Hypothesis Testing

Recall: the inferential step to conclude that the null hypothesis is false goes as follows: The data (or data more extreme) are very unlikely given that the null hypothesis is true. This means that:

(1) a very unlikely event occurred or

(2) the null hypothesis is false.

The inference usually made is that the null hypothesis is false. Importantly it doesnt prove the null hypothesis to be false.

Type I and II errors

There are two kinds of errors that can be made in hypothesis testing:

(1) a true null hypothesis can be incorrectly rejected

(2) a false null hypothesis can fail to be rejected.

The former error is called a ***Type I error*** and the latter error is called a ***Type II error***.

The probability of Type I error is always equal to the level of significance α (alpha) that is used as the standard for rejecting the null hypothesis .

Type II Error

- The probability of a Type II error is designated by the Greek letter beta (β).
- A Type II error is only an error in the sense that an opportunity to reject the null hypothesis correctly was lost.
- It is not an error in the sense that an incorrect conclusion was drawn since no conclusion is drawn when the null hypothesis is not rejected.

Types of Error

- A Type I error, on the other hand, is an error in every sense of the word. A conclusion is drawn that the null hypothesis is false when, in fact, it is true.
- Therefore, Type I errors are generally considered more serious than Type II errors.
- The probability of a Type I error (α) is set by the experimenter.
- There is a trade-off between Type I and Type II errors. The more an experimenter protects himself or herself against Type I errors by choosing a low level, the greater the chance of a Type II error.

Types of Error

- Requiring very strong evidence to reject the null hypothesis makes it very unlikely that a true null hypothesis will be rejected.
- However, it increases the chance that a false null hypothesis will not be rejected, thus increasing the likelihood of Type II error.
- The Type I error rate is almost always set at 0.05 or at 0.01, the latter being more conservative since it requires stronger evidence to reject the null hypothesis at the 0.01 level then at the 0.05 level.
- **Important**In this module, the significance level α can be assumed to be 0.05, unless explicitly stated otherwise.

Type I and II errors

These two types of errors are defined in the table below.

	True State: H0 True	True State: H0 False
Decision: Reject H0	Type I error	Correct
Decision: Do not Reject H0	Correct	Type II error

p-values

- In hypothesis tests, the difference between the observed value and the parameter value specified by H_0 is computed and the probability of obtaining a difference this large or large is calculated.
- The probability of obtaining data as extreme, or more extreme, than the expected value under the null hypothesis is called the ***p-value***.
- It is not the probability of the null hypothesis itself.
- Suppose if the probability value is 0.0175, this does not mean that the probability that the null hypothesis is either true (or false) is 0.0175.

p-values

- the p-value means that the probability of obtaining data as different or more different from the null hypothesis as those obtained in the experiment is 0.0175.
- If the p-value is less than the specified significance level, adjusted for the number of tails, then we reject the null hypothesis.

$$\text{is p-value} \leq \frac{\alpha}{k}?$$

The Hypothesis Testing Procedure

The second procedures is very similar to the first, but is more practicable for written exams, so we will use this one more. The first two steps are the same.

- Formally write out the null and alternative hypotheses (already described).
- Compute the test statistic
- Determine the ***critical value*** (described shortly)
- Make a decision based on the critical value.

Hypothesis Tests for single samples

- We could have inference procedures for single sample studies. We would base an argument on the either the sample mean or sample proportion as appropriate.
- A hypothesis test can be used to determine how “confident” we can be with our data in making that statements.
- The lower the significance level (The margin for Type I error) the stronger our data must be.
- Large samples lead to more confident conclusion.

Hypothesis Tests for single samples

- We could have either hypothesis test for the sample mean or the sample proportion, to test a statement about the population as a whole (i.e something about the population mean)
- We make our argument in the form of the null and alternative hypotheses.
- The Hypothesis testing procedure determines the strength of evidence in making our arguments.

Hypothesis Tests for single samples

- We simply follow the four step procedure.
- All of the components are the same used in confidence intervals.
- The critical value is simply a quantile from the Z or $t-$distribution.
- The standard errors are also as before. Although when performing a hypothesis test for proportions, we use the expected value under the null hypothesis, rather than point estimate. (reason beyond scope of course.)

Example 1 (a) Small Sample Hypothesis Test

- The manufacturer claims that average tube life for a particular brand of ultraviolet tube is 9,000 hr.
- Test this claim at the 5 percent level of significance against the alternative hypothesis that the mean life is not 9,000 hr
- We are given the following information: a sample of $n = 10$ tubes the mean operating life was $\bar{x} = 8,800$ hr. The sample standard deviation is be $s = 500$ hr.

