

7. Comparing Two Groups

Goal: Use CI and/or significance test to compare means (quantitative variable) or compare proportions (categorical variable)

	Group 1	Group 2	Estimate
Population mean	μ_1	μ_2	$\bar{y}_2 - \bar{y}_1$
Population proportion	π_1	π_2	$\hat{\pi}_2 - \hat{\pi}_1$

We conduct inference about the difference between the means or difference between the proportions (order irrelevant).

Does cell phone use while driving impair reaction times?

- Article in *Psych. Science* (2001, p. 462) describes experiment that randomly assigned 64 Univ. of Utah students to cell phone group or control group (32 each). Driving simulating machine flashed *red* or *green* at irregular periods. Instructions: Press brake pedal as soon as possible when detect red light.

See <http://www.psych.utah.edu/AppliedCognitionLab/>

- Cell phone group: Carried out conversation about a political issue with someone in separate room.
- Control group: Listened to radio broadcast

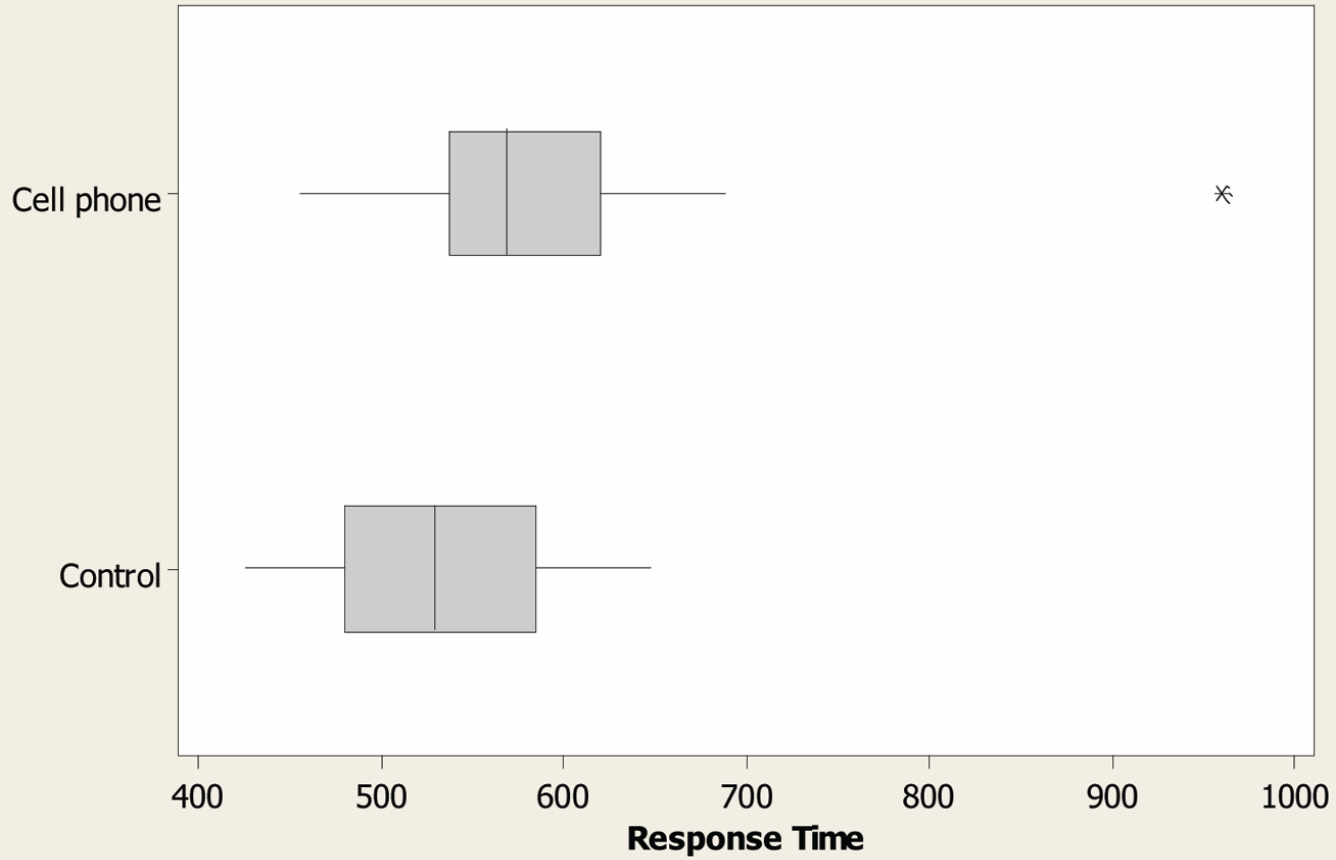
Outcome measure: mean response time for a subject over a large number of trials

- Purpose of study: Analyze whether (conceptual) population mean response time differs significantly for the two groups, and if so, by how much.

