

Chapter 7: Confidence Intervals

We begin with the mean-interval construction, then examine interpretation and precision before the first practice problem. The same ideas extend to independent groups and proportions, followed by sample-size planning and finite-population sampling. Three problems are included without solutions; the office-hours QR is a final appendix.

1. Parameters, estimators, and estimates

A **parameter** describes a population and is fixed, usually unknown. A **statistic** is calculated from a sample. A **point estimator** is a statistic used to estimate a parameter; its realized value is a **point estimate**. The estimator varies across repeated samples even though the parameter does not.

Target	Parameter	Point estimator
One population mean	μ	$\bar{X}$
Difference of two means	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$
One population proportion	p	$\hat{p} = X/n$
Difference of two proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$

Uppercase $X_i, \bar{X}, S$ denotes random quantities before observation; lowercase $x_i, \bar{x}, s$ denotes observed values. The symbol $\hat{p}$ is conventionally used both for the random sample proportion and its observed value; context determines which is meant. Write $q = 1 - p$ and $\hat{q} = 1 - \hat{p}$.

2. Chapter 6 facts used here

For independent, identically distributed observations with finite mean μ and variance $0 < \sigma^2 < \infty$,

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

These moments are exact. The **standard error** is the standard deviation of an estimator's sampling distribution. It measures variation among estimates, whereas σ measures variation among individual observations. The slides also write variance as V ; here $V = \text{Var}$.

Our normal notation is $N(\text{mean}, \text{standard deviation})$. For sufficiently accurate CLT approximation,

$$\bar{X} \overset{\text{approx}}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right), \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{\text{approx}}{\sim} N(0, 1).$$

For iid normal observations, the sample mean is exactly normal for every n , whether or not σ is known. A known σ makes the displayed standardized statistic computable. For a general population they are approximations; $n \geq 30$ is a course guideline, not a guarantee for arbitrary skewness or heavy tails. A large sample does not repair biased sampling or dependence.

3. What is new in Chapter 7?

Chapter 6 uses specified population parameters to predict sample behavior. Chapter 7 uses observed sample behavior to estimate unknown population parameters. The new step is to turn a sampling-distribution probability statement into an interval for the parameter, then quantify the method's coverage and precision.

Finite populations require a variance correction when sampling without replacement. Unknown population SDs require estimated standard errors. The same interval logic then applies to all four targets above.

Critical Values and Confidence Intervals

4. A central normal probability

For $Z \sim N(0, 1)$, keep density and cumulative probability distinct:

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

The density f is a height; Φ is an area. For $0 < \alpha < 1$, the course's **upper-tail** notation is

$$\Pr(Z > z_{\alpha/2}) = \frac{\alpha}{2}, \quad \Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}, \quad \Pr(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$

The confidence level is $100(1 - \alpha)\%$; total tail area α is split equally between two tails. Do not use z_α for a two-sided interval.

Confidence	90%	95%	98%	99%
$\alpha/2$	.05	.025	.01	.005
$z_{\alpha/2}$	1.644854	1.959964	2.326348	2.575829
Usual rounded value	1.645	1.96	2.33	2.576 (or 2.58)

For a normal observation, $\Pr(\mu - z_{\alpha/2}\sigma \leq X \leq \mu + z_{\alpha/2}\sigma) = 1 - \alpha$. This is an interval for an observation. To estimate μ using an average, replace the individual SD by the average's standard error.

5. Derive a confidence interval for a mean

Derivation. For iid observations and known σ , start with

$$\Pr\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha.$$

Write $c = z_{\alpha/2}\sigma/\sqrt{n} > 0$. The same event can be rearranged:

$$-c \leq \bar{X} - \mu \leq c \iff \bar{X} - c \leq \mu \leq \bar{X} + c.$$

Before sampling, the interval endpoints are random, so

$$\Pr\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha.$$

After observing $\bar{x}$, report

$$CI_{100(1-\alpha)\%}(\mu) = \left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

Coverage is exact for iid normal observations with known σ and exact critical values; otherwise it relies on an adequate normal approximation. For substantial finite-population sampling without replacement, use the corrected SE from the finite-population section and separately justify any normal approximation.

6. Center, margin of error, and width

The common Chapter 7 form is

$$\text{point estimate} \pm \underbrace{z_{\alpha/2} \times (\text{standard error})}_{\text{margin of error ME}}.$$

The full width is 2 ME, not ME. The endpoints have the same units as the parameter.

