \mu = 205.5) \approx 0.1345, \text{power} \approx 0.8655.$$



∴ enlarging the sample from 16 to 36 lifts the power from 0.5747 to 0.8655 (and cuts β from 0.4253 to 0.1345): the smaller standard error separates the two sampling distributions, so the same real shift of 5.5 mm is far easier to detect.

Example 4 — unscaffolded

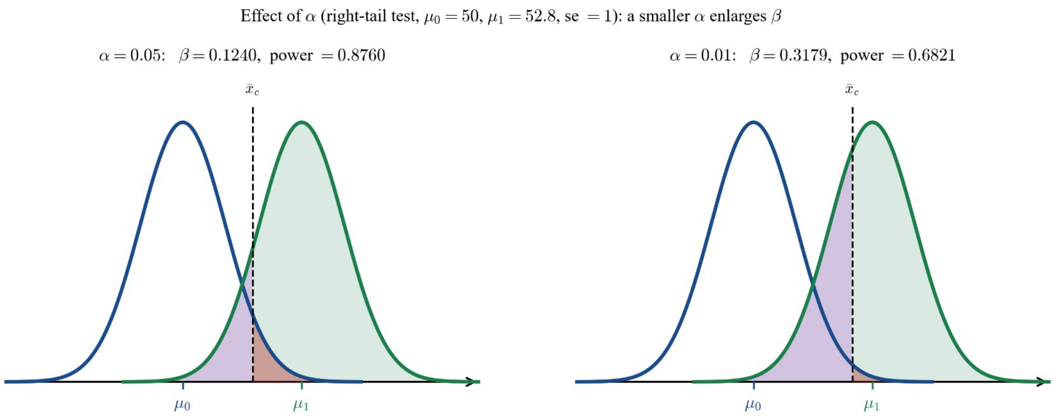
A process has target mean $\mu_0 = 50$ with $\sigma = 4$. A right-tailed test of $H_0 : \mu = 50$ against $H_1 : \mu > 50$ uses a sample of size $n = 16$, so $se = 1$. The true mean is $\mu_1 = 52.8$. Compare the Type II error and power when the test is run at $\alpha = 0.05$ and at $\alpha = 0.01$.

At $\alpha = 0.05$: $z^* = \Phi^{-1}(0.95) \approx 1.6449$, so $\acute{x}_c = 50 + 1.6449 \times 1 \approx 51.6449$. Under $\mu_1 = 52.8$, $\acute{X} \sim N(52.8, 1^2)$ and

$$\beta = P(\acute{X} < 51.6449 \,/\, \mu = 52.8) \approx 0.1240, \text{power} \approx 0.8760.$$

At $\alpha = 0.01$: $z^* = \Phi^{-1}(0.99) \approx 2.3263$, so $\acute{x}_c = 50 + 2.3263 \times 1 \approx 52.3263$, further from μ_0. Under $\mu_1 = 52.8$,

$$\beta = P(\acute{X} < 52.3263 \,/\, \mu = 52.8) \approx 0.3179, \text{power} \approx 0.6821.$$



∴ tightening the significance level from $\alpha = 0.05$ to $\alpha = 0.01$ makes a Type I error less likely but a Type II error **more** likely — β rises from 0.1240 to 0.3179 and the power falls from 0.8760 to 0.6821. This is the α–β trade-off at a fixed sample size.

Practice questions

Scaffolded

Q1. A new medicine is claimed to have a mean recovery time of $\mu_0 = 8$ days. Researchers test $H_0 : \mu = 8$ against $H_1 : \mu < 8$ (the medicine is faster) at significance level $\alpha = 0.05$. (a) Describe, in context, what a Type I error would be. (b) Describe, in context, what a Type II error would be. (c) State $P(\text{Type I error})$. (d) Explain what the *power* of this test measures.

Q2. A filling process has target mean $\mu_0 = 20$ with known $\sigma = 3$; the population is normal. A right-tailed test of $H_0 : \mu = 20$ against $H_1 : \mu > 20$ uses a sample of size $n = 9$ at $\alpha = 0.05$. The true mean is $\mu_1 = 22.2$. (a) Find $se = \sigma / \sqrt{n}$. (b) Find the critical value $\acute{x}_c = \mu_0 + z^* se$. (c) Find $\beta = P(\acute{X} < \acute{x}_c \,/\, \mu = 22.2)$. (d) State the power $1 - \beta$.

Q3. A cable has target mean breaking strain $\mu_0 = 500$ with $\sigma = 20$. A left-tailed test of $H_0 : \mu = 500$ against $H_1 : \mu < 500$ uses $n = 16$ at $\alpha = 0.05$. The true mean is $\mu_1 = 488$. (a) Find se. (b) Find $\acute{x}_c = \mu_0 - z^* se$. (c) Find $\beta = P(\acute{X} > \acute{x}_c \,/\, \mu = 488)$. (d) State the power.

Q4. A right-tailed test has $\mu_0 = 100$, $\sigma = 9$, $n = 9$ (so $se = 3$), against the true mean $\mu_1 = 105$. (a) At $\alpha = 0.05$ find $\acute{x}_c$, then β and the power. (b) At $\alpha = 0.10$ ($z^* \approx 1.2816$) find $\acute{x}_c$, then β and the power. (c) State which significance level gives the larger power. (d) State which gives the larger Type I error rate.

Q5. A right-tailed test has $\mu_0 = 60$, $\sigma = 8$, $\alpha = 0.05$, against the true mean $\mu_1 = 63.5$. (a) For $n = 16$, find se and $\acute{x}_c$. (b) Find the power for $n = 16$. (c) For $n = 64$, find se and $\acute{x}_c$. (d) Find the power for $n = 64$ and comment on the effect of the larger sample.

Q6. A two-tailed test of $H_0: \mu = 25$ against $H_1: \mu \neq 25$ uses $\sigma = 6$, $n = 9$ (so $se = 2$) at $\alpha = 0.05$, and $z^* = \Phi^{-1}(0.975) \approx 1.9600$. The true mean is $\mu_1 = 30$. (a) Find the two critical values $\mu_0 \pm z^* se$, giving the acceptance region. (b) State the distribution of $\bar{X}$ under $\mu_1 = 30$. (c) Find β, the probability the sample mean lands in the acceptance region. (d) State the power.

Unscaffolded

Q7. A smoke detector is tested. Treat H_0: "there is no fire" as the null hypothesis, with the alarm sounding (rejecting H_0) when smoke is sensed. (a) Describe a Type I error and a Type II error for this detector in context. (b) Explain why a manufacturer might deliberately choose a large α, and what this does to the power.

