

CHAPTER EIGHT

Tests of statistical significance

I. Conceptual exercises

A. Matching

1. e
2. i
3. f
4. g
5. d
6. h
7. a
8. c
9. b

B. True or false

1. T
2. F
3. F
4. T
5. F
6. F
7. T
8. T
9. F

C. Hypothetical situations

1. Language majors vs. non-majors

a. Null hypothesis:

$H_0: \bar{X}_{\text{majors}} = \bar{X}_{\text{non-majors}}$ The means of the majors and non-majors will be equal.

$H_1: \bar{X}_{\text{majors}} > \bar{X}_{\text{non-majors}}$ The mean of the majors will be larger than that of the non-majors.

b. Confidence intervals, based on z-scores, would be the most appropriate statistical test because the sample sizes are very large. The T-test could also be

used, but the appropriateness of this would depend on the comparability of the two groups, in terms of sample size and variances.

c. This is a multiple-group design.

d. We would probably use the 95% CI, if no decisions other than testing a research hypothesis will be made on the basis of the results. If the results will affect decisions about programs or the assessment, then the 99% CI would be more appropriate.

e. We would use a one-tailed probability, since we expect the mean of the majors to be larger than the mean of the non-majors.

f. We could not necessarily attribute the difference to the status—majors versus non-majors—of the students because we don't know how much French the students may already have known before they enrolled in the classes.

2. Listening comprehension—audio input only vs. audio and video input.

a. The null and research hypotheses would be as follows:

- If you have no reason to hypothesize that the video will either detract from or enhance performance:

$H_0: \bar{X}_{\text{audio-only}} = \bar{X}_{\text{audio+video}}$ The means of the two tests will be equal.

$H_1: \bar{X}_{\text{audio-only}} \neq \bar{X}_{\text{audio+video}}$ The means of the two tests will not be equal.

- If you hypothesize that the video will enhance performance:

$H_0: \bar{X}_{\text{audio-only}} = \bar{X}_{\text{audio+video}}$ The means of the two tests will be equal.

$H_1: \bar{X}_{\text{audio-only}} < \bar{X}_{\text{audio+video}}$ The mean of the audio plus video test will be larger than that of the audio-only test.

- If you hypothesize that the video will detract from performance:

$H_0: \bar{X}_{\text{audio-only}} = \bar{X}_{\text{audio+video}}$ The means of the two tests will be equal.

$H_1: \bar{X}_{\text{audio-only}} > \bar{X}_{\text{audio+video}}$ The mean of the audio plus video test will be smaller than that of the audio-only test.

- b. The t-test is the appropriate statistical test because the sample is small, and the same students are in both groups.
 - c. This is a single-group design.
 - d. We would use the 99% CI because we will make a decision about which format of input to use based on the results.
 - e. If we have no reason to expect performance on one test to be better than on the other, we will use a two-tailed probability. If we think the video input will enhance performance, then we will use a one-tailed test.
 - f. If all facets of measurement for the two tests (e.g. content, length of passage, administration) were the same except for the presence or absence of video input, then we could probably conclude that the difference is due to the additional video.
3. Performance of students from different disciplines (humanities, social sciences, life sciences and business) on English placement test
- a. The null and research hypotheses are as follows:

$H_0: \bar{X}_H = \bar{X}_{SS} = \bar{X}_{LS} = \bar{X}_B$ The means of all four groups will be equal.

$H_1: \bar{X}_H \neq \bar{X}_{SS} \neq \bar{X}_{LS} \neq \bar{X}_B$ The means of the four groups will not be equal.

- b. The appropriate statistical test is ANOVA. because more than two different groups are being compared.
- c. This is a multiple-group design.
- d. If we may make decisions about using different cut scores for placement for the different groups, then the 99% CI would be appropriate. If not, then 95% would be appropriate.
- e. We would use a two-tailed probability because we have no reason to expect the means of the groups to be ordered in any particular way.
- f. If the overall F-ratio is significant, then we would perform *post hoc* analyses. Even if these were significant for all groups, we could not necessarily attribute the differences to differences in the disciplines. If we have not taken a random sample of students from the entire student population, other characteristics of the students, such as gender, age, cultural background and native language might affect their performance.

III. Hand calculations with small data sets

- A. Use the data sets below to calculate z-tests and t-tests. Assume non-directional hypotheses and use .95 confidence level.