Confidence and Precision

7. What a confidence level means

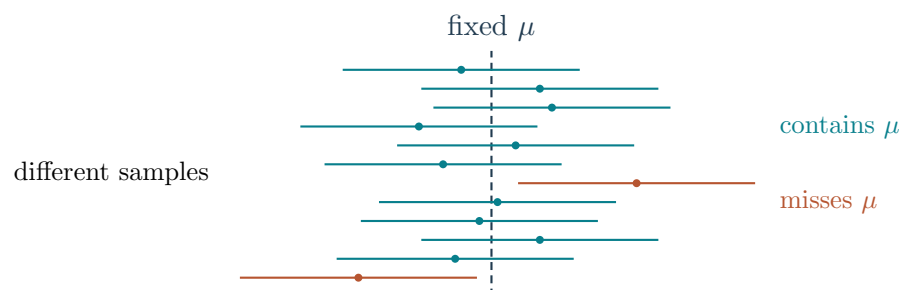
Imagine repeatedly taking random samples of the same size and constructing a confidence interval by the same method. A method with 95% coverage produces intervals that contain the **fixed population parameter** in 95% of such repetitions. For the large-sample intervals here, that coverage is approximate.

The interval changes from sample to sample. The parameter does not. Once the sample is observed, the endpoints are fixed and the interval either contains the parameter or misses it. We usually cannot tell which has happened. The confidence level describes the procedure, not a new probability distribution for the parameter.

An appropriate report is: “We are approximately 95% confident that the population mean is between L and U [units].” Specify the population and parameter. For a difference, specify which population is subtracted from which.

8. Common statements that do not follow

- A mean CI does **not** contain 95% of individual observations. It estimates a mean; individual variation uses σ , not $\sigma/\sqrt{n}$.
- It does **not** say 95% of sample means lie in the particular observed interval. A sampling-distribution probability and a confidence statement answer different questions.
- In this frequentist construction, an observed interval does **not** assign 95% probability to the fixed unknown parameter.
- A single interval cannot be guaranteed to contain the parameter. Nor must exactly 95 of the next 100 intervals contain it; finite repetitions fluctuate.
- A narrow interval does not rule out selection bias, measurement bias, or unmodeled dependence.



Each horizontal segment is one interval from a different sample; the dot is its estimate. This illustrative sketch is not calibrated to show exactly 95% successes. Coverage describes the long-run procedure; an observed interval either contains the fixed parameter or does not.

How Wide Is the Interval?

9. What controls the width?

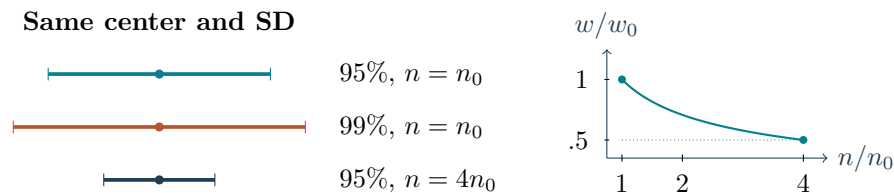
For a mean interval with a fixed SD value,

$$\text{ME} = z_{\alpha/2} \frac{\text{SD}}{\sqrt{n}}, \quad \text{width} = 2z_{\alpha/2} \frac{\text{SD}}{\sqrt{n}}.$$

Change, with other quantities fixed	Consequence
Higher confidence	Smaller tail area; larger critical value; wider interval.
Larger sample size	Smaller SE; narrower interval. Multiplying n by k divides width by $\sqrt{k}$.
Larger population or sample SD	Larger SE; wider mean interval.

To halve a usual mean-interval width at the same confidence and SD, use four times the sample size. A new sample can change $\bar{x}$ and s , so its center and realized width need not follow a fixed numerical prediction.

Increasing confidence trades precision for coverage. For an unrestricted mean, the trivial interval $\mathbb{R}$ has 100% coverage but provides no localization. This is why ordinary confidence intervals accept some risk of missing the parameter in exchange for useful finite width.



Left: wider confidence requires a larger critical value; more data reduce the planned width. The centers are aligned only to isolate width. Right: at fixed confidence and SD, relative width is $w/w_0 = \sqrt{n_0/n}$, so four times the sample size halves it. These are comparisons, not predictions of the center or SD in a future sample.

Practice 1

A mean, a margin, and a width

A laboratory takes a random sample of $n = 64$ independent water specimens from a large population. The sample mean concentration is $\bar{x} = 12.4$ mg/L and the population standard deviation is known to be $\sigma = 3.2$ mg/L. Assume the large-sample normal approximation for the sample mean is adequate. These data are illustrative.

Use $z_{0.025} = 1.96$ for 95% confidence and $z_{0.005} = 2.576$ for 99% confidence.

- (a) Define the population parameter being estimated. Construct its approximate 95% confidence interval. Identify the standard error, margin of error, and *full width*, including their units.
- (b) Interpret the interval and explain what the 95% confidence level means over repeated samples. A classmate says, “About 95% of individual water specimens have concentrations in this interval.” Is that what a confidence interval for a mean says? Explain.
- (c) For planning, suppose a future sample from the same population has $n = 256$, with $\sigma = 3.2$ mg/L still known. At 95% confidence, find its margin of error and full width. Compare both with part (a). Can you determine the future interval’s endpoints from this information alone?
- (d) Return to the *original* sample of 64 specimens. Construct the approximate 99% confidence interval. Explain why increasing the confidence level changes the width in this direction. Does changing the confidence level change the observed center?