Q8. A right-tailed test has $\mu_0 = 150$, $\sigma = 15$, $n = 25$, $\alpha = 0.05$. Find the critical value $\bar{x}_c$, then the probability of a Type II error and the power against the true mean $\mu_1 = 158$.

Q9. A left-tailed test has $\mu_0 = 40$, $\sigma = 9$, $n = 9$, $\alpha = 0.05$. Find $\bar{x}_c$, then β and the power against the true mean $\mu_1 = 36$.

Q10. A two-tailed test of $H_0: \mu = 1000$ against $H_1: \mu \neq 1000$ has $\sigma = 40$, $n = 16$, $\alpha = 0.05$ (so $z^* \approx 1.9600$). Find the acceptance region, then β and the power against the true mean $\mu_1 = 1030$.

Q11. A right-tailed test has $\mu_0 = 500$, $\sigma = 25$, $n = 25$, $\alpha = 0.05$. (a) Find the power against the true mean $\mu_1 = 508$. (b) Find the power against the true mean $\mu_1 = 516$. (c) Explain, referring to effect size, why the two powers differ.

Q12. A right-tailed test has $\mu_0 = 10$, $\sigma = 2$, $\alpha = 0.05$, against the true mean $\mu_1 = 10.9$. Find the power for a sample of size $n = 16$ and for a sample of size $n = 64$, and comment.

Q13. A right-tailed test has $\mu_0 = 75$, $\sigma = 12$, $n = 16$, against the true mean $\mu_1 = 79$. Find β and the power at $\alpha = 0.10$ ($z^* \approx 1.2816$) and at $\alpha = 0.01$ ($z^* \approx 2.3263$), and state the trade-off shown.

Q14. A right-tailed test has $\mu_0 = 100$, $\sigma = 10$ and $\alpha = 0.05$, and must have power at least 0.90 against the true mean $\mu_1 = 105$. Writing the requirement as $\beta \leq 0.10$ and standardising the critical value $\bar{x}_c = \mu_0 + z^* se$ under μ_1, turn the requirement into an inequality for se, hence find the smallest sample size n, and confirm the power it delivers. (You may use $\Phi^{-1}(0.90) \approx 1.2816$.)

Q15. A two-tailed test of $H_0: \mu = 200$ against $H_1: \mu \neq 200$ has $\sigma = 16$, $n = 64$, $\alpha = 0.01$ (so $z^* = \Phi^{-1}(0.995) \approx 2.5758$). Find the acceptance region, then β and the power against the true mean $\mu_1 = 207$.

Q16. A quality test of $H_0: \mu = 300$ against $H_1: \mu > 300$ uses $\sigma = 8$, $n = 25$ ($se = 1.6$) at $\alpha = 0.05$. The sample gives $\bar{x} = 302$. (a) Compute the test statistic z and state the decision at $\alpha = 0.05$. (b) Given that decision, state which type of error could have been made. (c) If the true mean is $\mu_1 = 303$, find the probability of that error and the power.

Q17. A right-tailed test has $\mu_0 = 80$, $\sigma = 6$, $n = 9$ ($se = 2$) at $\alpha = 0.05$, so $\bar{x}_c \approx 83.2897$. Find the value of the true mean μ_1 (with $\mu_1 > \mu_0$) against which the test has power exactly 0.80. (Hint: for power 0.80, $\beta = 0.20$, and $\Phi^{-1}(0.20) \approx -0.8416$.)

Q18. Light bulbs are advertised with mean life $\mu_0 = 1200$ hours and known $\sigma = 90$ hours. A consumer body tests $H_0: \mu = 1200$ against $H_1: \mu < 1200$ at $\alpha = 0.05$, using a random sample of $n = 36$ bulbs. (a) Describe a Type I error and a Type II error in this context. (b) Find the critical value $\bar{x}_c$. (c) If the true mean life is $\mu_1 = 1165$ hours, find β and the power. (d) The sample is enlarged to $n = 100$. Find the new power against $\mu_1 = 1165$, and comment.

Answers

Q1. (a) Concluding the medicine is faster ($\mu < 8$) when really $\mu = 8$; (b) failing to detect that it is faster when really $\mu < 8$; (c) $P(\text{Type I error}) = \alpha = 0.05$; (d) the probability the test correctly concludes $\mu < 8$ when the medicine really is faster, i.e. $1 - \beta$. **Q2.** (a) $se = 1$; (b) $\bar{x}_c \approx 21.6449$; (c) $\beta \approx 0.2894$; (d) power ≈ 0.7106. **Q3.** (a) $se = 5$; (b) $\bar{x}_c \approx 491.7757$; (c) $\beta \approx 0.2251$; (d) power ≈ 0.7749. **Q4.** (a) $\bar{x}_c \approx 104.9346$, $\beta \approx 0.4913$, power ≈ 0.5087; (b) $\bar{x}_c \approx 103.8447$, $\beta \approx 0.3501$, power ≈ 0.6499; (c) $\alpha = 0.10$; (d) $\alpha = 0.10$. **Q5.** (a) $se = 2$, $\bar{x}_c \approx 63.2897$; (b) power ≈ 0.5419; (c) $se = 1$, $\bar{x}_c \approx 61.6449$; (d) power ≈ 0.9682 — the larger sample raises the power. **Q6.** (a) $21.0801 < \bar{x} < 28.9199$; (b) $\bar{X} \sim N(30, 2^2)$; (c) $\beta \approx 0.2946$; (d) power ≈ 0.7054. **Q7.** (a) Type I: the alarm sounds when there is no fire (false alarm); Type II: the alarm stays silent when there is a fire (missed fire). (b) A large α makes the alarm more sensitive, so it misses fewer real fires (higher power / smaller β), at the cost of more false alarms. **Q8.** $\bar{x}_c \approx 154.9346$; $\beta \approx 0.1534$, power ≈ 0.8466. **Q9.** $\bar{x}_c \approx 35.0654$; $\beta \approx 0.6223$, power ≈ 0.3777. **Q10.** $980.4004 < \bar{x} < 1019.5996$; $\beta \approx 0.1492$, power ≈ 0.8508. **Q11.** (a) power ≈ 0.4821; (b) power ≈ 0.9400; (c) $\mu_1 = 516$ is a larger effect size, so the curves overlap less and the power is greater. **Q12.** $n = 16$: power ≈ 0.5616; $n = 64$: power ≈ 0.9747 — the larger sample raises the power. **Q13.** $\alpha = 0.10$: $\beta \approx 0.4794$, power ≈ 0.5206; $\alpha = 0.01$: $\beta \approx 0.8396$, power ≈ 0.1604. A smaller α lowers the Type I rate but raises β and lowers the power. **Q14.** $n = 35$; it delivers power $\approx 0.9054\ (\geq 0.90)$. **Q15.** $194.8483 < \bar{x} < 205.1517$; $\beta \approx 0.1777$, power ≈ 0.8223. **Q16.** (a) $z = 1.25 < 1.6449$, so do not reject H_0; (b) a Type II error; (c) $\beta \approx 0.4090$, power ≈ 0.5910. **Q17.** $\mu_1 \approx 84.9729$ (about 84.97). **Q18.** (a) Type I: concluding the bulbs are short-lived ($\mu < 1200$) when really $\mu = 1200$; Type II: failing to detect short life when really $\mu < 1200$. (b) $\bar{x}_c \approx 1175.3272$; (c) $\beta \approx 0.2456$, power ≈ 0.7544; (d) power ≈ 0.9876 — the larger sample greatly raises the power.

