Two Samples Hypothesis Testing

Author: John M. Cimbala, Penn State University Latest revision: 10 February 2010

Introduction

- In a previous learning module, we discussed how to perform hypothesis tests for a *single* variable *x*.
- Here, we extend the concept of hypothesis testing to the comparison of *two* variables x_A and x_B .

Two Samples Hypothesis Testing when *n* is *the same* for the two Samples

Two-tailed paired samples hypothesis test:

- In engineering analysis, we often want to test whether some *modification* to a system causes a *statistically significant change* to the system (the system is either improved or made worse).
- We conduct some experiments in which the sample mean \overline{x}_A of sample *A* (without the modification) is indeed *different* than the sample mean \overline{x}_B of sample *B* (with the modification). In other words, the modification appears to have led to a change, but is the change statistically significant?
- Here we discuss the simplest such statistical test a test of whether one sample of data has a significantly different predicted population mean compared to a second sample of data, and with the number of data points *n* being the *same* in the two samples.
- Statisticians refer to this case (equal *n* in the two samples) as a *paired samples hypothesis test*.
- The procedure is very similar to the single-sample hypothesis tests we have already discussed, except that we replace variable x by the *difference between the two variables*, $\delta = x_B x_A$.
- In a *two-tailed paired-samples hypothesis test*, we want to know whether there is a statistically significant *change* in the predicted population means of the two samples. We don't care if the change is positive or negative in a two-tailed hypothesis test we are concerned only about whether there is a change.
- From the definition of variable δ , we see that an appropriate null hypothesis is $\delta = 0$, i.e., there is *no change* in the population mean between the two samples (the least likely scenario). Thus, we set: [This is a *two*-tailed hypothesis test.]
 - *Null hypothesis*: Critical value is $\mu_0 = 0$; the least likely scenario is $\mu = \mu_0$ (there is *no* statistically significant change in the population means). [This is the least likely scenario since $\overline{x}_A \neq \overline{x}_B$.]
 - *Alternative hypothesis*: (opposite of the null hypothesis), $\mu \neq \mu_0$. In other words, either $\mu < \mu_0$ or $\mu > \mu_0$ (there *is* a statistically significant change in the population means). [This is the most likely scenario since $\overline{x}_A \neq \overline{x}_B$.]
- The critical *t*-statistic is calculated as previously, but using the sample mean of δ instead of *x*, and the sample standard deviation of δ instead of *x*, i.e., $t = \frac{\overline{\delta} \mu_0}{S_{\delta} / \sqrt{n}}$.
- The corresponding *p*-value is calculated as previously, based on the critical *t*-statistic. In this case we are considering a *two-tail hypothesis test*. *p* is calculated in Excel using the function **TDIST**(**ABS**(*t*),**df**,**2**), where df is the number of degrees of freedom, df = n 1, and the "2" specifies two tails.
- If Excel is not available, we can use tables; some modern calculators can also calculate the *p*-value.
- We formulate our conclusions (to 95% confidence level) based on the *p*-value:
 - If p < 0.05, we reject the null hypothesis because the least likely scenario ($\mu = \mu_0$) has less than a 5% chance of being true. Thus, we can state confidently that *there is a statistically significant change in the population mean of the variable, i.e.*, $\mu_A \neq \mu_B$.
 - If 0.05 , we*cannot reject or accept the null hypothesis* $because the least likely scenario (<math>\mu = \mu_0$) has more than a 5% chance of being true, but less than a 95% chance of being true. The *results are therefore inconclusive we should conduct more tests*.
 - If p > 0.95, we *accept the null hypothesis* because what we set as the least likely scenario ($\mu = \mu_0$) turns out to have more than a 95% chance of being true. Thus, we can state confidently that *there is no statistically significant change in the population mean of the variable, i.e.*, $\mu_A = \mu_B$.

<u>One-tailed paired samples hypothesis test</u>: [This is the more common one used in engineering analysis.]

- We assume here that our experiments yield $\overline{x}_B > \overline{x}_A$. In other words, the modification we made leads to an *improvement* in the mean between Sample A and Sample B. But is the improvement statistically significant?
- In a *one-tailed paired-samples hypothesis test*, we want to know whether there is a statistically significant *improvement* in the predicted population means of the two samples. From the definition of variable δ , we see

that an appropriate null hypothesis is $\delta < 0$, i.e., the modification caused the population mean between the two samples to *decrease* (the least likely scenario since we are assuming here that our experiments show that $\overline{x} > \overline{x}$). Thus, we get [This is a sum toiled hypothesis test].

