

Chapter 11

Hypothesis Testing for Two Samples

Detailed Solutions

11.1 Basic Concept

11.1 Independent vs. Paired Samples

Problem: Classify scenarios.

Solution:

- (a) **Scenario A (Tires): Paired.** The tires are on the *same* car. Factors like driving style and road conditions affect both tires equally.
- (b) **Scenario B (Salaries): Independent.** The engineers and accountants are distinct groups of people with no direct link between a specific engineer and a specific accountant.
- (c) **Scenario C (Diet): Paired.** Data is from the *same* person (Before vs. After).
- (d) **Scenario D (Concrete): Independent.** The batches are mixed separately; there is no logical pairing between Batch 1 of A and Batch 1 of B.

(a) Paired	(b) Independent	(c) Paired	(d) Independent
------------	-----------------	------------	-----------------

11.2 Choosing the Right Test Statistic

Problem: Choose Z , T_{pooled} , $T_{unpooled}$, T_{paired} .

Solution:

- (a) **Known Variances:** Use the Z-Test.
- (b) **Unknown but Equal Variances:** Use T_{pooled} (Pooled Variance T-test).
- (c) **Unknown and Unequal Variances:** Use $T_{unpooled}$ (Welch's T-test).

(d) **Before/After:** Use T_{paired} (Paired T-test).

- | | | | |
|-------|------------------|----------------------------|------------------|
| (a) Z | (b) T_{pooled} | (c) $T_{unpooled}$ (Welch) | (d) T_{paired} |
|-------|------------------|----------------------------|------------------|

11.2 Intermediate

11.3 Pooled T-Test: Polymer Strength

Problem: Form 1 ($n = 10, \bar{x} = 850, s = 20$) vs Form 2 ($n = 12, \bar{x} = 835, s = 15$). Equal var.

Solution:

(a) **Pooled Variance S_p^2 :**

$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{(9)(20^2) + (11)(15^2)}{10 + 12 - 2} \\ &= \frac{9(400) + 11(225)}{20} \\ &= \frac{3600 + 2475}{20} = \frac{6075}{20} = 303.75 \end{aligned}$$

$$S_p = \sqrt{303.75} \approx 17.428.$$

(b) **T-statistic:**

$$\begin{aligned} T &= \frac{\bar{x}_1 - \bar{x}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{850 - 835}{17.428 \sqrt{\frac{1}{10} + \frac{1}{12}}} \\ &= \frac{15}{17.428 \sqrt{0.1833}} \\ &= \frac{15}{17.428(0.428)} = \frac{15}{7.46} \approx 2.011 \end{aligned}$$

(c) **Decision:** $df = 20$. $\alpha = 0.05$ (One-tailed). Critical $t_{0.05,20} = 1.725$. Since $2.011 > 1.725$, we **Reject** H_0 . Formulation 1 is significantly stronger.

- (a) $S_p^2 = 303.75$
- (b) $T \approx 2.01$
- (c) Reject H_0 .

11.4 Paired T-Test: Algorithm Speed

Problem: Comparison on 5 datasets.

Solution:

(a) **Differences** $d = A - B$: $d = \{0.5, 0.5, 0.5, -0.2, 0.5\}$.

(b) **Statistics:** Mean $\bar{d} = \frac{1.8}{5} = 0.36$. Variance:

$$\begin{aligned}s_d^2 &= \frac{\sum(d_i - \bar{d})^2}{n - 1} \\&= \frac{(0.14^2 \times 4) + (-0.56^2)}{4} \\&= \frac{0.0784 + 0.3136}{4} = \frac{0.392}{4} = 0.098\end{aligned}$$

$$s_d = \sqrt{0.098} \approx 0.313.$$

(c) **Test** ($H_0 : \mu_d = 0$):

$$\begin{aligned}T &= \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{0.36}{0.313/\sqrt{5}} \\&= \frac{0.36}{0.140} \approx 2.57\end{aligned}$$

Critical $t_{0.025,4} = 2.776$ (Two-tailed). Since $2.57 < 2.776$, we **Fail to Reject** H_0 . The difference is not statistically significant at 0.05 (though close).

(d) **Why Paired?** The execution time depends heavily on the dataset complexity (Dataset 3 is slow for both, Dataset 1 is fast for both). Pairing removes this "dataset variability" noise, focusing only on the algorithm difference.

(b) $\bar{d} = 0.36, s_d = 0.31$
 (c) Fail to Reject ($T = 2.57 < 2.78$)

11.5 F-Test: Comparing Variances

Problem: Check $H_0 : \sigma_1^2 = \sigma_2^2$ for Polymer data ($s_1^2 = 400, s_2^2 = 225$).

Solution:

(a) **F-statistic:** Place larger variance on top.

$$F = \frac{s_1^2}{s_2^2} = \frac{400}{225} \approx 1.778$$

(b) **Critical Values:** $df_1 = 9, df_2 = 11, \alpha = 0.10$ (Two-tailed \rightarrow use 0.05 tables).
 $F_{0.05,9,11} \approx 2.90$. $F_{0.95,9,11} = 1/F_{0.05,11,9} \approx 1/3.10 = 0.32$.

