STAT 5000

Statistical Methods I

Week 3

Fall 2024

Dr. Danica Ommen

Unit 1

Model-Based Inference:

Hypothesis Testing

Scenario:

- Randomized Experiment
  - Recruit sample from a single population
  - Two treatments
  - Is there a difference in the mean value of the response variable between the two treatments?
- Observational Study
  - Two populations
  - One random sample from each population
  - Is there a difference in the mean value of the variable between the two populations?

Statistical Assumptions:

- $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ are i.i.d. $N(\mu_1, \sigma_1^2)$
- $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ are i.i.d. $N(\mu_2, \sigma_2^2)$
- Y_{1i} and Y_{2j} are independent for all i and j
- $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Key Components:

- Research Question & Hypotheses
- Test Statistic
- Sampling Distribution
- *p*-value
- Interpretation
- Post-hoc Assessment

Research Question & Hypotheses:

Research Question: should be clearly formulated
i.e., Is this person guilty of committing a crime?

Null Hypothesis: denoted H_0
this is the *status quo*
i.e., innocent

Alternative Hypothesis: denoted H_a
what you're trying to *prove*
should be formulated directly from the question
i.e., guilty

Hypotheses

- $H_0 : \mu_1 = \mu_2 \quad (\mu_1 - \mu_2 = 0)$
- H_a :
  - Left-tailed: $\mu_1 < \mu_2 \quad (\mu_1 - \mu_2 < 0)$
  - Right-tailed: $\mu_1 > \mu_2 \quad (\mu_1 - \mu_2 > 0)$
  - Two-tailed: $\mu_1 \neq \mu_2 \quad (\mu_1 - \mu_2 \neq 0)$
  - Should pick one based on the research question

Test Statistic

- This is like the *evidence* in the criminal trial example ...
- Compute the *observed value* from the study data

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Is it typical to see this value when H_0 is true?
- Or is this value really unlikely when H_0 is true?
- Compare this to a distribution of randomized values

Sampling Distribution

- Create the distribution of randomized values (assuming H_0 true)

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

  - If H_0 is true, $\mu_1 - \mu_2 = 0$
  - If H_0 is true, expect to observe T close to zero.
  - Unlikely to observe large deviations from zero if H_0 true.

p-value

Definition: The probability of getting a new test statistic that is at least as extreme as the one observed assuming H_0 is true

- The p-value is NOT the probability that H_0 is true
- Use the sampling distribution to compute
- Choose the "extreme" (or tail) based on H_a
- Compute area in tail where t is the cutoff

Computing p-values

$H_a : \mu_1 \neq \mu_2$

$$p\text{-value} = 2 * P(T_{n_1+n_2-2} > |t|)$$

$H_a : \mu_1 < \mu_2$

$$p\text{-value} = P(T_{n_1+n_2-2} < t)$$

$H_a : \mu_1 > \mu_2$

$$p\text{-value} = P(T_{n_1+n_2-2} > t)$$

Interpretation: **Scale-of-Evidence Framework**

p-value range	scale of evidence statement
$p > 0.1$	little to no evidence for H_a
$0.05 < p < 0.1$	borderline/weak evidence for H_a
$0.025 < p < 0.05$	moderate evidence for H_a
$0.001 < p < 0.025$	substantial/strong evidence for H_a
$p < 0.001$	overwhelming evidence for H_a

Interpretation: Statistical-Significance Framework

- Choose your significance level, α
- Reject H_0 if p-value $< \alpha$
  There is statistically significant evidence for H_a
- Fail to reject H_0 if if p-value $\geq \alpha$
  There is no statistically significant evidence for H_a
- Should always state your interpretation in the context of your study

Post-hoc Assessment: Errors

- If the p-value was small:
  - H_0 is true and we unluckily/randomly made an error
  - Type 1 error probability: $P(\text{reject } H_0 | H_0 \text{ true}) \leq \alpha$
  - H_0 is false (no error committed)
- If the p-value was large:
  - H_a is true and we unluckily/randomly made an error
  - Type 2 error probability: $P(\text{fail to reject } H_0 | H_0 \text{ false}) = \beta$
  - The power of a test is $1 - \beta$
  - H_0 is true (no error committed)

