Two-Sample Hypothesis Tests for the Difference of Means

AP Statistics · Mr. Merrick · February 22, 2026

We compare two population means by testing a claim about the parameter $\mu_1 - \mu_2$.
From two independent random samples (or two randomized groups), we compute:

$$\overline{x}_1,\ s_1,\ n_1 \qquad \text{and} \qquad \overline{x}_2,\ s_2,\ n_2.$$

For most AP problems, the null value is

$$H_0 : \mu_1 - \mu_2 = 0.$$

1) Two-Sample z test

When to use: σ_1, σ_2 known (rare).

Check conditions:

- Random: each sample is from a random sample or randomized experiment.
- Independence:
  - within groups: $n_1 \le 0.10N_1$ and $n_2 \le 0.10N_2$ (if sampling w/o replacement)
  - between groups: the two samples/groups are independent
- Normal/Large Sample: each population is Normal, or each n is large enough for CLT (≥ 30).

Test statistic (for $H_0 : \mu_1 - \mu_2 = \Delta_0$):

$$\boxed{Z = \frac{(\overline{x}_1 - \overline{x}_2) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}$$

Under H_0, $Z \sim N(0,1)$.

2) Two-Sample t test (Welch's)

When to use: σ_1, σ_2 unknown (typical).

Check conditions:

- Random: each sample is random or groups are randomized.
- Independence:
  - within groups: $n_1 \le 0.10N_1$ and $n_2 \le 0.10N_2$ (if sampling w/o replacement)
  - between groups: independent samples (or randomized groups)
- Normal/Large Sample:
  - if n_1 and/or n_2 are small: check each group's sample distribution is roughly symmetric with no outliers
  - if both are large (≥ 30): CLT supports the procedure

Standard error (estimated): $SE = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$

Test statistic (for $H_0 : \mu_1 - \mu_2 = \Delta_0$):

$$\boxed{T = \frac{(\overline{x}_1 - \overline{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$$

with Welch df:

$$df \;\approx\; \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

$$\boxed{\min\{n_1 - 1, n_2 - 1\} \le df \le n_1 + n_2 - 2}$$

3) Enrichment: Two-Sample t test with pooled variance (NOT required for AP)

Big idea: If the population variances are equal ($\sigma_1^2 = \sigma_2^2$), we can pool information from both groups to estimate the common variance.

Important: This is *not* needed for AP Statistics. The AP standard is Welch's two-sample t test.

In more advanced settings, equality of variances might be assessed by:

- comparing sample spreads (s_1 vs. s_2), or sample variances (s_1^2 vs. s_2^2),
- using a formal procedure such as Levene's test (beyond AP).

Pooled standard deviation:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \qquad s_p = \sqrt{s_p^2}.$$

Pooled standard error:

$$SE_p = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Pooled test statistic (for $H_0 : \mu_1 - \mu_2 = \Delta_0$):

$$\boxed{T = \frac{(\overline{x}_1 - \overline{x}_2) - \Delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{with} \quad df = n_1 + n_2 - 2}$$

For AP Stats: Use Welch's approximation unless the problem explicitly explores pooling (possible on an investigative task).

Example 1

A manufacturer compares the fill amounts of two machines. From historical calibration, the population standard deviations are known:

$$\sigma_1 = 1.2 \text{ mL}, \qquad \sigma_2 = 1.6 \text{ mL}.$$

A random sample of $n_1 = 40$ bottles from Machine 1 has mean $\overline{x}_1 = 502.3$ mL. A random sample of $n_2 = 35$ bottles from Machine 2 has mean $\overline{x}_2 = 500.9$ mL.
Test, at the $\alpha = 0.05$ level, whether Machine 1 fills *more* on average than Machine 2.

Solution. Step 1 — State

Let μ_1 be the true mean fill amount (in mL) for all bottles produced by Machine 1, and let μ_2 be the true mean fill amount (in mL) for all bottles produced by Machine 2.

We will test the hypotheses:

$$H_0 : \mu_1 - \mu_2 = 0 \qquad \text{vs.} \qquad H_a : \mu_1 - \mu_2 > 0.$$

Step 2 — Justify

We will use a two-sample z test for $\mu_1 - \mu_2$ because the population standard deviations are known ($\sigma_1 = 1.2$ mL and $\sigma_2 = 1.6$ mL).

- **Random:** The problem states that a random sample of $n_1 = 40$ bottles was selected from Machine 1 and a random sample of $n_2 = 35$ bottles was selected from Machine 2.
- **Independent:**
  - **Within groups:** Because the production runs are much larger than 40 and 35 bottles, it is reasonable to assume $n_1 = 40 \leq 0.10N_1$ and $n_2 = 35 \leq 0.10N_2$ (10% condition), so observations within each sample are approximately independent.
  - **Between groups:** Bottles from Machine 1 are produced separately from bottles from Machine 2, and the samples were taken independently, so the two samples are independent of each other.
- **Normal/Large Sample:** Both sample sizes are large ($n_1 = 40 \geq 30$ and $n_2 = 35 \geq 30$), so by the Central Limit Theorem the sampling distribution of $\overline{x}_1 - \overline{x}_2$ is approximately Normal.