- Data

Cell-phone group: $\bar{y}_1 = 585.2$ milliseconds, $s_1 = 89.6$

Control group: $\bar{y}_2 = 533.7$, $s_2 = 65.3$.



Types of variables and samples

- The outcome variable on which comparisons are made is the **response variable**.
- The variable that defines the groups to be compared is the **explanatory variable**.

Example: *Reaction time* is response variable

Experimental group is explanatory variable

(categorical var. with categories cell-phone, control)

Or, could express experimental group as “cell-phone use” with categories (yes, no)

- Different methods apply for **dependent samples** (natural matching between each subject in one sample and a subject in other sample, such as in “longitudinal studies,” which observe subjects repeatedly over time) and **independent samples** (different samples, no matching, as in “cross-sectional studies”).

Example: We later consider a separate experiment in which the *same subjects* formed the control group at one time and the cell-phone group at another time.

Example: Anorexia study, studying weight change for 3 groups (behavioral therapy, family therapy, control)

What would be an example of

independent samples?

dependent samples?

se for difference between two estimates (independent samples)

- The sampling distribution of the difference between two estimates is *approximately normal* (large n_1 and n_2) and has estimated

$$se = \sqrt{(se_1)^2 + (se_2)^2}$$

Example: Data on “Response times” has

32 using cell phone with mean 585.2, $s = 89.6$

32 in control group with mean 533.7, $s = 65.3$

What is se for difference between means of

$$585.2 - 533.7 = 51.4?$$

$$se_1 = s_1 / \sqrt{n_1} = 89.6 / \sqrt{32} = 15.84$$

$$se_2 = s_2 / \sqrt{n_2} = 65.3 / \sqrt{32} = 11.54$$

$$se = \sqrt{(se_1)^2 + (se_2)^2} = \sqrt{(15.84)^2 + (11.54)^2} = 19.6$$

(Note larger than each separate se. *Why?*)

So, the estimated difference of 51.4 has a margin of error of *about* $2(19.6) = 39.2$

95% CI is about 51.4 ± 39.2 , or (12, 91).

Interpretation: We can be 95% confident that population mean for cell phone group is between 12 milliseconds higher and 91 milliseconds higher than population mean for the control group.

(In practice, good idea to re-do analysis without outlier, to check its influence. What do you think would happen?)

CI comparing two proportions

- Recall se for a sample proportion used in a CI is

$$se = \sqrt{\hat{\pi}(1 - \hat{\pi}) / n}$$

- So, the se for the difference between sample proportions for two independent samples is

$$se = \sqrt{(se_1)^2 + (se_2)^2} = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}$$

- A CI for the difference between population proportions is

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm z \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}$$

As usual, z depends on confidence level, 1.96 for 95% confidence

Example: College Alcohol Study conducted by
Harvard School of Public Health
(<http://www.hsph.harvard.edu/cas/>)

Trends over time in percentage of binge drinking
(consumption of 5 or more drinks in a row for men and 4 or
more for women, at least once in past two weeks)
or activities influenced by it?

“Have you engaged in unplanned sexual activities
because of drinking alcohol?”

1993: 19.2% yes of $n = 12,708$

2001: 21.3% yes of $n = 8783$

What is 95% CI for change saying “yes”?

- Estimated change in proportion saying “yes” is $0.213 - 0.192 = 0.021$.

$$se = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} = \sqrt{\frac{(.192)(.808)}{12,708} + \frac{(.213)(.787)}{8783}} = 0.0056$$

95% CI for change in population proportion is

$$0.021 \pm 1.96(0.0056) = 0.021 \pm 0.011, \text{ or roughly } (0.01, 0.03)$$

We can be 95% confident that the population proportion saying “yes” was between about 0.01 larger and 0.03 larger in 2001 than in 1993.

Comments about CIs for difference between two population proportions

- If 95% CI for $\pi_2 - \pi_1$ is (0.01, 0.03), then 95% CI for $\pi_1 - \pi_2$ is (-0.03, -0.01). It is arbitrary what we call Group 1 and Group 2 and what the order is for comparing the proportions.
- When 0 is not in the CI, we can conclude that one population proportion is higher than the other.
(e.g., if all positive values when take Group 2 – Group 1, then conclude population proportion higher for Group 2 than Group 1)

- When 0 is in the CI, it is plausible that the population proportions are identical.