Estimated Standard Error

10. Estimating variability

When σ is unknown, estimate its variance using

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

For iid observations with finite variance, $E[S^2] = \sigma^2$: the estimator is **unbiased**. This is a statement about its average over repeated samples, not about each observed s^2 . The complete derivation appears in the final reference section before the fall-break problems.

Intuition. Estimating the mean from the same observations uses one degree of freedom: the deviations from $\bar{x}$ sum to zero. The correction $n - 1$ compensates for the tendency of deviations about a fitted center to be too small. Although S^2 is unbiased, S is not generally unbiased for σ .

11. The large-sample interval when σ is unknown

Estimate the standard error by $\widehat{SE}(\bar{X}) = s/\sqrt{n}$. For a sufficiently large iid sample with finite variance and an adequate approximation,

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \stackrel{\text{approx}}{\sim} N(0, 1), \quad \boxed{CI_{100(1-\alpha)\%}(\mu) \approx \bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}.}$$

Here approximation refers to coverage. Replacing σ with S does **not** give an exactly standard normal statistic at $n = 30$ or at any finite threshold. The course's large-sample rule is a practical approximation. Small-sample methods with unknown variance are beyond this chapter's normal-interval formulas.

12. A mean-interval calculation from the slides

The slide data are $n = 100$, $\bar{x} = 7.56$ grams/day and $s = 35$ grams/day. Assuming representative independent sampling and an adequate normal approximation,

$$\widehat{SE} = \frac{35}{\sqrt{100}} = 3.5, \quad ME = 1.96(3.5) = 6.86, \quad CI_{95\%} = [0.70, 14.42] \text{ grams/day}.$$

This estimates the population *mean*, not the range of individual intakes. The sample size is 100 throughout. In practice, substantial skewness can make the approximation questionable even with 100 observations; the formula requires more than a sample-size check.

13. Report the target, units, and uncertainty

State which population mean is being estimated, give both endpoints with units, and name the confidence level. The sample mean is the center; the estimated standard error measures sampling variation; the margin of error is the critical value times that SE. The full interval width is twice the margin.

For example, the interval above estimates a population *mean* intake. Its endpoints are not limits on individual intakes. A negative endpoint for a physically nonnegative quantity can signal a poor normal approximation or limited precision; the formula alone does not repair the underlying model.

Difference of Two Population Means

14. Define the contrast and sampling design

Let population j have mean μ_j and variance σ_j^2 , and draw independent random samples of sizes n_1, n_2 . Observations within each group must satisfy the applicable iid or justified sampling assumptions. The target and estimator are

$$\Delta = \mu_1 - \mu_2, \quad D = \bar{X}_1 - \bar{X}_2.$$

State this order first. $\Delta > 0$ means population 1 has the higher population mean; a positive observed D alone does not establish this. Reversing the order changes an interval $[L, U]$ to $[-U, -L]$.

15. Derive its mean and standard error

Linearity gives $E[D] = \mu_1 - \mu_2$. In general,

$$\text{Var}(D) = \text{Var}(\bar{X}_1) + \text{Var}(\bar{X}_2) - 2 \text{Cov}(\bar{X}_1, \bar{X}_2).$$

Independent groups have covariance zero, so

$$\text{Var}(D) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}, \quad \boxed{\text{SE}(D) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Variances add even though the means are subtracted. Standard deviations do not add. Equal sample sizes or equal population variances are *not* required.

For independent normal samples this difference is exactly normal. For nonnormal populations, both sample means need adequate normal approximations:

$$D \stackrel{\text{approx}}{\sim} N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

The course uses $n_1, n_2 \geq 30$ as a guideline; one large group does not automatically compensate for a poorly approximated small group.

16. Construct the interval

Standardize D about Δ , place the standardized quantity between $\pm z_{\alpha/2}$, and rearrange just as for one mean. With known population SDs,

$$\boxed{(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

With unknown SDs and adequate large samples, use the **unpooled** estimated standard error:

$$\widehat{\text{SE}}(D) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad \boxed{(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \widehat{\text{SE}}(D)}.$$

The unknown-SD normal interval has approximate coverage. Each group's SD must be squared, divided by its own sample size, and then included under the *same* square root.

17. Interpret the result and check independence

An interval entirely above zero supports a positive population-mean difference at the stated confidence level; entirely below zero supports a negative difference. If it contains zero, equality is compatible with the interval, but is not established. Its endpoints describe the magnitude and units of the difference.

Before/after measurements on the same individuals are paired: the covariance need not vanish. Do not apply the independent-groups formula automatically, or replace a difference CI with two separate mean CIs.