Table 8.1 *Data set 1*

	Group 1	Group 2
N	932	840
\bar{X}	34.7	36.2
s	5.95	6.08

$$s_{\bar{X}_1} = \frac{s_1}{\sqrt{N}} = \frac{5.95}{\sqrt{932}} = .195$$

$$s_{\bar{X}_2} = \frac{s_2}{\sqrt{N}} = \frac{6.08}{\sqrt{840}} = .210$$

$$CI_{.95} = \bar{X}_1 \pm 1.96 s_{\bar{X}_1} = 34.7 \pm .3822 = 34.32 - 35.08$$

$$CI_{.95} = \bar{X}_2 \pm 1.96 s_{\bar{X}_2} = 36.2 \pm .4116 = 35.79 - 36.61$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_{\bar{X}_1}^2 + s_{\bar{X}_2}^2}} = \frac{34.7 - 36.2}{\sqrt{.038 + .044}} = \frac{-1.5}{.286} = 5.245, \text{ sig at } p < .001, \text{ df} = 1,770 \text{ (two-tailed)}$$

We report the absolute difference between the two means.

NB: The notation “<<” is sometimes used to mean “much less than” when the test statistic is much larger than the table value for the .001 significance level. In this case, the obtained t of 5.245 is much larger than the table value of 3.373 for a significance level of .001 for a two-tailed test with df = 120 in Table B in the textbook.

1. Could the observed difference between the means be due to chance, and why?

No. Neither mean falls within the other's confidence interval, and the t-test is highly significant.

2. How sure are you of this?

95% using CIs, and 99.9% using the t-test.

Table 8.2 *Data set 2*

	Group 3	Group 4
N	36	40
\bar{X}	34.7	36.2
s	5.95	6.08

$$s_{\bar{X}_1} = \frac{s_1}{\sqrt{N}} = \frac{5.95}{\sqrt{36}} = .992$$

$$s_{\bar{X}_2} = \frac{s_2}{\sqrt{N}} = \frac{6.08}{\sqrt{40}} = .961$$

$$CI_{.95} = \bar{X}_1 \pm 1.96 s_{\bar{X}_1} = 34.7 \pm 1.94 = 32.76 - 36.64$$

$$CI_{.95} = \bar{X}_2 \pm 1.96 s_{\bar{X}_2} = 36.2 \pm 1.88 = 34.32 - 38.08$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_{\bar{X}_1}^2 + s_{\bar{X}_2}^2}} = \frac{34.7 - 36.2}{\sqrt{.984 + .924}} = \frac{-1.5}{1.381} = 1.086, \text{ NS, } df = 74$$

3. Could the observed difference between the means be due to chance, and why?

Yes. Both means fall within the other's confidence intervals, and the t-test is not significant.

4. How sure are you of this?

95%.

Table 8.3 *Data set 3*

	Group 5	Group 6
N	932	840
\bar{X}	34.7	36.2
s	15.62	16.25

$$s_{\bar{X}_1} = \frac{s_1}{\sqrt{N}} = \frac{15.62}{\sqrt{932}} = .512$$

$$s_{\bar{X}_2} = \frac{s_2}{\sqrt{N}} = \frac{16.25}{\sqrt{840}} = .561$$

$$CI_{.95} = \bar{X}_1 \pm 1.96 s_{\bar{X}_1} = 34.7 \pm 1.00 = 33.70 - 35.70$$

$$CI_{.95} = \bar{X}_2 \pm 1.96 s_{\bar{X}_2} = 36.2 \pm 1.10 = 35.10 - 37.30$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_{\bar{X}_1}^2 + s_{\bar{X}_2}^2}} = \frac{34.7 - 36.2}{\sqrt{.262 + .315}} = \frac{-1.5}{.760} = -1.974, \text{ NS, df} = 1,770 \text{ (two-tailed)}$$

(Note that the obtained t, 1.974, is slightly smaller than the table value of 1.980 for df = 120. If we used the row for df = ∞ , it would be significant at $p \leq .05$, since it is bigger than the critical value of 1.960. This is a case where it would be preferable to calculate t from the raw data, using SPSS, which would provide an exact probability.)

5. Could the observed difference between the means be due to chance, and why?

Based on CIs, no, because neither mean falls within the other's confidence interval. Based on the t-value, yes, since it is not significant.

6. How sure are you of this?

95%.