Fully-worked solutions

Q1. The test is $H_0: \mu = 8$ versus $H_1: \mu < 8$. (a) A **Type I error** is rejecting H_0 when it is true: concluding the mean recovery time is *less* than 8 days when in fact it equals 8 days. (b) A **Type II error** is not rejecting H_0 when it is false: failing to conclude the medicine is faster when the true mean recovery time really is below 8 days. (c) The Type I error rate is the significance level, $P(\text{Type I error}) = \alpha = 0.05$. (d) The **power** is $1 - \beta = P(\text{reject } H_0 / \mu = \mu_1)$ — the probability the test correctly concludes the medicine is faster when the true mean really is the lower value μ_1.

Q2. (a) $se = \frac{3}{\sqrt{9}} = \frac{3}{3} = 1$. (b) Right-tailed at $\alpha = 0.05$, $z^* = \Phi^{-1}(0.95) \approx 1.6449$, so $\bar{x}_c = 20 + 1.6449 \times 1 \approx 21.6449$; reject H_0 when $\bar{x} \geq 21.6449$. (c) Under $\mu_1 = 22.2$, $\bar{X} \sim N(22.2, 1^2)$, so $\beta = P(\bar{X} < 21.6449 / \mu = 22.2) \approx 0.2894$. (d) Power $= 1 - \beta \approx 0.7106$.

Q3. (a) $se = \frac{20}{\sqrt{16}} = \frac{20}{4} = 5$. (b) Left-tailed at $\alpha = 0.05$, $z^* \approx 1.6449$, so $\bar{x}_c = 500 - 1.6449 \times 5 \approx 491.7757$; reject H_0 when $\bar{x} \leq 491.7757$, accept when $\bar{x} > 491.7757$. (c) Under $\mu_1 = 488$, $\bar{X} \sim N(488, 5^2)$, so $\beta = P(\bar{X} > 491.7757 / \mu = 488) \approx 0.2251$. (d) Power $= 1 - \beta \approx 0.7749$.

Q4. Here $se = \frac{9}{\sqrt{9}} = 3$ and $\bar{X} \sim N(105, 3^2)$ under μ_1. (a) At $\alpha = 0.05$, $z^* \approx 1.6449$, so $\bar{x}_c = 100 + 1.6449 \times 3 \approx 104.9346$; then $\beta = P(\bar{X} < 104.9346 / \mu = 105) \approx 0.4913$ and power ≈ 0.5087. (b) At $\alpha = 0.10$, $z^* \approx 1.2816$, so $\bar{x}_c = 100 + 1.2816 \times 3 \approx 103.8447$; then $\beta = P(\bar{X} < 103.8447 / \mu = 105) \approx 0.3501$ and power ≈ 0.6499. (c) The larger significance level $\alpha = 0.10$ gives the larger power. (d) It also gives the larger Type I error rate, since $P(\text{Type I error}) = \alpha$.

Q5. $\mu_0 = 60$, $\sigma = 8$, $\alpha = 0.05$ ($z^* \approx 1.6449$), true mean $\mu_1 = 63.5$. (a) For $n = 16$, $se = \frac{8}{\sqrt{16}} = 2$ and $\bar{x}_c = 60 + 1.6449 \times 2 \approx 63.2897$. (b) Under $\mu_1 = 63.5$, $\bar{X} \sim N(63.5, 2^2)$, so $\beta = P(\bar{X} < 63.2897 / \mu = 63.5) \approx 0.4581$

and power ≈ 0.5419. (c) For $n=64$, $se=\frac{8}{\sqrt{64}}=1$ and $\acute{x}_c=60+1.6449\times 1\approx 61.6449$. (d) Now $\acute{X}\sim N(63.5,1^2)$, so $\beta=P(\acute{X}<61.6449 / \mu=63.5)\approx 0.0318$ and power ≈ 0.9682. Quadrupling the sample size halves the standard error, separating the curves and lifting the power from 0.5419 to 0.9682.

Q6. (a) With $se=\frac{6}{\sqrt{9}}=2$ and $z^*\approx 1.9600$, the critical values are $25\pm 1.9600\times 2$, that is $\acute{x}_c\approx 21.0801$ and $\acute{x}_c\approx 28.9199$; the acceptance region is $21.0801<\acute{x}<28.9199$. (b) Under $\mu_1=30$, $\acute{X}\sim N(30,2^2)$. (c) A Type II error is the sample mean landing in the acceptance region: $\beta=P(21.0801<\acute{X}<28.9199 / \mu=30)\approx 0.2946$. (d) Power $=1-\beta\approx 0.7054$.

Q7. (a) Taking H_0: "no fire", a **Type I error** is rejecting H_0 when it is true — the alarm sounds although there is no fire (a false alarm). A **Type II error** is not rejecting H_0 when it is false — the alarm stays silent although there really is a fire (a dangerous missed detection). (b) Choosing a large α makes the detector quicker to sound, enlarging the rejection region; this lowers the Type II error rate β and so raises the power (fewer missed fires). The cost is a higher Type I rate — more false alarms — which for a safety device is usually the acceptable trade.

Q8. $se=\frac{15}{\sqrt{25}}=3$; right-tailed at $\alpha=0.05$, $z^*\approx 1.6449$, so $\acute{x}_c=150+1.6449\times 3\approx 154.9346$. Under $\mu_1=158$, $\acute{X}\sim N(158,3^2)$, so $\beta=P(\acute{X}<154.9346 / \mu=158)\approx 0.1534$ and power $=1-\beta\approx 0.8466$.