Example: Lizard Infection Study

- Independent random samples of 15 lizards infected with a disease and 15 non-infected lizards
- Test null hypothesis that mean distance traveled in two minutes is the same for both populations
- t-test of $H_0 : \mu_1 = \mu_2$ versus $H_0 : \mu_1 \neq \mu_2$
  $t = \frac{-5.3733}{(7.4649)(\sqrt{\frac{1}{15}+\frac{1}{15}})} = -1.97$ on 28 d.f. $\Rightarrow$ p-value $= 0.0586$
- t-test of $H_0 : \mu_1 = \mu_2$ versus $H_0 : \mu_1 < \mu_2$
  $t = \frac{-5.3733}{(7.4649)(\sqrt{\frac{1}{15}+\frac{1}{15}})} = -1.97$ on 28 d.f. $\Rightarrow$ p-value $= 0.0293$

Example: Lizard Infection Study

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

The TTEST Procedure

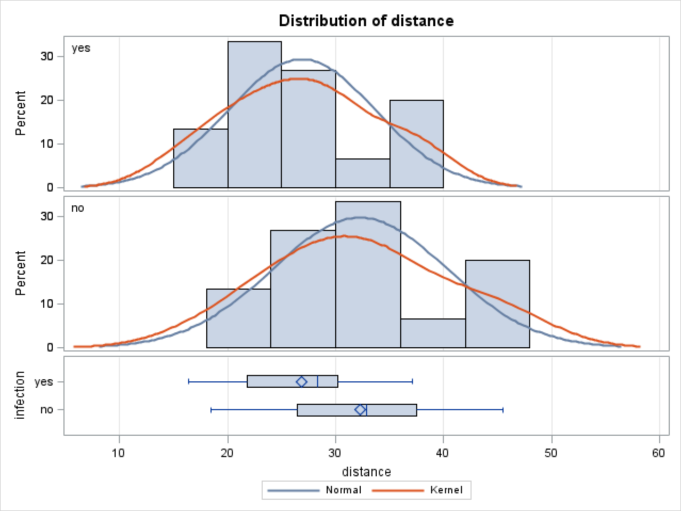
Variable: distance

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		

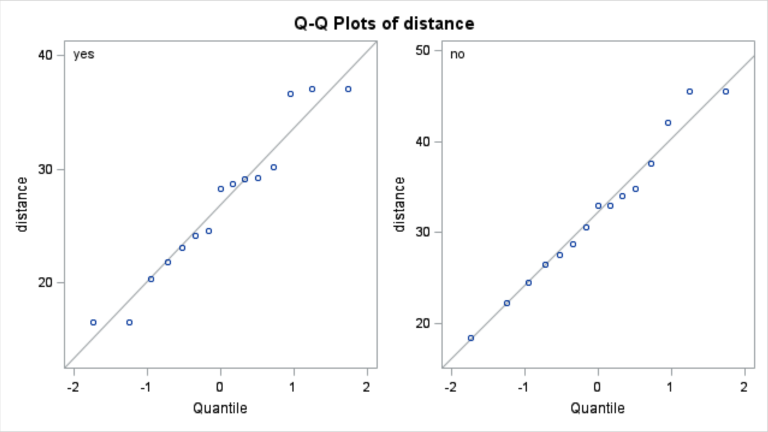
infection	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
yes		26.8600	23.0889	30.6311	6.8096	4.9855	10.7395
no		32.2333	27.7659	36.7008	8.0672	5.9062	12.7228
Diff (1-2)	**Pooled**	-5.3733	-10.9569	0.2102	7.4649	5.9240	10.0960
Diff (1-2)	**Satterthwaite**	-5.3733	-10.9640	0.2173			

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

Example: Lizard Infection Study



Example: Lizard Infection Study



Unit 1

Model-Based Inference:

Confidence Intervals

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?

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. $N(\mu_2, \sigma^2)$
  - So, population variances are equal
  - Y_{1i} and Y_{2j} are independent for all i and j

Confidence interval for $\mu_1 - \mu_2$

- 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 ...

Confidence interval for $\mu_1 - \mu_2$

- 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

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

- Solving for $\mu_1 - \mu_2$ gives

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

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

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

Difference in Means

- 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. $N(\mu_2, \sigma^2)$
  - So, population variances are equal
  - Y_{1i} and Y_{2j} are independent for all i and j
- Then, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

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

where $S_p = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$

Relationship Between Tests and Confidence Intervals

- Let $\delta = \mu_1 - \mu_2$ (A.K.A. the true effect size)
- A two-sided t-test will reject $H_0 : \delta = \mu_1 - \mu_2 = 0$ 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 $H_0 : \mu_1 - \mu_2 = \delta$ 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* quantity
- 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 <u>process</u> of constructing the intervals.

More Details ...