Table 8.4 Data set 4

	Group 7	Group 8
N	36	40
\bar{X}	34.7	36.2
s	15.62	16.25

$$s_{\bar{X}_1} = \frac{s_1}{\sqrt{N}} = \frac{15.62}{\sqrt{36}} = 2.603$$

$$s_{\bar{X}_2} = \frac{s_2}{\sqrt{N}} = \frac{16.25}{\sqrt{40}} = 2.569$$

$$CI_{.95} = \bar{X}_1 \pm 1.96 s_{\bar{X}_1} = 34.7 \pm 5.10 = 29.60 - 39.80$$

$$CI_{.95} = \bar{X}_2 \pm 1.96 s_{\bar{X}_2} = 36.2 \pm 5.04 = 31.16 - 41.24$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_{\bar{X}_1}^2 + s_{\bar{X}_2}^2}} = \frac{34.7 - 36.2}{\sqrt{6.776 + 6.600}} = \frac{-1.5}{3.657} = -.410, \text{ NS, df} = 74$$

7. Could the observed difference between the means be due to chance, and why?

According to the CI, yes, since both means fall within the other's confidence interval. According to the t-test, yes, since the value for t is not significant.

8. How sure are you of this?

95%.

Summary of z- and t-tests for data sets 1 – 4

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
N	932	840	36	40	932	840	36	40
\bar{X}	34.7	36.2	34.7	36.2	34.7	36.2	34.7	36.2

s	5.95	6.08	5.95	6.08	15.62	16.25	15.62	16.25
$s_{\bar{x}}$	0.195	0.210	0.992	0.961	0.512	0.561	2.603	2.569
z CI_{.95} \bar{X}_1	34.32 – 35.08		32.76 – 36.64		33.70 – 35.70		29.60 – 39.80	
z CI_{.95} \bar{X}_2	35.79 – 36.61		34.32 – 38.08		35.10 – 37.30		31.16 – 41.24	
reject H_0?	Yes		No		No		No	
t	5.245		1.086		1.974		0.410	
p	< < .001		NS		NS		NS	
reject H_0?	Yes		No		No		No	

B. Looking at the values you have calculated, answer the following questions:

1. What happens to the values for the $s_{\bar{x}}$ as N gets smaller?

$s_{\bar{x}}$ gets larger as N gets smaller.

2. What happens to the values for the $s_{\bar{x}}$ as s gets larger?

$s_{\bar{x}}$ gets larger as s gets larger.

3. What happens to the values for t as the N gets smaller?

t gets smaller as N gets smaller.

4. What happens to the values for t as s gets larger?

t gets smaller as s gets larger.

5. Do you reach the same conclusions for all four data sets? Why or why not?

No. For data set 1, we can reject the null hypothesis, while for data sets 2 and 4 we cannot. For data set 3, the z-test rejects the null hypothesis whereas the t-test cannot. Even though the means are the same for all these data sets, the sample sizes for data sets 2 and 4 are much smaller than those for data sets 1 and 3. Because they have smaller sample sizes, the standard errors of the means for

groups 2 and 4 are larger, and the CIs are thus wider. Similarly, the smaller sample sizes yield smaller t-values.

6. Do you reach the same conclusions based on the z-tests and t-tests? Why or why not?

With the exception of data set 3, yes. With data set 3, the table does not provide sufficient precision, in terms of the degrees of freedom, so that the obtained t-value falls just short of significance at $p \leq .05$.

V. SPSS exercises

The exercises below are based on the CTCS data set, 'CTCS.sav'. The SPSS programs for doing the analyses for all of these exercises are in the file 'CTCS-SPSS exercises.sps', while the results of the SPSS analyses are in the file 'CTCS-SPSS exercises.spo'.

A. CTCS: TOEFL preparation courses and performance TOEFL total.

1. State your hypotheses:

$H_0: \bar{X}_{T\text{-rep}} = \bar{X}_{\text{no-Tprep}}$ The TOEFL score means of the TOEFL preparation and no-preparation groups will be the same.

$H_1: \bar{X}_{T\text{-rep}} > \bar{X}_{\text{no-Tprep}}$ The TOEFL score mean of the TOEFL preparation group will be larger than that for the group with no TOEFL preparation.

2. What is the appropriate statistical test? Why?

Confidence intervals, based on z-scores, because the sample sizes are very large. Also, the sample sizes of the two groups are not the same.

3. What level of confidence will you use, and why?

95%, since it isn't clear that any decisions about the test or about people will be based on the results.

4. Will you use a one-tailed or two-tailed probability, and why?

One-tailed, because the research hypothesis is directional, that the TOEFL preparation group will perform better than the group with no preparation.

5. Are assumptions of the statistical test met? Why or why not?

Looking at the skewness and kurtosis of the two groups' scores in the results of the EXPLORE procedure, the distributions of both groups appear to be reasonably normal. The histograms, on the other hand, reveal that the distribution for the no-preparation group is nearly bimodal, with two peaks.