Q9. $se=\frac{9}{\sqrt{9}}=3$; left-tailed at $\alpha=0.05$, $z^*\approx 1.6449$, so $\acute{x}_c=40-1.6449\times 3\approx 35.0654$; reject when $\acute{x}\le 35.0654$. Under $\mu_1=36$, $\acute{X}\sim N(36,3^2)$, so $\beta=P(\acute{X}>35.0654 / \mu=36)\approx 0.6223$ and power $=1-\beta\approx 0.3777$. Even though the mean has fallen by 4, that is only 1.33 standard errors, so the test misses the drop more often than it detects it.

Q10. $se=\frac{40}{\sqrt{16}}=10$; two-tailed at $\alpha=0.05$, $z^*\approx 1.9600$, so the critical values are $1000\pm 1.9600\times 10$, that is ≈ 980.4004 and ≈ 1019.5996; the acceptance region is $980.4004<\acute{x}<1019.5996$. Under $\mu_1=1030$, $\acute{X}\sim N(1030,10^2)$, so $\beta=P(980.4004<\acute{X}<1019.5996 / \mu=1030)\approx 0.1492$ and power $=1-\beta\approx 0.8508$.

Q11. $se=\frac{25}{\sqrt{25}}=5$; right-tailed at $\alpha=0.05$, $z^*\approx 1.6449$, so $\acute{x}_c=500+1.6449\times 5\approx 508.2243$. (a) Under $\mu_1=508$, $\acute{X}\sim N(508,5^2)$, so $\beta=P(\acute{X}<508.2243 / \mu=508)\approx 0.5179$ and power ≈ 0.4821. (b) Under $\mu_1=516$, $\acute{X}\sim N(516,5^2)$, so $\beta=P(\acute{X}<508.2243 / \mu=516)\approx 0.0600$ and power ≈ 0.9400. (c) The alternative $\mu_1=516$ is 16 units above μ_0 (an effect size of 3.2 se), whereas $\mu_1=508$ is only 8 units above (1.6 se). The larger effect size makes the H_1 curve overlap the acceptance region far less, so the power is much greater.

Q12. Right-tailed at $\alpha=0.05$, $z^*\approx 1.6449$, true mean $\mu_1=10.9$. For $n=16$, $se=\frac{2}{\sqrt{16}}=0.5$ and $\acute{x}_c=10+1.6449\times 0.5\approx 10.8224$; under $\mu_1=10.9$, $\acute{X}\sim N(10.9,0.5^2)$, so $\beta=P(\acute{X}<10.8224 / \mu=10.9)\approx 0.4384$ and power ≈ 0.5616. For $n=64$, $se=\frac{2}{\sqrt{64}}=0.25$ and $\acute{x}_c=10+1.6449\times 0.25\approx 10.4112$; under $\mu_1=10.9$, $\acute{X}\sim N(10.9,0.25^2)$, so $\beta=P(\acute{X}<10.4112 / \mu=10.9)\approx 0.0253$ and power ≈ 0.9747. The larger sample sharply reduces the standard error, so the power rises from 0.5616 to 0.9747.

Q13. $se=\frac{12}{\sqrt{16}}=3$, true mean $\mu_1=79$, $\acute{X}\sim N(79,3^2)$. At $\alpha=0.10$, $z^*\approx 1.2816$, so $\acute{x}_c=75+1.2816\times 3\approx 78.8447$ and $\beta=P(\acute{X}<78.8447 / \mu=79)\approx 0.4794$, power ≈ 0.5206. At $\alpha=0.01$, $z^*\approx 2.3263$, so $\acute{x}_c=75+2.3263\times 3\approx 81.9790$ and $\beta=P(\acute{X}<81.9790 / \mu=79)\approx 0.8396$, power ≈ 0.1604. Reducing α from 0.10 to 0.01 lowers the Type I error rate but pushes the critical value further out, so β rises from 0.4794 to 0.8396 and the power falls from 0.5206 to 0.1604 — the α–β trade-off.

Q14. With $se=\dfrac{10}{\sqrt{n}}$ the critical value is $\acute{x}_c=100+z^* se$, where $z^*=\Phi^{-1}(0.95)$. Power ≥ 0.90 means $\beta \leq 0.10$, and standardising the boundary under $\mu_1=105$ turns this into

$$\beta=P\left(Z<\frac{\acute{x}_c-105}{se}\right)\leq 0.10 \Leftrightarrow \frac{\acute{x}_c-105}{se}\leq \Phi^{-1}(0.10).$$

Now $\acute{x}_c-105=z^* se-5$ and $\Phi^{-1}(0.10)=-\Phi^{-1}(0.90)$, so the condition becomes $z^*-\dfrac{5}{se}\leq -\Phi^{-1}(0.90)$, that is

$$\frac{5}{se}\geq z^*+\Phi^{-1}(0.90) \Leftrightarrow \sqrt{n}\geq \frac{\left(z^*+\Phi^{-1}(0.90)\right)\times 10}{5},$$

using $se=10/\sqrt{n}$. With $z^*\approx 1.6449$ and $\Phi^{-1}(0.90)\approx 1.2816$ this gives $\sqrt{n}\geq 5.8528$, so $n\geq 34.26$. Since n must be a whole number and a larger sample only raises the power, round **up** to $n=35$. Check: with $n=35$, $se=\dfrac{10}{\sqrt{35}}\approx 1.6903$, so $\acute{x}_c=100+1.6449\times 1.6903\approx 102.7803$; under $\mu_1=105$, $\acute{X}\sim N(105, 1.6903^2)$ gives $\beta=P(\acute{X}<102.7803 / \mu=105)\approx 0.0946$ and power $\approx 0.9054\geq 0.90$, as required. (With $n=34$ the power is only 0.8981, so 35 really is the smallest.)

Q15. $se=\dfrac{16}{\sqrt{64}}=2$; two-tailed at $\alpha=0.01$, $z^*=\Phi^{-1}(0.995)\approx 2.5758$, so the critical values are $200\pm 2.5758\times 2$, that is ≈ 194.8483 and ≈ 205.1517; the acceptance region is $194.8483<\acute{x}<205.1517$. Under $\mu_1=207$, $\acute{X}\sim N(207, 2^2)$, so $\beta=P(194.8483<\acute{X}<205.1517 / \mu=207)\approx 0.1777$ and power $=1-\beta\approx 0.8223$.