- Confidence interval widths depend on
  - the confidence level (which is related to significance α)
  - the value of σ
  - sample sizes n_1 and n_2
- For the lizard study, a 95% confidence interval for $\mu_1 - \mu_2$ is $-5.3733 \pm (2.048)(7.4649)\sqrt{\frac{1}{15} + \frac{1}{15}} \Rightarrow (-10.96, 0.21)$

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		

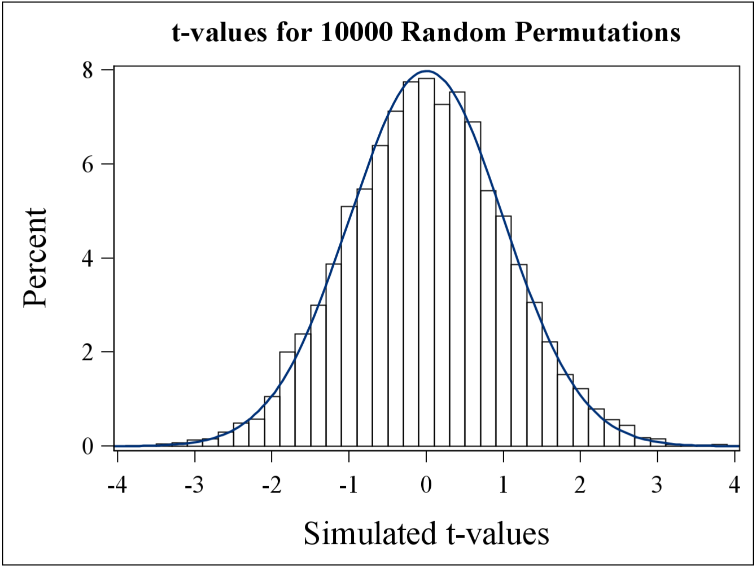
infection	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
yes		26.8600	23.0889	30.6311	6.8096	4.9855	10.7395
no		32.2333	27.7659	36.7008	8.0672	5.9062	12.7228
Diff (1-2)	**Pooled**	-5.3733	-10.9569	0.2102	7.4649	5.9240	10.0960
Diff (1-2)	**Satterthwaite**	-5.3733	-10.9640	0.2173			

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


t-values for 10000 Random Permutations
Percent
0
2
4
6
8
-4
-3
-2
-1
0
1
2
3
4
Simulated t-values

Unit 1

Model-Based Inference:

Sample Size Considerations

Sample Size Determination

- Frequently asked design questions:
  How many experimental units should I use?
- Four possible considerations:
  1. Include as many as you can afford or find
  2. Desired precision (standard error) of some estimator
  3. Width of a confidence interval
  4. Power of a hypothesis test

Based on Standard Error

Difference in Means

- Difference in population means ($\mu_1 - \mu_2$):
  $s.e.(\bar{Y}_1 - \bar{Y}_2) = S_p\sqrt{1/n_1 + 1/n_2}$
- Assuming $n_1 = n_2 = n$, we have:
  $s.e.(\bar{Y}_1 - \bar{Y}_2) = S_p\sqrt{2/n}$
- Specify an acceptable value for the s.e. and solve for n
  $$s.e. = \frac{\sqrt{2}S_p}{\sqrt{n}} \quad \Rightarrow \quad n = \frac{2S_p^2}{(s.e.)^2}$$
- Requires a value for S_p from:
  - a previous study
  - a pilot study
  - a guess

Based on Confidence Interval

Difference in Means

- Width of CI (assuming $n_1 = n_2 = n$) is
  $w = 2 \times t_{2(n-1),1-\alpha/2} S_p \sqrt{2/n}$
- Find n to achieve specified width
  $$n = 8\left(\frac{t_{2(n-1),1-\alpha/2} S_p}{w}\right)^2$$
- One difficulty is that n enters twice (sample size and df for t)
  - Compute initial value using $z_{1-\alpha/2}$ in place of the t-value
    $$n_0 = 8\left(\frac{z_{1-\alpha/2} S_p}{w}\right)^2$$
  - Then improve using $n = 8\left(\frac{t_{2(n_0-1),1-\alpha/2} S_p}{w}\right)^2$

Based on Hypothesis Test

Recall:

Four Possible Outcomes for Hypothesis Test

Decision	H_0 is true	H_0 is false
Reject H_0	**Type I Error**	Good Decision
Fail to reject H_0	Good Decision	**Type II Error**

Recall: Hypothesis Testing Errors

- Type I error $\Rightarrow$ reject H_0 when it is true
  - Probability of Type I error: α = P(reject H_0 | H_0 is true)
  - Threshold for p-value, significance level
  - Value of α is typically set to be small, such as 0.05.
- Type II error $\Rightarrow$ fail to reject H_0 when it is false
  - Probability of Type II error: β = P(fail to reject H_0 | H_0 is false)

Recall: Power

- Power of a statistical test = $1 - \beta$
- Function of a particular alternative to the null hypothesis: Power = $1 - \beta$= P(reject $H_0 | H_a$ true)
- For fixed α, power is determined by
  - How much the alternative deviates from the null hypothesis (true effect size, e.g., $\delta = \mu_1 - \mu_2$)
  - population variance (σ^2)
  - sample sizes (n_1, n_2)

We can use power to determine the sample size because, for any given test, type I error rate α, power $1-\beta$, standard deviation σ or S, effect size $\delta = \mu_1 - \mu_2$, and sample size n are all related. Specifying four enables us to calculate the fifth.