6. What are the results of your statistical test?

	No preparation	Preparation
N	110	1042
\bar{X}	487.68	508.60
s	65.502	55.887
$s_{\bar{X}}$	6.245	1.731
z CI_{.95} \bar{X}_1	475.30 – 500.06	
z CI_{.95} \bar{X}_2	505.20 – 512.00	
reject H_0?	yes	
t	3.228 (equal variances not assumed)	
p	p= .002	
reject H_0?	yes	

7. Can you reject the null hypothesis? Explain why or why not?

Yes, neither mean is within the other's CI_{.95}, and the t-value is significant at p = .002, df = 126

B. SPEAK and FCE Paper 5 ratings for Pronunciation.

1. State your hypotheses:

H_0 : $\bar{X}_{\text{SPEAK-pron}} = \bar{X}_{\text{FCE5-pron}}$ The means for SPEAK and FCE 5 ratings for pronunciation will be equal.

H_1 : $\bar{X}_{\text{SPEAK-pron}} \neq \bar{X}_{\text{FCE5-pron}}$ The means for SPEAK and FCE 5 ratings for pronunciation will not be equal.

2. Independent or dependent t-test? Why?

Dependent, because it's a single-group design—the same people took both tests.

3. What level of confidence will you use, and why?

95% because no decisions will be made about the test or people

4. Will you use a one-tailed or two-tailed probability, and why?

Two-tailed, because the research hypothesis is non-directional.

5. Are assumptions of the statistical test met? Why or why not?

Looking at the skewness and kurtosis, as well as the histograms, both distributions appear to be reasonably normal. However, because the two measures appear to be on different scales (SPEAK Pronunciation scores range from 0 – 3, while FCE 5 scores range from 0 – 5), the assumption of homogeneous variances needs to be checked. However, given the fact that this is a paired-sample, or single-group design, with equal sample sizes for each set of test scores, the t-test is fairly robust to differences in variances.

6. What is the value and significance of the t-ratio you obtained?

$t = 58.884$, $df = 1,282$, sig at $p = .000$ (two-tailed)

7. Can you reject the null hypothesis? Explain why or why not:

Yes, t is significant at $p = .000$.

- C. Differences in TOEFL total score across the eight different countries in which the CTCS was conducted.

1. State your hypotheses:

$H_0: \bar{X}_T = \bar{X}_E = \bar{X}_J = \bar{X}_H = \bar{X}_{Sp} = \bar{X}_B = \bar{X}_F = \bar{X}_{Sw}$ The means are all the same.

$H_1: \bar{X}_T \neq \bar{X}_E \neq \bar{X}_J \neq \bar{X}_H \neq \bar{X}_{Sp} \neq \bar{X}_B \neq \bar{X}_F \neq \bar{X}_{Sw}$ The means are all different.

2. What level of confidence will you use, and why?

95%, because no decisions will be made about individuals or the test on the basis of these results.

3. (Omitted)

4. Will you use a one-tailed or two-tailed probability, and why?

Two-tailed, since we have no reason to expect an ordering of the means across the sites.

5. Are assumptions of the ANOVA met? Why or why not?

Looking at the skewness and kurtosis of the scores from the different sites, as well as the histograms, the distributions appear to be reasonably normal. However, the test of homogeneity of variance is significant, so we will need to use the robust ANOVA statistics. If the initial F is significant, we will need to use *post-hoc* comparisons that are robust to violation of this assumption or that do not assume equal variances.

The tests were administered under standard conditions, so there is no reason to believe that the assumption of independence of responses was violated.

These are independent groups, from different countries.

6. What are the results of your ANOVA?

- a. F-ratio and significance for overall group differences

Welch's F' is 66.838 and the Brown-Forsythe F* is 75.337. Both are significant at $p = .000$.

- b. If the overall F-ratio was significant, what differences across the different regions were significant?

Region	TOEFL total mean	Sig diffs	
		Greater than	Lesser than
1. France	539.39	4,5,6,7,8	
2. Switzerland	534.85	5,6,7,8	
3. Chinese Hong Kong	528.88	5,6,7,8	
4. Brazil	520.86	6,7,8	1
5. Spain	511.18	6,7,8	1,2,3
6. Thailand	469.22		1,2,3,4,5
7. Egypt	465.98		1,2,3,4,5
8. Japan	457.20		1,2,3,4,5

c. Pattern of group differences

Both the Tamhane paired contrasts and the Ryan-Einot-Gabriel-Welsh F ranges indicate the groupings below, from high to low. Solid lines indicate significant differences, while dotted lines indicate overlap in groupings.

France
Switzerland

Chinese Hong Kong

Brazil
Spain

Thailand
Egypt
Japan