STAT 500

Two-Sample Inference: Interval Estimation

Scenario

- Randomized Experiment
  - Two treatments
  - Is there a difference in the mean value of the response variable between the two treatments?

- Observational Study
  - Two populations
  - One sample from each population
  - Is there a difference in the mean value of the variable between the two populations?

Interval estimation: confidence interval for $\mu_1 - \mu_2$

- Researchers may be interested in estimating the mean difference between the two treatments/populations under study.

- Assumptions
  - $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ are i.i.d. $N(\mu_1, \sigma^2)$
  - $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ are i.i.d. $N(\mu_2, \sigma^2)$
  - Y_{1j} and $Y_{2j'}$ are independent for all j and j'

Interval Estimation

- Based on the assumptions, we have the key result

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

- We can construct a confidence interval using this result.

Interval Estimation

- The idea is that, a $100(1-\alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$ contains the values of $\mu_1 - \mu_2$ for which

$$\left| \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \le t_{(n_1+n_2-2),1-\alpha/2}$$

- Solving for $\mu_1 - \mu_2$ gives

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{n_1+n_2-2,1-\alpha/2}\ S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Interval Estimation

$$
\begin{aligned}
1-\alpha &= Pr\left(t_{n_1+n_2-2,\alpha/2} \leq \frac{(\bar{Y}_1-\bar{Y}_2)-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq t_{n_1+n_2-2,1-\alpha/2}\right) \\
&= Pr\left(t_{n_1+n_2-2,\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq (\bar{Y}_1-\bar{Y}_2)-(\mu_1-\mu_2) \leq\right. \\
&\qquad\qquad \left. t_{n_1+n_2-2,1-\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\ \right) \\
&= Pr\left(-(\bar{Y}_1-\bar{Y}_2)+t_{n_1+n_2-2,\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq -(\mu_1-\mu_2) \leq\right. \\
&\qquad\qquad \left. -(\bar{Y}_1-\bar{Y}_2)+t_{n_1+n_2-2,1-\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\ \right)
\end{aligned}
$$

Interval Estimation

$$
\begin{aligned}
1-\alpha &= Pr\left((\bar{Y}_1-\bar{Y}_2)-t_{n_1+n_2-2,\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \geq (\mu_1-\mu_2) \geq \right. \\
&\qquad \left. (\bar{Y}_1-\bar{Y}_2)-t_{n_1+n_2-2,1-\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right) \\
&= Pr\left((\bar{Y}_1-\bar{Y}_2)-t_{n_1+n_2-2,1-\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq (\mu_1-\mu_2) \leq \right. \\
&\qquad \left. (\bar{Y}_1-\bar{Y}_2)+t_{n_1+n_2-2,1-\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)
\end{aligned}
$$

Relationship Between Tests and Confidence Intervals

- Let $\delta = \mu_1 - \mu_2$

- A two-sided t-test will reject $H_o : \mu_1 - \mu_2 = 0$ at the α level if and only if 0 is not in $100(1-\alpha)\%$ confidence interval for $\delta = \mu_1 - \mu_2$.

- A $100(1-\alpha)\%$ confidence interval can be constructed by including all values of δ such that data does not provide sufficient evidence to reject the null hypothesis $H_o : \mu_1 - \mu_2 = \delta$ relative to the two-sided alternative $H_a : \mu_1 - \mu_2 \neq \delta$ at the α significance level.

Frequentist Interpretation

- μ_1, μ_2, σ are fixed unknowns
- $\bar{Y}_1, \bar{Y}_2, S_p$ are random variables
- A confidence interval is a *random* interval
- Across a large number of repeated samples, $100(1-\alpha)\%$ of such intervals will contain the true value of $\mu_1 - \mu_2$ (a long-term frequency property).
- Any particular interval either contains the true value of $\mu_1 - \mu_2$ or not.
- The $100(1-\alpha)\%$ probability describes the process of constructing the intervals.

Interval Estimation

Classical frequentist interpretation

- Confidence interval widths depend on
  - the confidence level
  - the value of $\boldsymbol{\sigma}$
  - sample sizes $\boldsymbol{n_1}$ and $\boldsymbol{n_2}$

For the lizard study, a $\mathbf{95\%}$ confidence interval for $\boldsymbol{\mu_1 - \mu_2}$ is

$$\mathbf{-5.3733 \pm (2.048)(7.4649)(\sqrt{\tfrac{1}{15} + \tfrac{1}{15}} \quad \Rightarrow (-10.96, 0.21)}$$

T-test for Mean Distance for Two Minute Runs Sceloporis Occidentalis Lizards

The TTEST Procedure

Variable: distance

infection	N	Mean	Std Dev	Std Err	Minimum	Maximum
yes	15	26.8600	6.8096	1.7582	16.4000	37.1000
no	15	32.2333	8.0672	2.0829	18.4000	45.5000
Diff (1-2)		-5.3733	7.4649	2.7258		

infection	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
yes		26.8600	23.0889	30.6311	6.8096	4.9855	10.7395
no		32.2333	27.7659	36.7008	8.0672	5.9062	12.7228
Diff (1-2)	**Pooled**	-5.3733	-10.9569	0.2102	7.4649	5.9240	10.0960
Diff (1-2)	**Satterthwaite**	-5.3733	-10.9640	0.2173			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	-1.97	0.0586
Satterthwaite	Unequal	27.233	-1.97	0.0589

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	14	14	1.40	0.5343

Comparison of Randomization and Model Based Approaches

- Randomization tests
  - Require no model for population distributions
  - Are appropriate for randomized experiments
  - Require more computation
- Randomization CI's (based on test/interval relationship) require even more computation
- Model-based tests and CIs provide easily evaluated approximations to randomization inference
- Model-based approach can be useful for study design

Creative Writing Study
t-values for 10000 Permutations