Sampling Distribution of a Proportion

Recall the discrete-moment formulas: for a random variable Y with PMF f , $\mu = E[Y] = \sum_y yf(y)$, $E[Y^2] = \sum_y y^2 f(y)$, and $\text{Var}(Y) = E[Y^2] - \mu^2$, provided these moments exist. These apply to both binary outcomes and general discrete payouts or displacements.

18. A sample proportion is an average

For a binary characteristic, let $I_i = 1$ when sampled individual i has the characteristic and $I_i = 0$ otherwise. With independent observations and common success probability p ,

$$I_i \sim \text{Bernoulli}(p), \quad X = \sum_{i=1}^n I_i \sim \text{Bin}(n, p), \quad \hat{p} = \frac{X}{n} = \frac{1}{n} \sum_{i=1}^n I_i.$$

The target p is a population proportion, the count X is an integer, and $\hat{p}$ lies on the grid $0, 1/n, 2/n, \dots, 1$.

By scaling the binomial moments,

$$E[\hat{p}] = \frac{np}{n} = p, \quad \text{Var}(\hat{p}) = \frac{npq}{n^2} = \frac{pq}{n}, \quad \boxed{\text{SE}(\hat{p}) = \sqrt{\frac{pq}{n}}}, \quad q = 1 - p.$$

This is the sample-mean result applied to zero-one observations. For a simple random sample without replacement from a finite population, the count is hypergeometric rather than exactly binomial, and

$$\text{Var}(\hat{p}) = \frac{pq}{n} \frac{N - n}{N - 1}.$$

The binomial approximation is appropriate when the sampling fraction is small.

19. The normal approximation and its conditions

If both expected counts np and nq are sufficiently large,

$$\hat{p} \overset{\text{approx}}{\sim} N\left(p, \sqrt{\frac{pq}{n}}\right), \quad \frac{\hat{p} - p}{\sqrt{pq/n}} \overset{\text{approx}}{\sim} N(0, 1).$$

The slides use $np > 5$ and $nq > 5$ as a guideline. Both must be checked; a large n alone is insufficient if p is near 0 or 1. This is an approximation to a discrete distribution, not a claim that a proportion can truly lie outside $[0, 1]$.

20. A probability question uses the specified p

When p is given, use it in the center, SE, and expected-count checks. For example,

$$\Pr(a \leq \hat{p} \leq b) \approx \Phi\left(\frac{b - p}{\sqrt{pq/n}}\right) - \Phi\left(\frac{a - p}{\sqrt{pq/n}}\right).$$

In the slides' illustrative blue-eyes model, $p = .09$ and $n = 613$. Then $np = 55.17$, $nq = 557.83$, and

$$\text{SE}(\hat{p}) = \sqrt{\frac{(.09)(.91)}{613}} \approx .011559.$$

The direct normal approximation used in the slides gives

$$\Pr(\hat{p} \geq .11) \approx 1 - \Phi\left(\frac{.11 - .09}{.011559}\right) \approx \boxed{.0418}.$$

For an exact count calculation the event is $X \geq \lceil 613(.11) \rceil = 68$. A Chapter 6 continuity-corrected normal approximation would use boundary 67.5 on the count scale. State which approximation you use; do not mix a rounded SE with an unrelated z -score. The population percentage here is an assumed exercise value, not a verified demographic claim.

Confidence Interval for One Proportion

21. Estimation uses observed counts

To estimate an unknown population proportion p , observe x successes in a random sample of size n . Compute

$$\hat{p} = \frac{x}{n}, \quad \hat{q} = 1 - \hat{p}, \quad \widehat{\text{SE}}(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

The true SE depends on unknown p ; replacing it with $\hat{p}$ gives an estimated SE. This is different from a probability problem where p is specified.

22. Derive and apply the large-sample interval

Derivation. The approximate standardized quantity is

$$\frac{\hat{p} - p}{\sqrt{\hat{p}\hat{q}/n}} \stackrel{\text{approx}}{\sim} N(0, 1).$$

Place it between $-z_{\alpha/2}$ and $z_{\alpha/2}$ and rearrange for p . The Chapter 7 normal interval is

$$CI_{100(1-\alpha)\%}(p) \approx \left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}, \quad \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right].$$

This is often called a large-sample normal or Wald interval. Its margin of error is $z_{\alpha/2} \sqrt{\hat{p}\hat{q}/n}$ and its width is twice that value. Unlike a binomial tail approximation, this course CI formula does not insert a 0.5 continuity correction.

23. A calculation workflow

1. Define the target population and what counts as a success. Verify that the observations and the sampling method justify the model.
2. Record x and n . Compute $\hat{p} = x/n$ and $\hat{q} = 1 - \hat{p}$.
3. Check the course conditions $n\hat{p} = x > 5$ and $n\hat{q} = n - x > 5$. Also check independence or a small sampling fraction.
4. From the requested confidence level find α , then use $z_{\alpha/2}$, not z_{α} .
5. Calculate estimated SE, then margin of error, then both endpoints. Retain intermediate precision and round the reported interval sensibly.
6. Interpret the interval as estimating the *population proportion*. Convert proportions to percentages if useful, multiplying both endpoints by 100.