Based on Hypothesis Test

Difference in Means

For a t-test of $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$

- Equal sample sizes $n_1 = n_2 = n$,
- Type I error rate $= \alpha$,
- Power =1 $- \beta$ for detecting $\delta = \mu_1 - \mu_2$,
- Pooled estimate of the population variance S_p^2

the required sample size for each group is

$$n = \frac{\left(t_{2(n-1),1-\alpha/2} + t_{2(n-1),1-\beta}\right)^2 \left(2S_p^2\right)}{\delta^2}$$

Based on Hypothesis Test

Difference in Means

- As before, n enters twice. Use the same two-step approach. First compute $n_0 = \dfrac{(z_{1-\alpha/2} + z_{1-\beta})^2\,(2S_p^2)}{\delta^2}$
- Then update $n = \dfrac{(t_{2(n_0-1),1-\alpha/2} + t_{2(n_0-1),1-\beta})^2\,(2S_p^2)}{\delta^2}$
- Common to use power values of 80%, 90% or 95%. Just as arbitrary as using α=5%.
- Can adapt to one-sided alternative by replacing $\alpha/2$ with α in the previous formulas

Derivation:

- Based on the model assumptions, we have the key result

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

- While controlling the type I error rate at α, we reject H_0 if

$$\left|\frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}\right| > t_{(n_1+n_2-2),1-\alpha/2}$$

Derivation:

$$
\begin{aligned}
1-\beta &= P(\text{reject } H_0 : \mu_1-\mu_2=0 \,|\, \mu_1-\mu_2=\delta) \\
&= P\left(\left|\frac{(\bar{Y}_1-\bar{Y}_2)-0}{\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}}\right| > t_{(n_1+n_2-2),1-\alpha/2} \,\middle|\, \mu_1-\mu_2=\delta\right) \\
&= P\left(|(\bar{Y}_1-\bar{Y}_2)-0| > t_{(n_1+n_2-2),1-\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \,\middle|\, \mu_1-\mu_2=\delta\right) \\
&= P\left((\bar{Y}_1-\bar{Y}_2) > t_{(n_1+n_2-2),1-\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \,\middle|\, \mu_1-\mu_2=\delta\right) \\
&\quad + P\left((\bar{Y}_1-\bar{Y}_2) < -t_{(n_1+n_2-2),1-\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \,\middle|\, \mu_1-\mu_2=\delta\right)
\end{aligned}
$$

Derivation:

$$
\begin{aligned}
1-\beta = \\
P&\left(\frac{(\bar{Y}_1-\bar{Y}_2)-\delta}{\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}} > \frac{t_{(n_1+n_2-2),1-\alpha/2}\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}-\delta}{\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}} \,\middle|\, \mu_1-\mu_2=\delta\right) \\
+P&\left(\frac{(\bar{Y}_1-\bar{Y}_2)-\delta}{\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}} < \frac{-t_{(n_1+n_2-2),1-\alpha/2}\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}-\delta}{\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}} \,\middle|\, \mu_1-\mu_2=\delta\right)
\end{aligned}
$$

This implies that

$$
-t_{(n_1+n_2-2),1-\beta} \approx \frac{t_{(n_1+n_2-2),1-\alpha/2}\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}-\delta}{\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}}
$$

Derivation:

- With $n_1 = n$ and $n_2 = Cn$ and $\delta > 0$ we have

$$-t_{((1+C)n-2),1-\beta} \approx \frac{t_{((1+C)n-2),1-\alpha/2}\sqrt{S_p^2(\frac{1}{n}(1+\frac{1}{C}))} - \delta}{\sqrt{S_p^2(\frac{1}{n}(1+\frac{1}{C}))}}$$

- Solve for n:

$$n = \frac{(t_{((1+C)n-2),1-\alpha/2} + t_{((1+C)n-2),1-\beta})^2 S_p^2(1+\frac{1}{C})}{\delta^2}$$

- Setting $n_1 = n_2 = n$, we get

$$n = \frac{(t_{2(n-1),1-\alpha/2} + t_{2(n-1),1-\beta})^2 2S_p^2}{\delta^2}$$

Example: Bone Density Study

- Client's request:

  I want to do a randomized experiment to compare two treatments for preventing bone loss in elderly women. I want to use a 0.05-level t-test that has at least an 80% chance of detecting of a difference of 4 units between mean loss in bone density for the two treatments. How many subjects do I need?