(c) **Decision:** $0.32 < 1.778 < 2.90$. The test statistic falls inside the acceptance region.
We **Fail to Reject** H_0 . The assumption of equal variances is valid.

- | |
|--|
| (a) $F = 1.78$ |
| (c) Valid Assumption (Fail to Reject). |

11.6 Two Proportions: Marketing

Problem: A (40/500) vs B (60/600). Test $p_B > p_A$.

Solution:

- (a) **Pooled Proportion:** $\hat{p}_A = 0.08, \hat{p}_B = 0.10$.

$$\hat{p} = \frac{X_A + X_B}{n_A + n_B} = \frac{40 + 60}{500 + 600} = \frac{100}{1100} \approx 0.0909$$

- (b) **Z-statistic:**

$$\begin{aligned} Z &= \frac{\hat{p}_B - \hat{p}_A}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \\ &= \frac{0.10 - 0.08}{\sqrt{0.0909(0.9091)\left(\frac{1}{500} + \frac{1}{600}\right)}} \\ &= \frac{0.02}{\sqrt{0.0826(0.00366)}} = \frac{0.02}{\sqrt{0.000302}} \\ &= \frac{0.02}{0.0174} \approx 1.15 \end{aligned}$$

- (c) **Decision:** $P(Z > 1.15) \approx 0.125$. Since $0.125 > 0.05$, we **Fail to Reject** H_0 . Design B is not significantly better.

- | |
|---|
| (b) $Z = 1.15$ |
| (c) $P \approx 0.125$. Fail to Reject. |

11.3 Challenge

11.7 Derivation of Pooled Variance Estimator

Problem: Prove $E[S_p^2] = \sigma^2$.

Solution:

$$\begin{aligned} E[S_p^2] &= E\left[\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\right] \\ &= \frac{1}{n_1 + n_2 - 2} ((n_1 - 1)E[S_1^2] + (n_2 - 1)E[S_2^2]) \end{aligned}$$

Since sample variances are unbiased ($E[S^2] = \sigma^2$):

$$\begin{aligned} &= \frac{1}{n_1 + n_2 - 2} ((n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2) \\ &= \frac{\sigma^2(n_1 - 1 + n_2 - 1)}{n_1 + n_2 - 2} \\ &= \frac{\sigma^2(n_1 + n_2 - 2)}{n_1 + n_2 - 2} = \sigma^2 \end{aligned}$$

Proven.

11.8 Welch's T-Test (Unequal Variances)

Problem: $n_1 = 10, s_1 = 20$ and $n_2 = 12, s_2 = 5$.

Solution:

(a) **Calculations:** Pooled df = $10 + 12 - 2 = 20$.

Welch's df: Let $v_1 = s_1^2/n_1 = 400/10 = 40$. Let $v_2 = s_2^2/n_2 = 25/12 \approx 2.08$. Numerator = $(40 + 2.08)^2 = 1770.7$. Denominator = $\frac{40^2}{9} + \frac{2.08^2}{11} = 177.7 + 0.39 = 178.1$.

$$\nu = \frac{1770.7}{178.1} \approx 9.94 \rightarrow 9$$

(b) **Comparison:** Welch df (9) is much lower than Pooled df (20). This "penalizes" the test because the variances are very different (20^2 vs 5^2). We are less certain about the combined variance, so the effective sample size is closer to the smaller group (or the group with larger variance), leading to a higher critical t-value and a more conservative test.

- (a) Pooled df=20, Welch df=9
- (b) Reduces df to account for uncertainty due to unequal variances.

11.9 Battery Technology Comparison (Visual)

Problem: Boxplot Analysis.

Solution:

(a) **Visual Interpretation:**

- **Medians:** Supplier B's median (red line) is higher than Supplier A's.
- **IQRs:** Supplier A has a smaller box (less height), indicating less variability (more consistent). Supplier B's box is taller.
- **Overlap:** There is some overlap, but Supplier B's box is mostly shifted upwards.

(b) **Stats:** A: $\bar{x}_A = 252.0, s_A^2 = 18.22$. B: $\bar{x}_B = 265.8, s_B^2 = 28.15$.

(c) **Hypothesis Test:** $H_1 : \mu_B > \mu_A. F = 28.15/18.22 = 1.54. P > 0.05 \implies$
Assume Equal Variances. $S_p^2 = \frac{9(18.22) + 11(28.15)}{20} = 23.68 \implies S_p = 4.87$.

$$T = \frac{265.8 - 252.0}{4.87\sqrt{1/10 + 1/12}} = \frac{13.8}{2.08} \approx 6.63$$

Critical $t_{0.01,20} = 2.528$. Reject H_0 . Supplier B is significantly better.

(d) **99% CI:** $13.8 \pm 2.845(2.08) = 13.8 \pm 5.9 \implies [7.9, 19.7]$. Since lower bound > 0, it confirms the result.

(c) Reject H_0 ($T = 6.63$).

11.10 VR Safety Training (Paired)

Problem: Before/After for 8 workers.