- $\overline{x}_B > \overline{x}_A$). Thus, we set: [This is a *one*-tailed hypothesis test.]
- *Null hypothesis*: Critical value is $\mu_0 = 0$; the least likely scenario is $\mu < \mu_0$ (the population mean has decreased due to the modification, or $\mu_B < \mu_A$). [This is the least likely scenario since $\overline{x}_B > \overline{x}_A$.]
- Alternative hypothesis: $\mu > \mu_0$. In other words, there *is* a statistically significant increase in the population means, $\mu_B > \mu_A$). [This is the most likely scenario since $\overline{x}_B > \overline{x}_A$.]
- The critical *t*-statistic is calculated exactly as above for the two-tailed test.
- The corresponding *p*-value is calculated based on the critical *t*-statistic. In this case we are considering a *one-tail hypothesis test*. So, *p* is calculated in Excel using the function **TDIST**(**ABS**(*t*),**df**,**1**), where the "1" specifies one tail. You can also use the tables if Excel is not available; do *not* multiply *p* by 2 for a 1-tail test.
- For a one-tailed hypothesis test in which the null hypothesis is set to the least likely scenario, the *p*-value is limited in range from 0 to 0.5 (0% to 50%). Thus, we formulate our conclusions (to 95% confidence level) as follows:
 - If p < 0.05, we reject the null hypothesis because the least likely scenario ($\mu_B < \mu_A$) has less than a 5% chance of being true. Thus, we can state confidently that *there is a statistically significant increase in the population mean of the variable, i.e.*, $\mu_B > \mu_A$.
 - If 0.05 , we*cannot reject or accept the null hypothesis* $because the least likely scenario (<math>\mu_B < \mu_A$) has more than a 5% chance of being true, but less than a 50% chance of being true. The *results are therefore inconclusive we should conduct more tests*.
- For 99% confidence, substitute 0.01 for 0.05 in the above criteria.
- Excel has a built-in macro in <u>Data Analysis</u> that performs this type of hypothesis test automatically. It is called **t-Test: Paired Two Sample for Means**.
- The procedure is best illustrated by example, which we will do in class.

Two Sample Hypothesis Testing when *n* is *not the same* for the two Samples <u>Two-tailed un-paired samples hypothesis test</u>:

- Now consider the more general case in which the number of data points n_A in sample A is *not* the same as the number of data points n_B in sample B (e.g., $n_A = 10$ and $n_B = 15$).
- The analysis is similar to the above simpler case, except we need to combine the two samples in some appropriate manner to calculate the *t*-statistic.
- Consider the following general case:
 - **Sample** *A*: Number of data points = n_A , sample mean = \overline{x}_A , and sample standard deviation = S_A .
 - **Sample** *B*: Number of data points = n_B , sample mean = \overline{x}_B , and sample standard deviation = S_B .
 - Our goal is to predict whether there is a statistically significant difference between μ_A (the population mean of sample *A*) and μ_B (the population mean of sample *B*).
- Statisticians refer to this kind of hypothesis test as *hypothesis testing of two independent samples*.
- As usual, we set the null hypothesis and alternative hypothesis:
 - *Null hypothesis*: There is *no difference* between the population means, i.e., $\mu_A = \mu_B$. [This is the least likely scenario since $\overline{x}_A \neq \overline{x}_B$.] [This is a two-tailed hypothesis test.]
 - *Alternative hypothesis*: $\mu_A \neq \mu_B$. In other words, either $\mu_A > \mu_B$ or $\mu_A < \mu_B$. [This is the most likely scenario since $\overline{x}_A \neq \overline{x}_B$.]
- The critical *t*-statistic is formed using a *root sum of the squares* approach, similar to the way we handled

multiple uncertainties previously using RSS uncertainty analysis, namely, $t = \frac{1}{\sqrt{S}}$

- ly, $t = \frac{X_A X_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$.
- The corresponding *p*-value is calculated as previously, based on the critical *t*-statistic. In this case we are considering a *two-tail hypothesis test*. *p* is calculated in Excel using **TDIST(ABS(t),df,2)**, where df is the number of degrees of freedom, and the "2" specifies two tails. Use the tables if Excel is not available.
- But what should we use as the value of df? There are several options, and statisticians seem to disagree on which is best:

- The simplest option is to use the *average* df, $df = AVERAGE((n_A 1), (n_B 1))$
- Another option is to use $df = MIN((n_A 1), (n_B 1))$. In other words, we set df to the *smaller* of the two degrees of freedom calculated for the two independent samples.
- Another option is to use $df = MAX((n_A 1), (n_B 1))$. In other words, we set df to the *larger* of the two degrees of freedom calculated for the two independent samples.
- The "best" and most popular option is *Welch's equation*, df = NINT



where NINT(*x*) returns the integer *nearest* to real variable \overline{x} .

- With Welch's equation, df is calculated based on a *weighted average* of the two samples.
- *Note*: It is possible for Welch's equation to yield a value of df that is *larger* than either df_A or df_B .
- In Excel, the integer nearest to real number x is calculated using a built-in function, ROUND(x,0).
 So here, we would use ROUND(df,0) to obtain the nearest integer value of df.
- There are other equations for calculating df, but these are not listed here.
- As previously, we formulate our conclusions (to 95% confidence level) based on the *p*-value:
 - If p < 0.05, we reject the null hypothesis because the least likely scenario ($\mu_A = \mu_B$) has less than a 5% chance of being true. Thus, we can state confidently that *there is a statistically significant change in the population mean of the variable, i.e.*, $\mu_A \neq \mu_B$.
 - If 0.05 , we*cannot reject or accept the null hypothesis* $because the least likely scenario (<math>\mu_A = \mu_B$) has more than a 5% chance of being true, but less than a 95% chance of being true. The *results are therefore inconclusive we should conduct more tests*.
 - If p > 0.95, we *accept the null hypothesis* because what we set as the least likely scenario ($\mu_A = \mu_B$) turns out to have more than a 95% chance of being true. Thus, we can state confidently that *there is no statistically significant change in the population mean of the variable, i.e.,* $\mu_A = \mu_B$.

One-tailed un-paired samples hypothesis test: [This is the more common one used in engineering analysis.]

- We assume here that our experiments yield $\overline{x}_B > \overline{x}_A$. In other words, the modification we made leads to an *improvement* in the mean between Sample A and Sample B. But is the improvement statistically significant?
- For this one-tailed hypothesis test, we set the null hypothesis and alternative hypothesis:
 - *Null hypothesis*: There is *a decrease in* the population means, i.e., $\mu_B < \mu_A$. [This is the least likely scenario since $\overline{x}_B > \overline{x}_A$.] [This is a one-tailed hypothesis test.]
 - Alternative hypothesis: $\mu_B > \mu_A$. [This is the most likely scenario since $\overline{x}_B > \overline{x}_A$.]
- The critical *t*-statistic is calculated exactly as above for the two-tailed test.
- The corresponding *p*-value is calculated based on the critical *t*-statistic. In this case we are considering a *one-tail hypothesis test*. So, *p* is calculated in Excel using the function **TDIST**(**ABS**(*t*),**df**,**1**), where the "1" specifies one tail. Use the tables if Excel is not available.
- As discussed previously, for a one-tailed hypothesis test in which the null hypothesis is set to the *least likely scenario*, the *p*-value is limited in range from 0 to 0.5 (0% to 50%). we therefore formulate our conclusions (to 95% confidence level) based on the *p*-value:
 - If p < 0.05, we reject the null hypothesis because the least likely scenario ($\mu_B < \mu_A$) has less than a 5% chance of being true. Thus, we can state confidently that *there is a statistically significant increase in the population mean of the variable, i.e.*, $\mu_B > \mu_A$.
 - If 0.05 , we*cannot reject or accept the null hypothesis* $because the least likely scenario (<math>\mu_B < \mu_A$) has more than a 5% chance of being true, but less than a 50% chance of being true. The *results are therefore inconclusive we should conduct more tests*.
- For 99% confidence, substitute 0.01 for 0.05 in the above criteria.
- Excel has a macro, **t-Test: Two-Sample Assuming Unequal Variances**, to conduct this type of hypothesis test, as will be demonstrated in class. *Note*: Excel uses Welch's equation to calculate df in this macro.
- Again, the procedure is best illustrated by example, which we will do in class.