- After some discussion the client explains that she would like to use treatment groups of the same size

Example: Bone Density Study

- Further discussion about previous studies of bone density loss in elderly women suggests that the measured responses should be approximately normally distributed for each treatment $\Rightarrow$ a t-test could be used
- The null hypothesis is $H_0 : \mu_1 = \mu_2$
- The alternative is two-sided and the deviation of interest is $\delta = |\mu_1 - \mu_2| = 4$
- Type I error level is $\alpha = 0.05$
- Power = 1-β = 0.80, and β=0.20
- Review of previous studies suggests $S_p^2 \approx 25$

Example: Bone Density Study

- Compute initial sample size value

$$
\begin{aligned}
n_0 &= \frac{(z_{0.975} + z_{0.80})^2\,(2S_p^2)}{\delta^2} \\
&= \frac{(1.96 + 0.841)^2\,(2 \times 25)}{(4)^2} \\
&= 24.52 \\
&\Rightarrow 25
\end{aligned}
$$

Example: Bone Density Study

- Then compute

$$
\begin{aligned}
n &= \frac{\left(t_{48,0.975} + t_{48,0.80}\right)^2 \left(2S_p^2\right)}{\delta^2} \\
&= \frac{(2.01 + 0.849)^2\,(2 \times 25)}{(4)^2} \\
&= 25.54 \\
&\Rightarrow 26
\end{aligned}
$$

- Use 26 subjects in each treatment group

Example: Bone Density Study

> *Instead of performing a test, suppose the client in the previous example wanted to construct a 95% confidence interval for the difference in the mean bone density loss for the two treatments of width 4 units*

Goal: $(\bar{Y}_1 - \bar{Y}_2) \pm 2$

- Width of CI (assuming $n_1 = n_2 = n$) is
  $w = 2 \times t_{2(n-1),1-\alpha/2} S_p \sqrt{2/n}$
- Find n to achieve specified width
  $n = 8 \left(\frac{t_{2(n-1),1-\alpha/2} S_p}{w} \right)^2$

Example: Bone Density Study

- $\alpha = 0.05$ and $S_p^2 \approx 25$
- First compute

$$n_0 = 8\left(\frac{z_{0.975} \times S_p}{w}\right)^2 = 8\left(\frac{1.96 \times 5}{4}\right)^2 = 48.02 \quad \Rightarrow \quad 49$$

- Then improve using

$$n = 8\left(\frac{t_{96,0.975} \times S_p}{w}\right)^2 = 8\left(\frac{1.99 \times 5}{4}\right)^2 = 49.50 \quad \Rightarrow \quad 50$$

- Use 50 subjects in each group

Example: Bone Density Study

> *Suppose the client simply wanted to use enough subjects to make the standard error for the estimated difference of the means no larger than 1.0.*

Goal: $s.e.(\bar{Y}_1 - \bar{Y}_2) = S_p\sqrt{\frac{2}{n}} = 1$

- $S_p^2 \approx 25$
- Compute $n = \dfrac{2 \times S_p^2}{(1)^2} = \dfrac{2(25)}{1} = 50$
- Use 50 subjects in each group

Designing Comparative Studies

- Consider unequal group sizes in experiments when
  - One treatment is much more expensive than another
  - One treatment is limited or not readily available
  - Variation in responses differs between treatments
- Unequal sample sizes in observational studies are more common
- Previously only considered $n_1 = n_2 = n$, why?
  - Could consider $n_1 = 2n_2$, same concept, modified formula
  - For fixed value of $n_1 + n_2$, $n_1 = n_2$ gives estimator of difference in means with smallest s.e. (assuming homogeneous variances)

Designing Comparative Studies

- Distribution of the t-statistic is not distorted by unequal variances when $n_1 = n_2$
- Size for one group may be limited by availability or higher cost, but
  $n_1 = 10,\ n_2 = 40$ is better than $n_1 = 10,\ n_2 = 10$
  $n_1 = 10,\ n_2 = 40$ is not as good as $n_1 = 25,\ n_2 = 25$

Questions?

Contact me:

Email: dmommen@iastate.edu
Visit Student Office Hours: Thur @ 10-11 am