24. Interpretation and limits

A 95% procedure has approximately 95% repeated-sampling coverage under its assumptions. It does not say that 95% of sampled individuals possess the characteristic. Nor does the confidence level become a probability assigned to p after observing the interval.

An interval such as $[\cdot40, \cdot46]$ means 40% to 46% of the population, with a margin of error of 3 *percentage points*. A percentage-point difference is not a relative percentage change.

The success/failure guideline is not a guarantee of accurate coverage. With rare outcomes, small counts, or proportions near 0 or 1, the normal interval can perform poorly or extend outside $[0, 1]$. Truncating an endpoint does not by itself fix the method. Such cases require a different justified method rather than mechanically applying this chapter's approximation.

Difference of Two Population Proportions

25. Two independent groups

Let x_j successes occur among n_j sampled individuals in group j , with $\hat{p}_j = x_j/n_j$ and $\hat{q}_j = 1 - \hat{p}_j$. The target is $\Delta = p_1 - p_2$ and the estimator is $\hat{p}_1 - \hat{p}_2$. For independent groups,

$$E[\hat{p}_1 - \hat{p}_2] = p_1 - p_2, \quad \text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}.$$

Each group's proportion is a Bernoulli sample mean. The variances add because the groups are independent. For a probability calculation with specified p_1, p_2 , use the true SE:

$$\hat{p}_1 - \hat{p}_2 \stackrel{\text{approx}}{\sim} N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right).$$

Check all four expected counts $n_1 p_1, n_1 q_1, n_2 p_2, n_2 q_2 > 5$, as well as the sampling assumptions.

26. Construct the confidence interval

When p_1, p_2 are unknown, estimate them separately. The **unpooled** CI uses

$$\widehat{\text{SE}} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}, \quad \boxed{(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \widehat{\text{SE}}}.$$

Check all four *observed* counts $x_1, n_1 - x_1, x_2, n_2 - x_2 > 5$. Do not pool the proportions into one common estimate: the interval estimates an unknown difference without assuming that it is zero. Unequal sample sizes are allowed.

27. Bond-proposal example from the slides

For independent samples, the developing section has $n_1 = 50$, $\hat{p}_1 = .76$; the rest of the city has $n_2 = 100$, $\hat{p}_2 = .65$. The success/failure counts are 38, 12, 65, 35, all greater than 5. Then

$$\hat{p}_1 - \hat{p}_2 = .11, \quad \widehat{\text{SE}} = \sqrt{\frac{(.76)(.24)}{50} + \frac{(.65)(.35)}{100}} = \sqrt{.005923} \approx .076961.$$

Using the slides' rounded 99% critical value 2.58,

$$CI_{99\%}(p_1 - p_2) \approx .11 \pm 2.58(.076961) \approx \boxed{[-.0886, .3086]}.$$

We are approximately 99% confident that the developing-section population support proportion minus the rest-of-city proportion is between -8.86 and 30.86 percentage points. Because zero is included, this interval does not establish a direction of difference. It also does not establish equality.

28. Keep design, target, and units straight

Two measurements on the same people do not form independent groups. Large sample sizes do not remove the resulting covariance. A difference of proportions describes an absolute change on the proportion scale: .11 is 11 percentage points, not automatically an 11% relative increase.

The observed contrast alone gives no measure of sampling uncertainty. Its CI combines the contrast with both groups' sample sizes and variation. Reverse the group order by negating the center and switching/negating the endpoints.

Choosing a Sample Size

29. Specify precision before collecting data

Choose a confidence level $100(1 - \alpha)\%$ and a desired **margin of error** $\varepsilon > 0$. If a problem specifies a total interval width w , use $\varepsilon = w/2$. Solve the margin-of-error inequality and round *up* to the next integer. These calculations plan precision under the model; they do not guarantee representative sampling or exact normal coverage.

30. Planning for one population mean

Derivation. For independent sampling with known or planned SD σ ,

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \varepsilon \iff \sqrt{n} \geq \frac{z_{\alpha/2} \sigma}{\varepsilon} \iff \boxed{n \geq \left(\frac{z_{\alpha/2} \sigma}{\varepsilon} \right)^2}.$$

If σ is unknown, use a defensible prior estimate or pilot SD. The resulting sample size is conditional on that planning value; the actual study's estimated SD may differ. If asked for *additional* observations, first find the required *total* n , then subtract the number already collected.

The slides suggest estimating σ by a range divided by 4. This is a rough planning heuristic when the stated range represents approximately the central 95% of a roughly normal population, since that span is about 4σ . A normal population has no finite full range, and an observed minimum-to-maximum range depends on sample size. Arbitrary hard bounds or strongly skewed data do not justify this rule. The slides' illustrative range 17 to 60 gives $(60 - 17)/4 = 10.75$ only as such a rough planning estimate.