Example: Suppose 95% CI for change in population proportion (2001 – 1993) is (-0.01, 0.03)

“95% confident that population proportion saying yes was between 0.01 *smaller* and 0.03 *larger* in 2001 than in 1993.”

- There is a significance test of $H_0: \pi_1 = \pi_2$ that the population proportions are identical (i.e., difference $\pi_1 - \pi_2 = 0$), using test statistic $z = (\text{difference between sample proportions})/se$

For unplanned sex in 1993 and 2001,

$$z = \text{diff.}/se = 0.021/0.0056 = 3.75$$

$$\text{Two-sided P-value} = 0.0002$$

This seems to be *statistical significance* without *practical significance!*

Details about test on pp. 189-190 of text; use se_0 which pools data to get better estimate under H_0

(We study this test as a special case of “chi-squared test” in next chapter, which deals with possibly many groups, many outcome categories)

- The theory behind the CI uses the fact that sample proportions (and their differences) have approximate normal sampling distributions for large n 's, by the Central Limit Theorem, assuming randomization)
- In practice, formula works ok if at least 10 outcomes of each type for each sample (Note: We don't use t dist. for inference about proportions; however, there are specialized small-sample methods, e.g., using binomial distribution)

Quantitative Responses: Comparing Means

- Parameter: $\mu_2 - \mu_1$
- Estimator: $\bar{y}_2 - \bar{y}_1$
- Estimated standard error: $se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
 - Sampling dist.: Approximately normal (large n 's, by CLT)
 - CI for independent random samples *from two normal population distributions* has form

$$(\bar{y}_2 - \bar{y}_1) \pm t(se), \text{ which is } (\bar{y}_2 - \bar{y}_1) \pm t \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Formula for df for t -score is complex (later). If both sample sizes are at least 30, can just use z-score

Example: GSS data on “number of close friends”

Use gender as the explanatory variable:

486 females with mean 8.3, $s = 15.6$

354 males with mean 8.9, $s = 15.5$

$$se_1 = s_1 / \sqrt{n_1} = 15.6 / \sqrt{486} = 0.708$$

$$se_2 = s_2 / \sqrt{n_2} = 15.5 / \sqrt{354} = 0.824$$

$$se = \sqrt{(se_1)^2 + (se_2)^2} = \sqrt{(0.708)^2 + (0.824)^2} = 1.09$$

Estimated difference of $8.9 - 8.3 = 0.6$ has a margin of error of $1.96(1.09) = 2.1$, and 95% CI is

0.6 ± 2.1 , or $(-1.5, 2.7)$.

- We can be 95% confident that the population mean number of close friends for males is between 1.5 less and 2.7 more than population mean number of close friends for females.
- Order is arbitrary. 95% CI comparing means for females – males is $(-2.7, 1.5)$
- When CI contains 0, it is plausible that the difference is 0 in the population (i.e., population means equal)
- Here, normal population assumption clearly violated. For large n 's, no problem because of CLT, and for small n 's the method is robust. (But, means may not be relevant for very highly skewed data.)
- Alternatively could do significance test to find strength of evidence about whether population means differ.

Significance Tests for $\mu_2 - \mu_1$

- Typically we wish to test whether the two population means differ
(null hypothesis being no difference, “no effect”).
- $H_0: \mu_2 - \mu_1 = 0$ ($\mu_1 = \mu_2$)
- $H_a: \mu_2 - \mu_1 \neq 0$ ($\mu_1 \neq \mu_2$)
- Test Statistic:

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se} = \frac{\bar{y}_2 - \bar{y}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Test statistic has usual form of

(estimate of parameter – null hypothesis value)/standard error.

- P -value: 2-tail probability from t distribution
- For 1-sided test (such as $H_a: \mu_2 - \mu_1 > 0$), P -value = one-tail probability from t distribution (but, not robust)
- Interpretation of P -value and conclusion using α -level same as in one-sample methods

(e.g., Suppose P -value = 0.58. Then, under supposition that null hypothesis true, probability = 0.58 of getting data like observed or even “more extreme,” where “more extreme” determined by H_a)

Example: Comparing female and male mean number of close friends, $H_0: \mu_1 = \mu_2$ $H_a: \mu_1 \neq \mu_2$

Difference between sample means = $8.9 - 8.3 = 0.6$

$se = 1.09$ (as in CI calculation)

Test statistic $t = 0.6/1.09 = 0.55$

P -value = $2(0.29) = 0.58$.