Q16. (a) $se=\dfrac{8}{\sqrt{25}}=1.6$, so $z=\dfrac{\acute{x}-\mu_0}{se}=\dfrac{302-300}{1.6}=1.25$. Since the critical value is $z^*\approx 1.6449$ and $1.25<1.6449$, the result is not significant: **do not reject** H_0. (b) Because we did not reject H_0, the only possible error is a **Type II error** (failing to detect a real increase). (Rejecting H_0 would risk a Type I error; not rejecting risks a Type II error.) (c) The critical value on the $\acute{x}$ scale is $\acute{x}_c=300+1.6449\times 1.6\approx 302.6318$. Under $\mu_1=303$, $\acute{X}\sim N(303, 1.6^2)$, so $\beta=P(\acute{X}<302.6318 / \mu=303)\approx 0.4090$ and power $=1-\beta\approx 0.5910$.

Q17. For power 0.80 the Type II error is $\beta=0.20$. Under the alternative the sample mean is $\acute{X}\sim N(\mu_1, 2^2)$, and a Type II error is $\acute{X}<\acute{x}_c\approx 83.2897$, so $\beta=P(\acute{X}<83.2897 / \mu=\mu_1)=0.20$. Standardising the boundary under μ_1,

$$\frac{83.2897-\mu_1}{2}=\Phi^{-1}(0.20)\approx -0.8416,$$

so $\mu_1\approx 83.2897+0.8416\times 2\approx 84.9729$. Thus the test has power exactly 0.80 against a true mean of about 84.97: shifts larger than this are detected with probability greater than 0.80, smaller shifts with less.

Q18. $se=\dfrac{90}{\sqrt{36}}=15$; the test is left-tailed at $\alpha=0.05$, $z^*\approx 1.6449$. (a) A **Type I error** is rejecting H_0 when true: concluding the mean life is below 1200 hours when it really is 1200. A **Type II error** is not rejecting H_0 when false: failing to detect that the true mean life is below 1200 hours. (b) $\acute{x}_c=1200-1.6449\times 15\approx 1175.3272$; reject H_0 when $\acute{x}\leq 1175.3272$. (c) Under $\mu_1=1165$, $\acute{X}\sim N(1165, 15^2)$, so $\beta=P(\acute{X}>1175.3272 / \mu=1165)\approx 0.2456$ and power $=1-\beta\approx 0.7544$. (d) With $n=100$, $se=\dfrac{90}{\sqrt{100}}=9$, so $\acute{x}_c=1200-1.6449\times 9\approx 1185.1963$; under $\mu_1=1165$, $\acute{X}\sim N(1165, 9^2)$ gives $\beta=P(\acute{X}>1185.1963 / \mu=1165)\approx 0.0124$ and power ≈ 0.9876. Enlarging the sample from 36 to 100 cuts the standard error from 15 to 9 and raises the power from 0.7544 to 0.9876, so the larger sample is far more likely to detect the 35-hour shortfall.

Topic 5 Test A — Technology-active

One bound reference and one approved technology (CAS calculator or software) are permitted. Reading time: 3 minutes. Writing time: 45 minutes. Total 40 marks. In all questions where a numerical answer is required, an exact value must be given unless otherwise specified. Where more than one mark is available, appropriate working must be shown.

Question 1 (7 marks)

An artisan bakery produces sourdough loaves whose mass X grams is modelled by $X \sim N(760, 15^2)$, and the loaves are baked independently of one another.

a. As a loaf cools it loses 4% of its mass, so its cooled mass is $C = 0.96X$. Find $E(C)$ and $sd(C)$. 2 marks

b. A gift box holds 13 independently baked (uncooled) loaves; let $T = X_1 + X_2 + \cdots + X_{13}$ be the total mass. Find $E(T)$ and $sd(T)$. A trainee claims that $sd(T) = 13 \times 15 = 195$ g. Explain why the trainee is wrong. 3 marks

c. Find $P(T < 9800)$, correct to four decimal places. 2 marks

Question 2 (6 marks)

On a delivery run, the distance a courier van travels in town is $X \sim N(40, 5^2)$ km and the distance on the highway is $Y \sim N(120, 10^2)$ km, with X and Y independent. The van uses fuel at 0.09 litres per town kilometre and 0.06 litres per highway kilometre, so the fuel used on a run is

$$L = 0.09X + 0.06Y \text{ litres.}$$

a. Find $E(L)$ and $Var(L)$. 3 marks

b. State the distribution of L and find $P(L > 12)$, correct to four decimal places. 2 marks

c. The van makes two independent runs of this type in a day. Find the standard deviation of the total fuel used, correct to four decimal places. 1 mark

Question 3 (6 marks)

The time X minutes a barista takes to make one coffee has mean $\mu = 3.5$ and standard deviation $\sigma = 1.2$. The distribution of X is **not** normal. A random sample of $n = 50$ coffees is timed, and $\bar{X}$ is the sample mean time.

a. State the mean and the standard deviation of $\bar{X}$, giving the standard deviation correct to four decimal places. 2 marks

b. Explain why $\bar{X}$ is approximately normally distributed even though X is not, and state the approximate distribution of $\bar{X}$. 2 marks

c. Find the approximate probability that the sample mean time exceeds 3.8 minutes, correct to four decimal places. 2 marks

Question 4 (7 marks)

The fill of a carton of milk is normally distributed with a known standard deviation of $\sigma = 8$ mL. A quality inspector measures a random sample of $n = 25$ cartons and finds a sample mean of $\bar{x} = 1002$ mL.

a. Find a 95% confidence interval for the population mean fill μ, giving the endpoints correct to two decimal places. 3 marks

b. Interpret this confidence interval in the context of the cartons. 1 mark

c. The inspector wants a 95% confidence interval whose margin of error is at most 2 mL. Find the smallest sample size n that achieves this. 3 marks

Question 5 (7 marks)

A tyre maker states that its tyres have a mean tread life of $\mu = 52\,000$ km, and the tread life is known to have standard deviation $\sigma = 4000$ km. A consumer group suspects that the true mean tread life is **less** than the stated value. It tests a random sample of $n = 40$ tyres and finds a sample mean of $\bar{x} = 50\,800$ km. A test is carried out at the 5% level of significance.

a. Write down the null hypothesis H_0 and the alternative hypothesis H_1, and state whether the test is one-tailed or two-tailed. 2 marks

b. Calculate the value of the test statistic z, correct to four decimal places. 2 marks

c. Find the p-value, correct to four decimal places, and state the conclusion of the test in context. 3 marks

Question 6 (7 marks)

A machine fills bags of flour; the fill mass is normally distributed with a known standard deviation of $\sigma = 6$ g. Each hour a random sample of $n = 9$ bags is weighed and the sample mean $\bar{x}$ is used to test $H_0 : \mu = 500$ against $H_1 : \mu > 500$ at the 5% level of significance; H_0 is rejected (and the machine stopped) when $\bar{x}$ is large enough.