31. Planning for one population proportion

Derivation. If a defensible planning value of p is available, use

$$z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \leq \varepsilon \iff \boxed{n \geq \left(\frac{z_{\alpha/2}}{\varepsilon} \right)^2 p(1-p)}.$$

Without prior information, make the conservative choice $p = .5$ because

$$p(1-p) = \frac{1}{4} - \left(p - \frac{1}{2} \right)^2 \leq \frac{1}{4}.$$

Thus the largest possible planned SE, and therefore required sample size, occurs at $p = .5$:

$$\boxed{n \geq \frac{z_{\alpha/2}^2}{4\varepsilon^2}}.$$

This does not assert that the unknown proportion equals .5; it protects the planned margin of error against any value of p within the normal-approximation calculation. A future sample's $\hat{p}$ is unavailable when planning, though a prior or pilot estimate can support a less conservative calculation.

32. The slides' 95% / one-percentage-point target

For 95% confidence and margin of error at most .01, use $z_{.025} \approx 1.96$:

$$n \geq \frac{1.96^2}{4(.01)^2} = \boxed{9604}.$$

The target is one percentage point of margin on either side, hence two percentage points of total width. Retain precision until the final ceiling; if a supplied table uses a rounded critical value, state that convention. The normal approximation and sampling assumptions still need checking in the eventual study.

Finite-Population Sampling

33. Sampling without replacement

Derivation. Consider a fixed population of $N > 1$ values $y_1, \dots, y_N$. Its mean and variance, with divisor N , are

$$\mu = \frac{1}{N} \sum_{j=1}^N y_j, \quad \sigma^2 = \frac{1}{N} \sum_{j=1}^N (y_j - \mu)^2.$$

Take a **simple random sample without replacement** of size $n \leq N$: each subset of size n is equally likely. Each sampled position has mean μ and variance σ^2 , but different positions are not independent. Selecting one value changes which values remain.

Linearity still gives $E[\bar{X}] = \mu$. For distinct positions $i \neq j$,

$$\text{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}.$$

To see the negative sign, sum centered products over all ordered distinct population pairs:

$$\sum_{a \neq b} (y_a - \mu)(y_b - \mu) = \left[\sum_a (y_a - \mu) \right]^2 - \sum_a (y_a - \mu)^2 = -N\sigma^2.$$

Divide by the $N(N-1)$ equally likely ordered pairs. Hence

$$\begin{aligned} \text{Var}(\bar{X}) &= \frac{1}{n^2} \left[n\sigma^2 + n(n-1) \left(-\frac{\sigma^2}{N-1} \right) \right] \\ &= \boxed{\frac{\sigma^2}{n} \frac{N-n}{N-1}}, \quad \text{SE}(\bar{X}) = \boxed{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}. \end{aligned}$$

The factor $(N-n)/(N-1)$ is the **finite population correction to variance**; the SD uses its square root. It is close to 1 when the sampling fraction n/N is small. If $n = N$, the sample is a census, $\bar{X} = \mu$, and the sampling variance is zero. This exact variance formula alone does not imply a normal shape.

34. The small-population example from the slides

For $\{3, 6, 9, 12, 15\}$, $N = 5$, $\mu = 9$, and $\sigma^2 = 18$. Sampling $n = 3$ gives $\binom{5}{3} = 10$ equally likely subsets. Their averages are

Sample	(3, 6, 9)	(3, 6, 12)	(3, 6, 15)	(3, 9, 12)	(3, 9, 15)
$\bar{x}$	6	7	8	8	9
Sample	(3, 12, 15)	(6, 9, 12)	(6, 12, 15)	(6, 9, 15)	(9, 12, 15)
$\bar{x}$	10	9	11	10	12

Thus the discrete PMF (also denoted by lowercase f in the slides) is

$\bar{x}$	6	7	8	9	10	11	12
$f(\bar{x})$	.1	.1	.2	.2	.2	.1	.1

It gives $E[\bar{X}] = 9$ and $\text{Var}(\bar{X}) = 3$, agreeing with $(18/3)(2/4) = 3$. Independence would instead give $18/3 = 6$, overestimating the variance.

If the question asks for **ordered** samples without replacement, there are $N!/(N-n)!$ equally likely sequences, rather than $\binom{N}{n}$ subsets. Each subset has $n!$ orders and the same sample mean in every order. Either representation gives the same distribution of $\bar{X}$ when probabilities are counted consistently. List ordered sequences when ordered samples are requested.

Choosing and Checking the Method

35. Formula map for independent large-population sampling

Every interval below has center $\hat{\theta}$ and margin $z_{\alpha/2} \times \text{SE}$. Use true SDs when known; estimated SEs lead to large-sample approximate coverage.