If null hypothesis true of equal population means, would not be unusual to get samples such as observed.

For $\alpha = 0.05$, not enough evidence to reject null.

It is plausible that the population means are identical.

For $H_a: \mu_1 < \mu_2$, P -value = 0.29

For $H_a: \mu_1 > \mu_2$ P -value = $1 - 0.29 = 0.71$

Equivalence of CI and Significance Test

“ $H_0: \mu_1 = \mu_2$ rejected (not rejected) at α -level in favor of
 $H_a: \mu_1 \neq \mu_2$ ”

equivalent to

“100(1 - α)% CI for $\mu_1 - \mu_2$ does not contain 0 (contains 0)”

Example for $\alpha = 0.05$: P -value = 0.58, so

“we do not reject H_0 of equal population means”

95% CI of (-1.5, 2.7) contains 0.

Alternative inference comparing means assumes equal population standard deviations

- We will not consider formulas for this approach here (in Sec. 7.5 of text), as it's a special case of "analysis of variance" methods studied later in Chapter 12.

This CI and test uses t distribution with

$$df = n_1 + n_2 - 2$$

- We will see how software displays this approach and the one we've used that does not assume equal population standard deviations.

Example: Exercise 7.30, p. 213. Improvement scores for
therapy A: 10, 20, 30
therapy B: 30, 45, 45

A: mean = 20, $s_1 = 10$

B: mean = 40, $s_2 = 8.66$

Data file, which we input into SPSS and analyze

Subject	Therapy	Improvement
1	A	10
2	A	20
3	A	30
4	B	30
5	B	45
6	B	45

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
improvement	Equal variances assumed	.000	1.000	-2.619	4	.059
	Equal variances not assumed			-2.619	3.920	.060

Independent Samples Test

		t-test for Equality of Means			
		Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
				Lower	Upper
improvement	Equal variances assumed	-20.00000	7.63763	-41.20545	1.20545
	Equal variances not assumed	-20.00000	7.63763	-41.37722	1.37722

Test of $H_0: \mu_1 = \mu_2$ $H_a: \mu_1 \neq \mu_2$

Test statistic $t = (40 - 20)/7.64 = 2.62$

When $df = 4$, P -value = $2(0.0294) = 0.059$.

For one-sided $H_a: \mu_1 < \mu_2$ (i.e., predict before study that therapy B is better), P -value = 0.029

With $\alpha = 0.05$, insufficient evidence to reject null for two-sided H_a , but can reject null for one-sided H_a and conclude therapy B better.

(but remember, must choose H_a ahead of time!)

How does software get df for “unequal variance” method?

- When allow $\sigma_1^2 \neq \sigma_2^2$ recall that

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- The “adjusted” degrees of freedom for the t distribution approximation is (Welch-Satterthwaite approximation) :

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

Some comments about comparing means

- *One-sided t* tests are *not* robust against severe violations of the normality assumption, when n is relatively small. (Better to use “nonparametric” methods (which do not assume a particular form of population distribution) for one-sided inference when normal population assumption is badly violated, invalidating t inferences; see text Sec. 7.7)
- CI shows whether plausible values are near or far from H_0 in practical terms.

- When groups have similar variability, a summary measure of *effect size* is

$$\text{effect size} = \frac{\text{mean}_1 - \text{mean}_2}{\text{standard deviation in each group}}$$

- **Example:** The therapies had sample means of 20 for A and 40 for B and standard deviations of 10 and 8.66. If the standard deviation in each group is 9 (say), then

$$\text{effect size} = (20 - 40)/9 = -2.2$$

Mean for therapy B estimated to be about two std. dev's larger than the mean for therapy A, a *large* effect.

Example: Which study showed the largest effect?

1. $\bar{y}_1 = 20, \bar{y}_2 = 30, s = 10$

2. $\bar{y}_1 = 200, \bar{y}_2 = 300, s = 100$

3. $\bar{y}_1 = 20, \bar{y}_2 = 25, s = 2$

Comparing Means with Dependent Samples

- Setting: Each sample has the same subjects (as in longitudinal studies or crossover studies) or *matched pairs* of subjects
- Then, it is *not* true that for comparing two statistics,

$$se = \sqrt{(se_1)^2 + (se_2)^2}$$

- Must allow for “correlation” between estimates (Why?)
- Data: y_i = difference in scores for subject (pair) i
- Treat data as single sample of difference scores, with sample mean \bar{y}_d and sample standard deviation s_d and parameter μ_d = population mean difference score.
- In fact, $\mu_d = \mu_2 - \mu_1$