Since the conditions are met, a two-sample z test is appropriate.

Step 3 — Carry Out

Test statistic:

$$Z = \frac{(\overline{x}_1 - \overline{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$\overline{x}_1 - \overline{x}_2 = 502.3 - 500.9 = 1.4$$

$$SE = \sqrt{\frac{1.2^2}{40} + \frac{1.6^2}{35}} = \sqrt{0.0360 + 0.0731} = \sqrt{0.1091} \approx 0.330$$

$$Z = \frac{1.4}{0.330} \approx 4.24$$

p-value (right-tailed):

$$p = P(Z \geq 4.24) \approx 0.00001 \text{ (essentially 0)}.$$

Step 4 — Conclude

Because $p < \alpha = 0.05$, we reject H_0.

There is convincing evidence that Machine 1 has a higher true mean fill amount than Machine 2.

Example 2

A school compares weekly study time for students in two different programs. Two independent random samples are taken.

Group	n	$\overline{x}$ (hours)	s (hours)
Program A	18	6.8	1.9
Program B	14	5.4	2.3

Test, at the $\alpha = 0.10$ level, whether the true mean weekly study time differs between the two programs.

Solution. Step 1 — State

Let μ_A be the true mean weekly study time (in hours) for all students in Program A, and let μ_B be the true mean weekly study time (in hours) for all students in Program B.

We will test:

$$H_0 : \mu_A - \mu_B = 0 \qquad \text{vs.} \qquad H_a : \mu_A - \mu_B \neq 0.$$

Step 2 — Justify

We will use a two-sample t test for $\mu_A - \mu_B$ using Welch's approximation because the population standard deviations are not known and we use s_A and s_B as estimates.

- **Random:** The problem states that two independent random samples of students were taken from the two programs.
- **Independent:**
  - **Within groups:** Each program contains many more than 18 and 14 students, so it is reasonable to assume $n_A = 18 \leq 0.10N_A$ and $n_B = 14 \leq 0.10N_B$ (10% condition). Thus observations within each sample are approximately independent.
  - **Between groups:** The samples are from two different programs and were selected independently, so the two samples are independent.
- **Normal/Large Sample:** The sample sizes are below 30 ($n_A = 18$ and $n_B = 14$), so we cannot rely on the CLT alone. We would check that each group's sample distribution is roughly symmetric with no outliers (using histograms/boxplots or a graph provided). Assuming no strong skewness and no outliers, Normality is reasonable.

Since the conditions are met, a Welch two-sample t test is appropriate.

Step 3 — Carry Out

$$T = \frac{(\overline{x}_A - \overline{x}_B) - 0}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

$$\overline{x}_A - \overline{x}_B = 6.8 - 5.4 = 1.4$$

$$SE = \sqrt{\frac{1.9^2}{18} + \frac{2.3^2}{14}} = \sqrt{0.2006 + 0.3779} = \sqrt{0.5785} \approx 0.7606$$

$$T = \frac{1.4}{0.7606} \approx 1.84$$

Welch df:

$$df \approx 25.1 \;\Rightarrow\; df \approx 25.$$

Two-sided p-value:

$$p = 2P\left(t_{25} \geq 1.84\right) \approx 0.078.$$

Step 4 — Conclude

Because $p = 0.078 < \alpha = 0.10$, we reject H_0.

There is convincing evidence that the true mean weekly study time differs between Program A and Program B.

Example 3

A nutritionist compares sodium content (mg) for two brands of soup. Independent samples are taken.

Brand	n	$\overline{x}$ (mg)	s (mg)
Brand 1	10	710	48
Brand 2	9	742	55

Test, at the $\alpha = 0.05$ level, whether Brand 1 has a *lower* true mean sodium content than Brand 2.

Solution. Step 1 — State

Let μ_1 be the true mean sodium content (in mg) for all cans of Brand 1 soup, and let μ_2 be the true mean sodium content (in mg) for all cans of Brand 2 soup.

We will test:

$$H_0 : \mu_1 - \mu_2 = 0 \qquad \text{vs.} \qquad H_a : \mu_1 - \mu_2 < 0.$$

Step 2 — Justify

We will use a Welch two-sample t test because the population standard deviations are not known and we use s_1 and s_2 as estimates.