Parameter	Center	Standard error used in the CI
μ , known σ	$\bar{x}$	$\sigma/\sqrt{n}$
μ , unknown σ	$\bar{x}$	$s/\sqrt{n}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{s_1^2/n_1 + s_2^2/n_2}$
p	$\hat{p}$	$\sqrt{\hat{p}(1-\hat{p})/n}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$

For two known population SDs, replace s_1, s_2 by σ_1, σ_2 . For a substantial sampling fraction without replacement, the relevant finite-population variance correction must also be considered.

36. Recognize the question before choosing a formula

- **Sampling probability:** the population parameter is specified; find the chance an estimator lies in an event. Use the true parameter in the sampling mean and SE, standardize, and evaluate Φ .
- **Confidence interval:** the parameter is unknown; data give an estimate and a known or estimated SE. Center the interval at the statistic and interpret coverage for the population parameter.
- **Sample-size design:** confidence and target margin are specified before the new study. Choose a planning variance, solve an inequality, and take a ceiling.
- **Finite-population distribution:** for a small population, enumerate equally likely samples and combine those with the same statistic value. Exact moments need not imply a normal sampling distribution.

The normal intervals in this chapter do not replace later small-sample procedures, and they do not constitute a full hypothesis-testing framework. The sampling distributions, derivations, and conditions determine which result is appropriate.

37. Recover a required or implied sample size

If an interval for a mean uses known or planned SD a , its margin m gives $n = (z_{\alpha/2}a/m)^2$. For a proportion with known estimate $\hat{p}$, it gives $n = z_{\alpha/2}^2 \hat{p}(1-\hat{p})/m^2$. Rounded reported margins may only determine an approximate n . A design requirement $\text{ME} \leq \varepsilon$ instead requires an inequality and an upward integer rounding. A margin alone cannot determine both n and $\hat{p}$.

Final Checks

38. Work backward from an interval or margin of error

For the symmetric normal intervals used here, a reported interval $[L, U]$ gives

$$\hat{\theta} = \frac{L + U}{2}, \quad \text{ME} = \frac{U - L}{2}, \quad \text{SE used} = \frac{\text{ME}}{z_{\alpha/2}}.$$

For a proportion interval with known n and margin m , put $A = n(m/z_{\alpha/2})^2$. Then

$$\hat{p}(1 - \hat{p}) = A \implies \boxed{\hat{p} = \frac{1 \pm \sqrt{1 - 4A}}{2}}.$$

The two roots are complementary and have the same SE. Context, such as whether the observed proportion exceeds one half, may select a root; the margin alone does not. Check $0 \leq A \leq 1/4$ and remember that $n\hat{p}$ must be an integer count. A rounded reported margin generally determines approximate, not exact, roots or counts.

39. Chapter 6 sum tools

For iid X_i with mean μ and variance $0 < \sigma^2 < \infty$, a total $T_n = \sum_i X_i$ satisfies

$$\text{E}[T_n] = n\mu, \quad \text{Var}(T_n) = n\sigma^2, \quad \text{SD}(T_n) = \sqrt{n}\sigma, \quad T_n \overset{\text{approx}}{\sim} N(n\mu, \sqrt{n}\sigma).$$

The moments are exact; the normal shape needs its usual justification. A fixed fee C per trial changes the total to $T_n - nC$: its mean becomes $n(\mu - C)$, while variance remains $n\sigma^2$. Write a gain/loss event before substituting a cutoff, then standardize it. For example,

$$\Pr(T_n - nC \leq a) \approx \Phi\left(\frac{a - n(\mu - C)}{\sigma\sqrt{n}}\right).$$

For a target probability $0 < r < 1$, set the standardized cutoff equal to $\Phi^{-1}(r)$ and solve for the unknown quantity. This is a sampling-probability calculation, not a confidence interval. For independent identically distributed discrete draws, a pair (x, y) has probability $f(x)f(y)$. Sum these probabilities over all pairs with $x + y = t$ to get $\Pr(X_1 + X_2 = t)$; distinct totals need not be equally likely.

40. Checks that prevent common errors

1. State the target parameter, population(s), units, and subtraction order. Do not substitute a count, sample mean, or sample proportion for the unknown parameter itself.
2. Check the sampling design before using independence. Use finite-population correction for substantial sampling without replacement. Distinguish paired from independent samples.
3. Distinguish individual SD, estimator SE, margin of error, and total width. Variances add for independent differences; SDs do not.
4. Use both tails for a two-sided CI. Higher confidence widens an interval; increasing sample size narrows it, holding the other inputs fixed.
5. For proportions, check success and failure counts separately in every group. Use each group's own proportion in a difference CI.
6. Normally retain intermediate precision, but follow any explicit rounding or continuity-correction instructions in the problem. Round required sample sizes upward.

Why the Sample Variance Uses $n - 1$

41. Estimating the population variance

Derivation. For iid observations with finite variance and $n \geq 2$, define

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

The statistic S^2 is an **unbiased estimator** of σ^2 : $E[S^2] = \sigma^2$. Unbiased means that its average value over repeated samples equals the target; it does not mean every estimate equals the target.

