Two-Sample Inference (Lectures 8-10)

BIOSTAT 201A Fall 2025

Discussion 5 – October 31, 2025

Cindy J. Pang

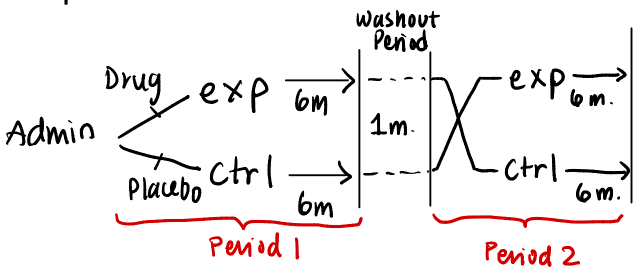
Outline

1. Motivation for Two Sample Inference Problems
2. Two Sample Inference for Paired Data (Lecture 8)
3. Two Sample Inference for Independent Samples (Lecture 9 & 10)

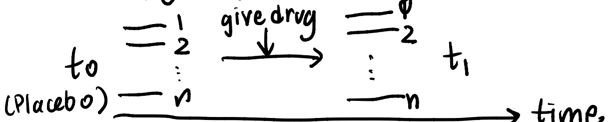
Two Sample Problems

- Goal of Two Sample Problems:
  - (1) Compare 2 Samples
  - (2) Identify Important Covariates (i.e. Males vs. Females, Young vs. Old, etc)
- Types
  - 1. Stratified Randomization (Independent Samples)
  - 2. Matched-Pairing (Dependent Samples)
    - Unpaired is more likely to reject but why should we still use a paired test?

      if there is a reason to do do matched pairs; we SHOULD do it, bc it helps w/ controlling for covariates (Goal #2)
    - Situations where we do matched-pairs in clinical research:
      - 1. Cross-Over Design



      - 2. Pre-Post Design



Two-Sample Problems	Dependent Groups (Paired Data)
Hypotheses	Let $\Delta = \mu_1 - \mu_2$ $H_0: \Delta = 0$ vs. $H_1: \Delta \neq 0$
Data	$\begin{array}{c\|cc\|l} i & x_{1i} & x_{2i} & d_i \\ \hline 1 & x_{11} & x_{21} & d_1 = x_{11} - x_{21} \\ 2 & x_{12} & x_{22} & d_2 = x_{12} - x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ n & x_{1n} & x_{2n} & d_n = x_{1n} - x_{2n} \end{array}$ where $d_i \sim N(\Delta, \sigma_d^2)$
Assumptions	The population of differences is normally distributed with mean Δ and variance σ_d^2
Test Statistic	$t = \dfrac{\bar{d}}{S_d/\sqrt{n}}$
Decision Rule	$\|t\| > t_{df=n-1,\, 1-\alpha/2} \Rightarrow$ Reject H_0 $\|t\| < t_{df=n-1,\, 1-\alpha/2} \Rightarrow$ Fail to reject H_0
Confidence Interval	$\bar{d} \pm t_{n-1,\, 1-\alpha/2} \dfrac{S_d}{\sqrt{n}}$ where $SE(d) = \dfrac{S_d}{\sqrt{n}}$

	Raw scale urinary protein (g/24 hr)		Diff
Patient	Before	After	
1	25.6	10.1	15.5
2	17.0	5.7	11.3
3	16.0	5.6	⋮
4	10.4	3.4	
5	8.2	6.5	
6	7.9	0.7	
7	5.8	6.1	
8	5.4	4.7	
9	5.1	2.0	
10	4.7	2.9	1.8

$\bar{d} = 5.84$

$s_d = 5.29$

Renal Disease. We are interested to see if the raw scale of urinary protein changes after 8 weeks.

(a) Identify the appropriate statistical procedure to do this. Explain.

Paired T-test since data are paired; dependence

(b) Did urinary protein change after 8 weeks? Use a hypothesis test to determine this.

$H_0: \Delta = 0$ vs. $\Delta \neq 0$

"no change" "changed"

Assuming the population of differences is normally distributed with mean Δ and variance σ_d^2,

$$TS = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{5.84}{5.29/\sqrt{10}} = 3.49, \quad t_{df=9,0.975} = 2.26$$

Since $TS = 3.49 > 2.26$, we $\underline{\text{reject}}$ the null hypothesis and conclude that there is a change in urinary protein after 8wks.