a. Explain what a Type I error would mean in this context, and write down its probability. 2 marks

b. Show that H_0 is rejected when $\bar{x} \geq 503.29$ g (correct to two decimal places). 2 marks

c. The machine has in fact drifted so that the true mean fill is $\mu = 504$ g. Find the probability of a Type II error and the power of the test, each correct to four decimal places. 3 marks

End of Topic 5 Test A

Test A — answers and solutions

Q1. (a) $C = 0.96X$ is a linear function aX with $a = 0.96$, so

$$E(C) = 0.96\,E(X) = 0.96(760) = 729.6\,\text{g}, \; sd(C) = |0.96|\sigma = 0.96(15) = 14.4\,\text{g}.$$

(b) The 13 loaves are **independent**, so their variances add:

$$E(T) = 13(760) = 9880\,\text{g}, \; Var(T) = 13(15^2) = 13(225) = 2925, \; sd(T) = \sqrt{2925} \approx 54.0833\,\text{g}.$$

The trainee has treated T as $13X$ — thirteen times a **single** loaf's mass — which would give $Var(13X) = 13^2(225) = 38025$ and $sd = 195$ g. But T is the sum of 13 **different independent** loaves, for which the variances (not the standard deviations) add: $sd(T) = \sqrt{13} \times 15 = \sqrt{2925} \approx 54.08$ g, far smaller than 195 g.

(c) $T \sim N(9880, 2925)$. Standardising, $z = \dfrac{9800 - 9880}{\sqrt{2925}} = \dfrac{-80}{54.0833} = -1.4792$, so

$$P(T < 9800) = P(Z < -1.4792) \approx 0.0695.$$

Q2. (a) For the linear combination $L = 0.09X + 0.06Y$ of independent variables,

$$E(L) = 0.09(40) + 0.06(120) = 3.6 + 7.2 = 10.8\,\text{litres},$$

$$Var(L) = 0.09^2(5^2) + 0.06^2(10^2) = 0.0081(25) + 0.0036(100) = 0.2025 + 0.36 = 0.5625.$$

(b) A linear combination of independent normal variables is normal, so $L \sim N(10.8, 0.5625)$ with $sd(L) = \sqrt{0.5625} = 0.75$. Standardising, $z = \dfrac{12 - 10.8}{0.75} = 1.6$, so

$$P(L > 12) = P(Z > 1.6) = 1 - P(Z \le 1.6) \approx 0.0548.$$

(c) Two independent runs give a total $F = L_1 + L_2$ with $Var(F) = 2(0.5625) = 1.125$, so $sd(F) = \sqrt{1.125} \approx 1.0607$ litres.

Q3. (a) $\acute{X}$ has the same mean as X and standard deviation $\sigma / \sqrt{n}$:

$$E(\acute{X}) = \mu = 3.5\,\text{min}, \; sd(\acute{X}) = \frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{50}} \approx 0.1697\,\text{min}.$$

(b) By the **central limit theorem**, for a large sample ($n = 50$) the sample mean of any distribution with finite mean and variance is approximately normal, regardless of the shape of X. Hence $\acute{X} \approx N(3.5, 0.1697^2)$ (that is, $N(3.5, 1.2^2/50)$).

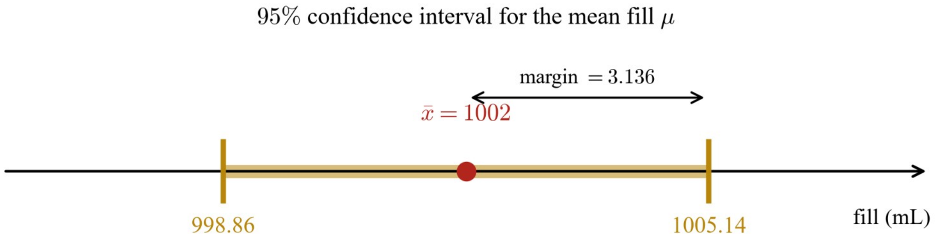
(c) Standardising, $z = \dfrac{3.8 - 3.5}{0.1697} = 1.7678$, so

$$P(\acute{X} > 3.8) \approx P(Z > 1.7678) \approx 0.0385.$$

Q4. (a) With known $\sigma = 8$, $n = 25$ and $z^* = 1.9600$ for 95% confidence, the margin of error is

$$z^* \frac{\sigma}{\sqrt{n}} = 1.9600 \times \frac{8}{\sqrt{25}} = 1.9600 \times 1.6 = 3.136\,\text{mL}.$$

The confidence interval is $\acute{x} \pm 3.136 = 1002 \pm 3.136$, i.e. $(998.86, 1005.14)$ mL.



(b) We are 95 % confident that the interval from 998.86 mL to 1005.14 mL contains the true mean fill μ. Equivalently, if many samples of 25 cartons were taken and a 95 % interval computed from each, about 95 % of those intervals would contain μ.

(c) The margin of error must satisfy $1.9600 \times \frac{8}{\sqrt{n}} \leq 2$, so

$$\sqrt{n} \geq \frac{1.9600 \times 8}{2} = 7.84 \Rightarrow n \geq 7.84^2 = 61.4656.$$

Since n must be a whole number and the margin decreases as n increases, the smallest sample size is $n = 62$.

Q5. (a) The group suspects the mean is **less** than the stated value, so

$$H_0 : \mu = 52\,000\,, H_1 : \mu < 52\,000\,,$$

a **one-tailed** (lower-tail) test.

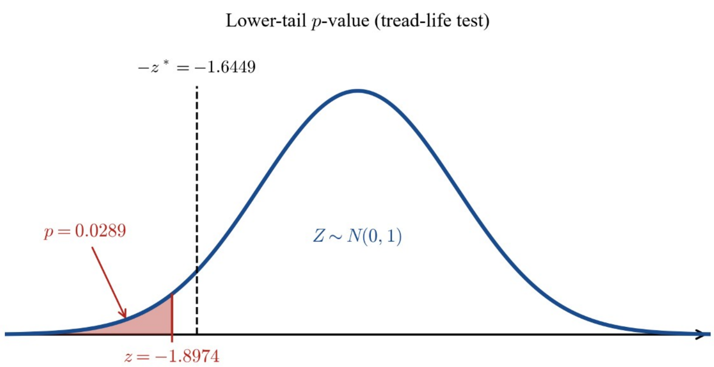
(b) The standard error is $\frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{40}} = 632.4555$, so

$$z = \frac{\acute{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{50\,800 - 52\,000}{632.4555} = \frac{-1200}{632.4555} = -1.8974.$$

(c) For a lower-tail test the p-value is the left-tail area:

$$p = P(Z \leq -1.8974) \approx 0.0289.$$

Since $0.0289 < 0.05$, we **reject** H_0. At the 5 % level there is evidence that the mean tread life is less than the stated 52 000 km.