Solution:

(a) **Inappropriate Test:** Independent T-test assumes groups are distinct. Here, scores are correlated (a smart worker scores high on both). We must pair to remove "worker ability" noise.

(b) **Diffs:** $d = \{10, 8, -1, 10, 12, 7, 10, 4\}$. $\bar{d} = 7.5. s_d = 4.408$.

(c) **Test:**

$$T = \frac{7.5}{4.408/\sqrt{8}} = \frac{7.5}{1.558} = 4.81$$

Critical $t_{0.05,7} = 1.895$. Reject H_0 . Training works.

(d) **95% CI:** $7.5 \pm 2.365(1.558) = 7.5 \pm 3.68 = [3.82, 11.18]$. Min expected improvement ≈ 3.8 points.

(c) Reject H_0 ($T = 4.81$).

11.11 E-Commerce A/B Testing

Problem: A (160/2000) vs B (200/2000).

Solution:

(a) $\hat{p}_A = 0.08, \hat{p}_B = 0.10.$

(b) **Z-Test:** Pooled $\hat{p} = 360/4000 = 0.09. SE = \sqrt{0.09(0.91)(2/2000)} = \sqrt{0.0000819} \approx 0.00905.$

$$Z = \frac{0.10 - 0.08}{0.00905} = 2.21$$

$P(Z > 2.21) = 0.0136.$ Since $0.0136 < 0.05,$ Reject $H_0.$ Page B is better.

(c) **CI Difference:** $SE_{diff} = \sqrt{\frac{0.1(0.9)}{2000} + \frac{0.08(0.92)}{2000}} = 0.00904. 0.02 \pm 1.96(0.00904) = 0.02 \pm 0.0177 = [0.0023, 0.0377].$

(d) **Revenue:** Min improvement = 0.0023 (0.23%). Visitors = 100,000. Extra clicks = $100,000 \times 0.0023 = 230.$ Extra Revenue = $230 \times \$5 = \$1,150.$

(b) $P = 0.0136.$ Reject $H_0.$

(d) Min \$1,150 per month.

11.12 Type II Error Derivation

Problem: β for Two Sample Z-test.

Solution:

(a) **Definition:** $\beta = P(\text{Fail to Reject } H_0 \mid H_1 \text{ is true}).$

(b) **Derivation:** Reject if $\bar{X}_1 - \bar{X}_2 > \delta_0 + Z_\alpha \sigma_{diff}.$ Fail to reject if $\bar{X}_1 - \bar{X}_2 \leq \delta_0 + Z_\alpha \sigma_{diff}.$ Under $H_1, (\bar{X}_1 - \bar{X}_2) \sim N(\Delta, \sigma_{diff}^2).$ Standardize using $\Delta:$

$$\begin{aligned} \beta &= P\left(Z < \frac{(\delta_0 + Z_\alpha \sigma_{diff}) - \Delta}{\sigma_{diff}}\right) \\ &= P\left(Z < Z_\alpha - \frac{\Delta - \delta_0}{\sigma_{diff}}\right) \\ &= \Phi\left(Z_\alpha - \frac{\Delta - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) \end{aligned}$$

Proven.

11.4 Application

11.13 Interpreting Excel Output

Problem: Machine A vs B.

Solution:

- (a) **Hypothesis Check:** $H_1 : \mu_A \neq \mu_B$ is a **Two-tail** test. Use "P(T<=t) two-tail".
- (b) **Decision:** $P_{two-tail} = 0.044$. Since $0.044 < 0.05$, we **Reject** H_0 . There is a significant difference.
- (c) **Critical Value:** $|t_{stat}| = 2.129$. $t_{crit} = 2.074$. Since $2.129 > 2.074$, the stat is in the rejection region. Confirmed.
- (d) **Pooled Variance Calculation:** $s_A^2 = 16.5$, $s_B^2 = 18.2$. $n = 12$.

$$\begin{aligned} S_p^2 &= \frac{11(16.5) + 11(18.2)}{22} \\ &= \frac{16.5 + 18.2}{2} = 17.35 \end{aligned}$$

Matches Excel exactly.

(b) $P = 0.044$. Significant.

11.14 Python: Sample Size (Power Analysis)

Problem: $d = 0.5$, Power = 0.8.

Solution:

- (a) **Output:** Approx 63.something \rightarrow 64. Must round UP because sample size must be integer, and 63 would yield power slightly < 0.8 .
- (b) **Graph Check (n=20):** Looking at the plot for $n = 20$, Power is around 0.3 – 0.4. This is **unacceptable**. You are more likely to miss the effect than find it.
- (c) **Smaller Effect ($d = 0.2$):** Sample size would **increase drastically**. Detecting a smaller signal amidst the same noise requires much more data to be certain.

11.15 Python: A/B Testing

Problem: Proportions Z-Test.

Solution:

- (a) **P-value:** From previous manual calc, approx 0.0136.
- (b) **Interpretation:** If both pages were equal, there is only a 1.36% chance we'd see B beating A by this much (20 conversions) just by luck.
- (c) **Two-sided:** P-value would **double** to approx 0.027. This is because a two-sided test checks for difference in *either* direction ($A \neq B$), splitting α into two tails.