(c) Compute a 95% confidence interval for the differences in urinary protein change. How does your created confidence interval support your conclusions in part (b)?

$$\bar{d} \pm t_{9,0.975}\frac{s_d}{\sqrt{n}} = 5.84 \pm 2.62\left(\frac{5.29}{\sqrt{10}}\right) = (2.06, 9.62)$$

Since our CI does not include 0, we conclude that there is a significant difference in urinary protein change like we did in part (b)

Two-Sample Problems	Independent Groups		
	Equal Variances	Unequal Variances	Test for Equality of Variances
Hypotheses	$H_0: \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0$ $H_1: \mu_1 \neq \mu_2 \Leftrightarrow \mu_1 - \mu_2 \neq 0$		$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$
Data	Sample 1: $\bar{X}_1, S_1^2, n_1$ Sample 2: $\bar{X}_2, S_2^2, n_2$		
Assumptions	$X_{1i} \sim N(\mu_1, \sigma^2), X_{2i} \sim N(\mu_2, \sigma^2)$ where $\sigma^2 = \sigma_1^2 = \sigma_2^2$	$X_{1i} \sim N(\mu_1, \sigma_1^2)$ $X_{2i} \sim N(\mu_2, \sigma_2^2)$	
Test Statistic	$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$	$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	$F = \frac{S_2^2}{S_1^2}$ where $S_2 \geq S_1$
Decision Rule	$\lvert t\rvert \geq t_{df=n_1+n_2-2, 1-\alpha/2} \Rightarrow$ Reject H_0 $\lvert t\rvert < t_{df=n_1+n_2-2, 1-\alpha/2} \Rightarrow$ Fail to Reject H_0	Use Satterwaithe's Approx. for df s.t. $d' = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}$ and round down to nearest integer, d'' Then $\lvert t\rvert \geq t_{d'', 1-\alpha/2} \Rightarrow$ Reject H_0 $\lvert t\rvert < t_{d'', 1-\alpha/2} \Rightarrow$ Fail to Reject H_0	$F > F_{n_2-1, n_1-1, 1-\alpha/2}$ or $F < F_{n_2-1, n_1-1, \alpha/2}$ $\Rightarrow$ Reject H_0 o.w. $\Rightarrow$ Fail to Reject H_0
Confidence Interval	$(\bar{X}_1 - \bar{X}_2) \pm t_{n_1+n_2-2, 1-\alpha/2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$(\bar{X}_1 - \bar{X}_2) \pm t_{d'', 1-\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$	

Nutrition. The mean ±1 sd of ln [calcium intake (mg)] among 25 females, 12 to 14 years of age, below the poverty level is 6.56 ± 0.64. Similarly, the mean ± 1 sd of ln [calcium intake (mg)] among 40 females, 12 to 14 years of age, above the poverty level is 6.80 ± 0.76. Let X_{1i} = ln of females below poverty, X_{2i} = ln of females above poverty.

where $\bar{X}_1 = 6.56$, $S_1 = 0.64$, $n_1 = 25$

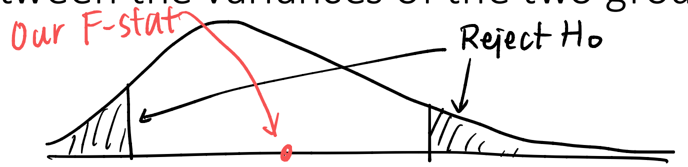
$\bar{X}_2 = 6.80$, $S_2 = 0.76$, $n_2 = 40$

$\alpha = 0.05$

(a) Test for a significant difference between the variances of the two groups.

$H_0: \sigma_1^2 = \sigma_2^2$ vs. $H_1: \sigma_1^2 \neq \sigma_2^2$

$F = S_2^2 / S_1^2 = 1.41$



$F_{39,24,0.025} = 0.49$ $\quad F_{39,24,0.975} = 2.15$

Since our test statistic is between $F_{39,24,0.025}$ and $F_{39,24,0.975}$ we fail to reject Ho and conclude that the variances of the two groups are not significantly different.

(b) Test for whether there is a significant difference in means between the two groups.

$H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$ w/ pooled variance (equality of variances s.t. $\sigma^2 = \sigma_1^2 = \sigma_2^2$)

$$TS = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ where } S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{(25-1)(0.64)^2 + (40-1)(0.76)^2}{25+40-2}} = 0.717, \; t_{crit} = t_{63,0.975} = 1.99$$

$$= \frac{6.56 - 6.80}{0.717\sqrt{\frac{1}{25} + \frac{1}{40}}}$$

$$= -1.314$$

Since $|TS| < 1.99$, we fail to reject Ho and conclude that there is no significant difference in mean calcium intake between females age 12-14 above and below the poverty line.

(c) Create a 95% Confidence Interval for the difference in means between the two groups.

$(\bar{X}_1 - \bar{X}_2) \pm t_{crit} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (-0.605, 0.125)$ → includes 0 → no significant diff b/t groups.