Q6. (a) A Type I error is rejecting H_0 when it is in fact true — here, **stopping the machine (concluding it is over-filling) when the mean fill is actually** 500 **g**. Its probability equals the significance level, $\alpha = 0.05$.

(b) The standard error is $\frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{9}} = 2$. For an upper-tail test at $\alpha = 0.05$ the critical z-value is $z^* = \Phi^{-1}(0.95) = 1.644854$, so the critical value of the sample mean is

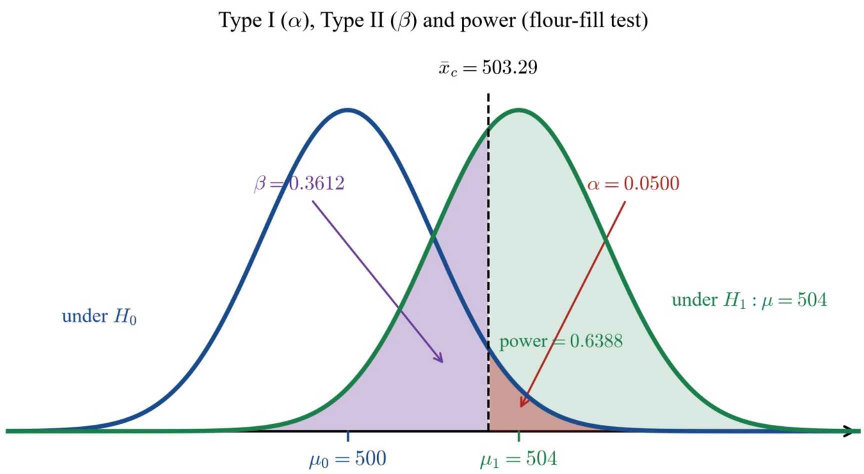
$$\acute{x}_c = \mu_0 + z^* \frac{\sigma}{\sqrt{n}} = 500 + 1.644854 \times 2 = 503.289708 \approx 503.29 \text{ g}.$$

Hence H_0 is rejected when $\acute{x} \geq 503.29$ g.

(c) If the true mean is $\mu = 504$, then $\acute{X} \sim N(504, 2^2)$. A Type II error is failing to reject H_0, i.e. $\acute{X} < \acute{x}_c$. Because the answer is required to four decimal places, carry the unrounded critical value $\acute{x}_c = 503.289708$ from part (b) rather than the rounded 503.29:

$$\beta = P(\acute{X} < 503.289708 / \mu = 504) = P\left(Z < \frac{503.289708 - 504}{2}\right) = P(Z < -0.355146) \approx 0.3612.$$

The power is $1 - \beta \approx 0.6388$ — the probability of correctly detecting the drift to 504 g.



Topic 5 Test B — Technology-free

Technology-free. No calculator. Reading time: 3 minutes. Writing time: 45 minutes. Total 40 marks. In all questions where a numerical answer is required, an exact value must be given. Where more than one mark is available, appropriate working must be shown.

Question 1 (7 marks)

A random variable X has mean $E(X)=4$ and variance $Var(X)=3$.

a. Find $E(2X+5)$ and $Var(2X+5)$. 2 marks

b. Let X_1, X_2, X_3 be three independent observations of X. Find $E(X_1+X_2+X_3)$ and $Var(X_1+X_2+X_3)$. 2 marks

c. Find $Var(3X)$, and explain why $Var(3X)$ is not equal to $Var(X_1+X_2+X_3)$. 3 marks

Question 2 (7 marks)

Two independent normal random variables are $A \sim N(50, 4^2)$ and $B \sim N(30, 3^2)$.

a. Find the distribution of $A+B$, giving its mean and standard deviation. 2 marks

b. Find the distribution of $A-B$, and explain why its standard deviation is the same as that of $A+B$. 2 marks

c. Given that $P(Z<1)=0.84$, where $Z \sim N(0,1)$, find $P(A+B>85)$. 3 marks

Question 3 (6 marks)

The mass of a hen's egg from a free-range farm is normally distributed with mean $\mu=60$ g and standard deviation $\sigma=9$ g. A random sample of $n=9$ eggs is taken, and $\bar{X}$ is the sample mean mass.

a. State the distribution of $\bar{X}$, giving its mean and standard deviation. 2 marks

b. Given that $P(Z \leq 2)=0.9772$, find $P(\bar{X}>66)$. 2 marks

c. State the standard deviation of $\bar{X}$ if the sample size were increased to $n=36$, and describe the effect on the spread of $\bar{X}$. 2 marks

Question 4 (7 marks)

A population is normally distributed with a known standard deviation of $\sigma=15$. A random sample of $n=36$ has sample mean $\bar{x}=82$. Use $z^*=1.96$ for 95% confidence.

a. Find the 95% confidence interval for the population mean μ. 3 marks

b. State the width of this confidence interval. 1 mark

c. Find the smallest sample size n for which a 95% confidence interval would have a margin of error of at most 2.45. 3 marks

Question 5 (7 marks)

A company claims that the mean mass of its chocolate bars is $\mu=100$ g. The mass is normally distributed with a known standard deviation of $\sigma=4$ g. A consumer takes a random sample of $n=64$ bars, finds a sample mean of $\bar{x}=99$ g, and tests, at the 5% level, whether the mean mass **differs** from 100 g.

a. Write down H_0 and H_1. 1 mark

b. Show that the value of the test statistic is $z=-2$. 2 marks

c. The two-tail critical values at the 5% level are ± 1.96. State the conclusion of the test, with a reason. 2 marks

d. Given that $P(Z<2)=0.9772$, find the p-value for this two-tail test and confirm the conclusion. 2 marks

Question 6 (6 marks)

A machine fills bags to a target mean mass of $\mu_0=25$ kg; the fill mass is normally distributed with a known standard deviation of $\sigma=0.6$ kg. To check for over-filling, a random sample of $n=9$ bags is weighed and $H_0:\mu=25$ is tested against $H_1:\mu>25$. The rule is to reject H_0 when the sample mean satisfies $\bar{x}\geq 25.329$ kg.

a. Using the critical value $z^*=1.645$, show that this rejection rule corresponds to a 5% level of significance. 2 marks

b. State, in context, what a Type I error would be here, and write down its probability. 2 marks

c. Suppose the true mean has risen to $\mu=25.529$ kg. Given that $P(Z<1)=0.8413$, find the probability of a Type II error and the power of the test. 2 marks

End of Topic 5 Test B

Test B — answers and solutions

Q1. (a) Using $E(aX+b)=a\mu+b$ and $Var(aX+b)=a^2\sigma^2$ with $a=2, b=5$:

$$E(2X+5)=2(4)+5=13, Var(2X+5)=2^2(3)=12.$$

(b) For independent observations the means add and the variances add:

$$E(X_1+X_2+X_3)=3(4)=12, Var(X_1+X_2+X_3)=3(3)=9.$$

(c) $Var(3X)=3^2 Var(X)=9(3)=27$. This is **not** the same as $Var(X_1+X_2+X_3)=9$. The sum $X_1+X_2+X_3$ is of three **different independent** copies of X, whose variances add to $3\sigma^2=9$. In contrast $3X$ is **one** observation scaled by 3, which multiplies the variance by $3^2=9$, giving $9\sigma^2=27$. Adding independent copies grows the variance in proportion to the count; scaling grows it with the square of the factor.