Example: Cell-phone study also had experiment with same subjects in each group (data on p. 194 of text)

For this “matched-pairs” design, data file has the form

Subject	Cell_no	Cell_yes
1	604	636
2	556	623
3	540	615
... (for 32 subjects)		

Sample means are 534.6 milliseconds without cell phone
585.2 milliseconds, using cell phone

We reduce the 32 observations to 32 difference scores,

$$636 - 604 = 32$$

$$623 - 556 = 67$$

$$615 - 540 = 75$$

....

and analyze them with standard methods for a single sample

$$\bar{y}_d = 50.6 = 585.2 - 534.6, \quad s_d = 52.5 = \text{std dev of } 32, 67, 75 \dots$$

$$se = s_d / \sqrt{n} = 52.5 / \sqrt{32} = 9.28$$

For a 95% CI, $df = n - 1 = 31$, t -score = 2.04

We get $50.6 \pm 2.04(9.28)$, or (31.7, 69.5)

- We can be 95% confident that the population mean using a cell phone is between 31.7 and 69.5 milliseconds higher than without cell phone.
- For testing $H_0 : \mu_d = 0$ against $H_a : \mu_d \neq 0$, the test statistic is

$$t = (\bar{y}_d - 0)/se = 50.6/9.28 = 5.5, df = 31,$$

Two-sided P -value = 0.000005, so there is extremely strong evidence against the null hypothesis of no difference between the population means.

In class, we will use SPSS to

- Run the dependent-samples t analyses
- Plot cell_yes against cell_no and observe a strong positive correlation (0.814), which illustrates how an analysis that ignores the dependence between the observations would be inappropriate.
- Note that one subject (number 28) is an outlier (unusually high) on both variables
- With outlier deleted, SPSS tell us that $t = 5.26$, $df = 30$ for comparing means ($P = 0.00001$) for comparing means, 95% CI of (29.1, 66.0). The previous results were not influenced greatly by the outlier.

SPSS output for original dependent-samples t analysis (including the outlier)

Paired Samples Test						
		Paired Differences				
					95% Confidence Interval of the Difference	
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper
Pair 1	cell_yes - cell_no	50.62500	52.48579	9.27826	31.70186	69.54814

Paired Samples Test				
		t	df	Sig. (2-tailed)
Pair 1	cell_yes - cell_no	5.456	31	.000

Some comments

- Dependent samples have advantages of (1) controlling sources of potential bias (e.g., balancing samples on variables that could affect the response), (2) having a smaller *se* for the difference of means, when the pairwise responses are highly positively correlated (in which case, the difference scores show less variability than the separate samples)
- With dependent samples, why can't we use the *se* formula for independent samples?

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Ex. (artificial, but makes the point)

Weights before and after anorexia therapy

Subject	Before	After	Difference
1	115	122	7
2	91	98	7
3	100	107	7
4	132	139	7

...

Lots of variability for each group of observations, but *no* variability for the difference scores

If you plot $x =$ weight before against $y =$ weight after, what do you see?

- The **McNemar test** (pp. 201-203) compares proportions with dependent samples
- **Fisher's exact test** (pp. 203-204) compares proportions for small independent samples
- Sometimes it's more useful to compare groups using *ratios* rather than *differences* of parameters

Example: U.S. Dept. of Justice reports that proportion of adults in prison is about 900/100,000 for males, 60/100,000 for females

Difference: $900/100,000 - 60/100,000 = 840/100,000 = 0.0084$

Ratio: $[900/100,000]/[60/100,000] = 900/60 = 15.0$

In applications in which the proportion refers to an undesirable outcome (e.g., most medical studies), the ratio is called the *relative risk*

A few summary questions

1. Give an example of (a) independent samples, (b) dependent samples
2. Give an example of (a) response var., (b) categorical explanatory var., and identify whether response is quantitative or categorical and state the appropriate analyses.
3. Suppose that a 95% CI for difference between Massachusetts and Texas in the population proportion supporting legal same-sex marriage is (0.15, 0.22).
 - a. Population proportion of support is higher in Texas
 - b. Since 0.15 and 0.22 < 0.50, less than half the population supports legal same-sex marriage.
 - c. The 99% CI could be (0.17, 0.20)
 - d. It is plausible that population proportions are equal.
 - e. P-value for testing equal population proportions against two-sided alternative could be 0.40.
 - f. We can be 95% confident that the sample proportion of support in MA is between 0.15 higher and 0.22 higher than in TX.