- **Random:** The problem does not explicitly state random sampling or random assignment. To proceed with inference, we must assume each sample is a random sample from its population. Without randomization, conclusions may not generalize.
- **Independent:**
  - **Within groups:** It is reasonable to assume each brand's production is much larger than the sample sizes, so $n_1 = 10 \leq 0.10N_1$ and $n_2 = 9 \leq 0.10N_2$ (10% condition). Thus observations within each sample are approximately independent.
  - **Between groups:** The samples come from two different brands and were taken independently, so the two samples are independent.
- **Normal/Large Sample:** Both sample sizes are small ($n_1 = 10$ and $n_2 = 9$), so we cannot rely on the CLT. We would examine each sample's distribution (histograms/boxplots or a graph provided) to verify they are roughly symmetric with no outliers and not strongly skewed. Assuming this is true, Normality is reasonable.

Since the conditions are met, a Welch two-sample t test is appropriate.

Step 3 — Carry Out

$$T = \frac{(\overline{x}_1 - \overline{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\overline{x}_1 - \overline{x}_2 = 710 - 742 = -32$$

$$SE = \sqrt{\frac{48^2}{10} + \frac{55^2}{9}} = \sqrt{230.4 + 336.1} = \sqrt{566.5} \approx 23.80$$

$$T = \frac{-32}{23.80} \approx -1.35$$

Welch df:

$$df \approx 16.0 \;\Rightarrow\; df \approx 16.$$

Left-tailed p-value:

$$p = P\left(t_{16} \leq -1.35\right) \approx 0.098.$$

Step 4 — Conclude
Because $p = 0.098 > \alpha = 0.05$, we fail to reject H_0.
There is not convincing evidence that Brand 1 has a lower true mean sodium content than Brand 2.

Example 4 (Enrichment: pooled t test)

(Not needed for AP.) A researcher believes two populations have equal variances. Independent samples produce:

Group	n	$\overline{x}$	s
Group 1	22	15.2	3.1
Group 2	20	12.9	2.9

Test, at the $\alpha = 0.05$ level, whether Group 1 has a higher true mean than Group 2.

Solution. Step 1 — State

Let μ_1 be the true mean response for the population represented by Group 1 and let μ_2 be the true mean response for the population represented by Group 2.

We will test:

$$H_0 : \mu_1 - \mu_2 = 0 \qquad \text{vs.} \qquad H_a : \mu_1 - \mu_2 > 0.$$

Step 2 — Justify

We will use a pooled two-sample t test (enrichment) because the researcher believes the population variances are equal, so it is reasonable (for enrichment purposes) to treat $\sigma_1^2 = \sigma_2^2$ and pool the sample variances.

- **Random:** The problem does not explicitly state random sampling or random assignment. To proceed with inference, we must assume each sample is a random sample from its population. Without randomization, conclusions may not generalize.
- **Independent:**
  - **Within groups:** It is reasonable to assume each population is much larger than the sample sizes, so $n_1 = 22 \le 0.10N_1$ and $n_2 = 20 \le 0.10N_2$ (10% condition). Thus observations within each sample are approximately independent.
  - **Between groups:** The samples were taken independently from two different groups, so the two samples are independent.
- **Normal/Large Sample:** The sample sizes are below 30 ($n_1 = 22$ and $n_2 = 20$), so we would check that each group's sample distribution is roughly symmetric with no outliers (using graphs if provided). Assuming that condition holds, using a t procedure is reasonable.
- **Equal variances (pooling condition):** For a pooled procedure, we additionally need it to be reasonable that $\sigma_1^2 = \sigma_2^2$. The sample standard deviations are close ($s_1 = 3.1$ and $s_2 = 2.9$), which supports the assumption of similar variances (in practice one might also use Levene's test, beyond AP).

Since the conditions are met (including the equal-variances assumption for enrichment), a pooled two-sample t test is appropriate.

Step 3 — Carry Out

Pooled standard deviation:

$$s_p^2 = \frac{(22-1)3.1^2 + (20-1)2.9^2}{22+20-2} = \frac{21(9.61) + 19(8.41)}{40} = \frac{361.60}{40} = 9.04 \qquad \Rightarrow \qquad s_p = \sqrt{9.04} \approx 3.007$$

Test statistic:

$$T = \frac{(\overline{x}_1 - \overline{x}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\overline{x}_1 - \overline{x}_2 = 15.2 - 12.9 = 2.3$$

$$SE_p = s_p\sqrt{\frac{1}{22} + \frac{1}{20}} = 3.007\sqrt{0.09545} \approx 0.929$$

$$T = \frac{2.3}{0.929} \approx 2.48$$

Degrees of freedom:

$$df = n_1 + n_2 - 2 = 40$$

Right-tailed p-value:

$$p = P\left(t_{40} \geq 2.48\right) \approx 0.009.$$

Step 4 — Conclude
Because $p = 0.009 < \alpha = 0.05$, we reject H_0.
There is convincing evidence that Group 1 has a higher true mean than Group 2.