Q2. (a) A sum of independent normals is normal, with means and variances adding:

$$A+B \sim N(50+30, 4^2+3^2)=N(80,25), sd=\sqrt{25}=5.$$

(b) For a difference of independent normals the mean subtracts but the variances still **add**:

$$A-B \sim N(50-30, 4^2+3^2)=N(20,25), sd=5.$$

The standard deviation equals that of $A+B$ because $Var(A-B)=Var(A)+Var(B)=Var(A+B)$: subtracting an independent variable adds its variance just as adding it does.

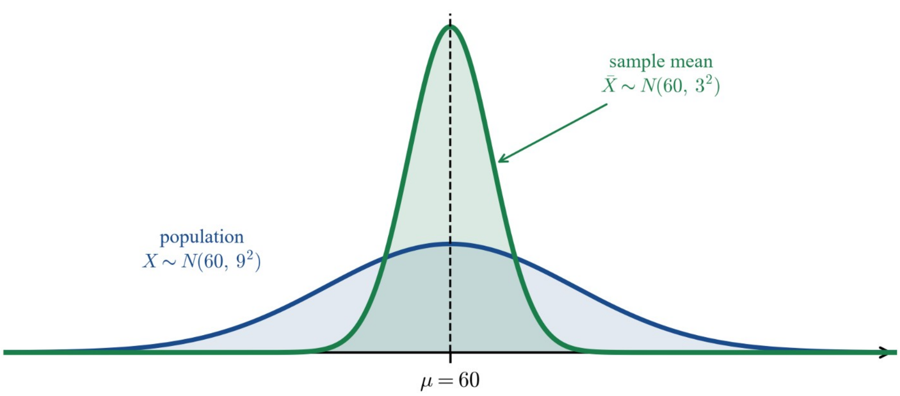
(c) $A+B \sim N(80,5^2)$, so standardising the boundary 85 gives $z=\frac{85-80}{5}=1$. Then

$$P(A+B>85)=P(Z>1)=1-P(Z<1)=1-0.84=0.16.$$

Q3. (a) For a sample mean of a normal population, $\acute{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ with

$$sd(\acute{X})=\frac{\sigma}{\sqrt{n}}=\frac{9}{\sqrt{9}}=\frac{9}{3}=3, \text{so } \acute{X} \sim N(60,3^2).$$

Population vs sampling distribution of $\bar{X}$ $(n=9)$: narrower by $1/\sqrt{n}$



(b) Standardising, $z=\frac{66-60}{3}=2$, so

$$P(\acute{X}>66)=P(Z>2)=1-P(Z\le 2)=1-0.9772=0.0228.$$

(c) With $n=36$, $sd(\bar{X})=\frac{9}{\sqrt{36}}=\frac{9}{6}=1.5$. Since $\sqrt{36}=6$ is twice $\sqrt{9}=3$, the standard deviation is **halved** (from 3 to 1.5): the larger sample makes the sampling distribution of $\bar{X}$ narrower, so $\bar{X}$ is more tightly concentrated about $\mu=60$.

Q4. (a) The standard error is $\frac{\sigma}{\sqrt{n}}=\frac{15}{\sqrt{36}}=\frac{15}{6}=2.5$, so the margin of error is $1.96\times2.5=4.9$. The interval is

$$\bar{x}\pm4.9=82\pm4.9=(77.1,86.9).$$

(b) The width is $2\times4.9=9.8$ (equivalently $86.9-77.1=9.8$).

(c) Require $1.96\times\frac{15}{\sqrt{n}}\le2.45$. Then

$$\sqrt{n}\ge\frac{1.96\times15}{2.45}=\frac{29.4}{2.45}=12\Rightarrow n\ge12^2=144.$$

The smallest sample size is $n=144$.

Q5. (a) "Differs from 100 g" names no direction, so

$$H_0:\mu=100,H_1:\mu\ne100.$$

(b) The standard error is $\frac{\sigma}{\sqrt{n}}=\frac{4}{\sqrt{64}}=\frac{4}{8}=0.5$, so

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}=\frac{99-100}{0.5}=-2.$$

(c) The test is two-tailed with critical values ±1.96. Since $|z|=2>1.96$, the test statistic lies in the rejection region, so we **reject** H_0: there is evidence at the 5% level that the mean mass differs from 100 g.

(d) The single tail beyond $|z|=2$ has area $P(Z>2)=1-0.9772=0.0228$. The two-tail p-value is

$$p=2\times0.0228=0.0456.$$

Since $0.0456<0.05$, we reject H_0, confirming the conclusion in part (c).

Q6. (a) The standard error is $\frac{\sigma}{\sqrt{n}}=\frac{0.6}{\sqrt{9}}=\frac{0.6}{3}=0.2$. Standardising the boundary $\bar{x}=25.329$ gives

$$z=\frac{25.329-25}{0.2}=\frac{0.329}{0.2}=1.645=z^*.$$

For an upper-tail test the area beyond $z^*=1.645$ is 0.05, so the rule rejects H_0 with probability 0.05 when $\mu=25$; the significance level is 5%.

(b) A Type I error is rejecting H_0 when it is true — here, **concluding that the machine is over-filling (mean above 25 kg) when the true mean is actually 25 kg**. Its probability is the significance level, 0.05.

(c) If $\mu=25.529$ then $\bar{X}\sim N(25.529,0.2^2)$. A Type II error is failing to reject H_0, i.e. $\bar{X}<25.329$:

$$\beta=P(\bar{X}<25.329/\mu=25.529)=P\left(Z<\frac{25.329-25.529}{0.2}\right)=P(Z<-1)=1-0.8413=0.1587.$$

The power is $1-\beta=1-0.1587=0.8413$.