The denominator $n - 1$ accounts for estimating the center from the same observations. Algebra gives

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

Using $E[Y^2] = \text{Var}(Y) + (E[Y])^2$,

$$\begin{aligned} E[S^2] &= \frac{1}{n-1} \left[\sum_{i=1}^n E[X_i^2] - n E[\bar{X}^2] \right] \\ &= \frac{1}{n-1} \left[n(\sigma^2 + \mu^2) - n \left(\frac{\sigma^2}{n} + \mu^2 \right) \right] \\ &= \frac{(n-1)\sigma^2}{n-1} = \boxed{\sigma^2}. \end{aligned}$$

Dividing the squared-deviation sum by n instead would have expectation $(n-1)\sigma^2/n$. The unbiasedness result is about S^2 , *not* $S = \sqrt{S^2}$; taking a square root does not preserve expectations. Nevertheless, s is a useful estimate of σ .

42. Do not mix variance conventions

The finite-population section defines a finite-population variance with divisor N . If s^2 estimates the finite-population variance with divisor $N - 1$, the corresponding estimated variance of the sample mean is $(1 - n/N)s^2/n$. State the sampling design and variance convention before applying a finite-population correction.

43. Keep the claims separate

- $E[S^2] = \sigma^2$ is an exact expectation identity under iid sampling with finite variance. It does not require a normal population.
- A normal approximation for $\bar{X}$ is a separate distributional claim. Estimating its SE does not make an approximate result exact.
- For iid normal observations with unknown σ , the studentized mean has a t distribution; those methods are developed in Chapter 9.

Fall-break problem 2

Comparing means and allocating a sample

A laboratory measures protein concentration in cultures produced by two processes. It takes an independent random sample from each process, with a different culture for every observation. The illustrative results are:

	Process 1	Process 2
Sample size	$n_1 = 64$	$n_2 = 100$
Sample mean (mg/mL)	$\bar{x}_1 = 5.9$	$\bar{x}_2 = 5.5$
Sample SD (mg/mL)	$s_1 = 1.6$	$s_2 = 1.5$

Both population means and SDs are unknown. Assume observations are independent within and between samples, both populations are much larger than the samples, and the normal approximations for the sample means are adequate. Use $z_{0.025} = 1.96$.

- Define $\Delta = \mu_1 - \mu_2$ in context and give its point estimate. Starting from independence, write the variance of $\bar{X}_1 - \bar{X}_2$ and calculate its estimated standard error. Why do the variance contributions add although the means are subtracted?
- Construct an approximate 95% confidence interval for Δ . Show its margin of error and endpoints, with units.
- A researcher says, “Process 1 has the larger sample mean, so its population mean must be larger.” Assess the claim using your interval. Does including zero establish equality? Explain what 95% confidence describes.
- The laboratory can measure 36 additional independent cultures, all from either Process 1 or Process 2. For planning, hold the SD estimates at 1.6 and 1.5. Calculate the anticipated standard error under each option and choose the option with the smaller anticipated margin of error. Explain why this does not determine the future interval’s center.

Fall-break problem 3

An unknown rate and a specified probability

A monitoring program calls a returned water sample *usable* if it meets a prespecified handling checklist. In a random sample of 240 returned kits, 168 samples are usable. Let p be the population proportion of usable samples among this program's returned kits. Assume independent Bernoulli outcomes with a common probability p , and a population much larger than the sample. These data are illustrative.

Use $z_{0.025} = 1.96$ and let Φ denote the standard normal CDF.

- (a) Estimate p and check the observed success and failure counts. Construct an approximate 95% confidence interval for p , showing the estimated standard error and margin of error. Interpret the endpoints as percentages and identify the population they describe.
- (b) Now consider a *separate hypothetical model* in which the population proportion is known to be $p = 0.75$. For a new independent sample of $n = 240$, find the approximate probability $\Pr(\hat{p} \leq 0.70)$. Check the expected counts, standardize, and use a direct normal approximation without continuity correction.
- (c) Explain why the standard errors in parts (a) and (b) use different proportions. Which question estimates a fixed unknown parameter, and which assigns a probability to a random sample result?
- (d) *Think ahead using the sample-size planning section.* A new program wants a planned 95% normal-interval margin of error of at most 4 percentage points. No reliable advance estimate of its population proportion is available. Find the smallest integer sample size that meets this target under the worst-case choice of p . Justify that choice and state whether your answer is a required *total* or a number to add to 240.

Office-hours availability

Help choose a regular time that works for you.

Use your **full name**.

Mark **every time** you can regularly attend
in a typical week, using **Eastern Time**.



Scan the code or open:

<https://www.when2meet.com/?39188547-qtuxi>

I will compare the class's overlap with my availability.

Names and availability are visible to people with the poll link.