Example 1 (b)

- $H_0: \ \mu = 9000$ (Average life span is 9000 hours.)
- $H_1: \ \mu \neq 9000$ (Average life span is not 9000 hours.)

- The observed difference is -200 hours. (i.e. 8,800 - 9,000 hours)
- The standard error is determined from formulae.

$$S.E.(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{500}{\sqrt{10}} = 158.1139$$

Example 1 (c) : Test Statistic and Critical Value

$$TS = \frac{8800 - 9000}{158.11}$$

- The test statistic $TS = -1.265$
- The CV is determined with $\alpha = 0.05$ and $k = 2$ (column = $\alpha/k = 0.025$).
- The sample is small n = 10 $df = n - 1 = 9$ (i.e. row =9).
- Therefore $CV = 2.262$
- (Remark: If the sample was large, we could use $CV = 1.96$).

Example 1 (d): Decision Rule

- **Decision:**Is $|TS| > CV$? Is $1.265 > 2.262$?
- No. We fail to reject the null hypothesis.
- There is not enough evidence to say that the mean lifespan is not 9000 hours.

The Paired t-test

A paired t-test is used to compare two population means where you have two samples in which observations in one sample can be ***paired*** with observations in the other sample.

Examples of where this might occur are:

- Before-and-after observations on the same subjects (e.g. students diagnostic test results before and after a particular module or course).
- A comparison of two different methods of measurement or two different treatments where the measurements/treatments are applied to the ***same*** subjects.

The difference between two paired measurements is known as a ***case-wise*** difference.

The Paired t-test

- We will often be required to compute the case-wise differences, the average of those differences and the standard deviation of those difference.
- The mean difference for a set of differences between paired observations is

$$\bar{d} = \frac{\sum d_i}{n}$$

- The computational formula for the standard deviation of the differences between paired observations is

$$s_d = \sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n-1}}$$

- It is nearly always a small sample test.

The Paired t-test

- μ_d mean value for the population of case-wise differences.
- The null hypothesis is that that $\mu_d = 0$
- Given $\bar{d}$ mean value for the sample of differences, and s_d standard deviation of the differences for the paired sample data, we can compute this test in the same manner as a one-sample test for the mean

Example 2: Paired Difference (a)

- An automobile manufacturer collects mileage data for a sample of $n = 10$ cars in various weight categories using a standard grade of gasoline with and without a particular additive.
- Of course, the engines were tuned to the same specifications before each run, and the same drivers were used for the two gasoline conditions (with the driver in fact being unaware of which gasoline was being used on a particular run).
- Given the mileage data on the next slide, test the hypothesis that there is no difference between the mean mileage obtained with and without the additive, using the 5 percent level of significance

Example 2: Paired Difference (b)

car	with additive	without additive	d_i	d_i^2
1	36.7	36.2	0.5	0.25
2	35.8	35.7	0.1	0.01
3	31.9	32.3	-0.4	0.16
4	29.3	29.6	-0.3	0.09
5	28.4	28.1	0.3	0.09
6	25.7	25.8	-0.1	0.01
7	24.2	23.9	0.3	0.09
8	22.6	22.0	0.6	0.36
9	21.9	21.5	0.4	0.16
10	20.3	20.0	0.3	0.09

Example 2: Paired Difference (c)

- The average of the case wise differences is computed as

$$\bar{d} = \frac{\sum d_i}{n}$$

$$\bar{d} = \frac{0.05 + 0.1 - 0.4 + \ldots + 0.30}{10} = 0.17$$

- Also, using last column, $\sum d_i^2 = (0.25 + 0.01 + 0.16 + \ldots + 0.09) = 1.31$

Example 2: Paired Difference (d)

Sample standard deviation of the case-wise differences:

$$s_d = \sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n-1}}$$

We know the following:

- The sample size n which is 10.
- The average of the case-wise differences. $\bar{d} = 0.17$
- $\sum d_i^2 = 1.31$

Example 2: Paired Difference (e)

Sample standard deviation of the case-wise differences://

$$s_d = \sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n-1}}$$

$$s_d = \sqrt{\frac{1.31 - 10(0.17)^2}{9}} = 0.337$$

The standard error:

$$S.E.(\bar{d}) = s_d/\sqrt{n} = \frac{0.0337}{3.16} = 0.107$$

Example 2: Paired Difference (f)

Null and Alternative Hypotheses:

- That is, the null hypothesis is:
  $H_0 : \mu_d = 0$ Additive makes no difference to performance
  $H_1 : \mu_d \neq 0$ Additive makes a significant difference to performance

Test Statistic:

$$TS = \frac{0.17 - 0}{0.107} = 1.59$$