King Fahd University of Petroleum & Minerals

DEPARTMENT OF MATHEMATICAL SCIENCES

Technical Report Series

TR 364

December 2006

On the Two-sample t-Test and Simple Linear Regression

M.H. Omar & B. Yushau

DHAHRAN 31261 ● SAUDI ARABIA ● www.kfupm.edu.sa/math/ ● E-mail: mathdept@kfupm.edu.sa

On the Two-sample t-Test and Simple Linear Regression

By
M.H. Omar & B. Yushau
Department of Mathematical Sciences
King Fahd University of Petroleum & Minerals
E-mails: omarmh@kfupm.edu.sa, byushau@kfupm.edu.sa.

Abstract

Statistical inferences on means of two independent groups are often taught in introductory statistics courses, so is simple linear regression. However, it is not obvious that these two seemingly different procedures have something in common. In this note we highlight some common grounds between these procedures with the hope that it will enhance the teaching of these two topics as well as consolidate students understanding of these concepts.

1. Introduction

In the traditional approach of teaching Introductory Statistics course, students will be exposed to statistical inference about two population means before learning simple linear regression. These two topics often appear disjointed for beginning students. In addition, it seems no write-up exists at the introductory level, to the best of our knowledge, for instructors who want to put some extra effort to highlight the similarities and differences between these two concepts. Therefore, students often understand them to be two different concepts with nothing in common. While advanced courses such as regression analysis or design of experiments may allude to some common grounds, these courses are either never taken by non-statistics majors or are taken at a later stage of the statistics curriculum.

In this note, we present both topics with the aim of underlining their similarities so as to provide instructors with an insight to bridge the pedagogical gap during instruction. One potential advantage of drawing attention to this is that students may have more unified understanding of statistical procedures and invariably deeper comprehension of the two concepts.

2. Two Independent samples t-test

The independent sample t-test is often used whenever there is a test for equivalence of two independent population means. This topic is usually taught right after single sample t-test and before simple linear regression as is commonly organized as such in most introductory statistics textbooks (e.g. Groebner, Shannon, Fry, & Smith (2005).

In case the population variances are not known, the general form of this test statistic is

$$t = \frac{\bar{Y}_B - \bar{Y}_A - \left(\mu_B - \mu_A\right)}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}} \qquad (2.1)$$

with degree of freedom, $df = \dfrac{\left(\left(\frac{s_A^2}{n_A}\right) + \left(\frac{s_B^2}{n_B}\right)\right)^2}{\frac{\left(\frac{s_A^2}{n_A}\right)^2}{n_A - 1} + \frac{\left(\frac{s_B^2}{n_B}\right)^2}{n_B - 1}}$.

The following assumptions are usually made and are crucial for this test statistic to take the form of a *t*-distribution with the said degrees of freedom.

(1) the Y_i are samples from two independent groups

(2) the Y_i are independently and identically distributed (iid) from a normal distribution with mean μ_{y_i} and $\sigma_{y_i}^2$, and

(3) the variance of Y_i for each group ($\sigma_{y_A}^2$ and $\sigma_{y_B}^2$) is unknown.

There are several special cases for the test statistic. For instance, when the population variances are known and therefore are used instead of the sample variances in (2.1), then the test statistic follows a standard normal distribution.

Another special case is when the two variances are unknown but can be assumed to be equal to each other. In this case, the test statistic is

$$t = \frac{\bar{Y}_B - \bar{Y}_A - \left(\mu_B - \mu_A\right)}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \qquad (2.2)$$

where $s_p = \sqrt{\dfrac{\left(n_A - 1\right)s_1^2 + \left(n_B - 1\right)s_2^2}{n_A + n_B - 2}}$, with degree of freedom $df = n_A + n_B - 2$. In addition to the three assumptions mentioned earlier, this case requires the assumption that although the population variances are unknown, they can be assumed to be equal.

The third special case is when the sample sizes are large. Under the central limit theorem, if the sample size for each group is large enough (for example when $n \geq 30$), the test statistic in (2.1) and (2.2) can be approximated by the standard normal distribution.

3. Simple Linear Regression.

Simple linear regression is used when there are two variables where one is used to predict values of the other. The model for this procedure is often expressed as:

$$Y = B_0 + B_1 X + e \qquad (3.1)$$

where $B_0 =$ intercept, $B_1 =$ slope of model, and e = error.

With this model, the variation in *Y* scores can be broken up into two parts as follows

$$SS_{total} = SS_{regression} + SS_{error}$$

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}_i\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \qquad (3.2)$$

where $df_{total} = n-1$, $df_{regression} = 1$, and $df_{error} = n-2$.

A measure of overall variation in the model is given by

$$MS_{error} = \frac{SS_{error}}{n-2} \qquad (3.3)$$

and a measure of overall model-data fit is given by

$$R^2 = \frac{SS_{regression}}{SS_{total}} \qquad (3.4).$$

To test the significance of this sample linear regression model, it suffices to test for the non-zero slope of the regression line. In this case, the following test statistic is used;

$$t = \frac{B_1 - \beta_{1_0}}{s_{B_1}} = \frac{B_1 - 0}{\sqrt{MS_{error}}/\sqrt{S_{xx}}} \qquad (3.5)$$

with degrees of freedom, $df = n-2$.

This test statistic takes on the form of a *t*-distribution with this degree of freedom when all assumptions of the simple linear regression model are met. These crucial assumptions are:

(1) The error terms are independent

(2) The error terms are from a normal distribution with mean 0 and variance σ_e^2

(3) The variance of the error terms is constant at all levels of X

(4) The means of Y given $X, E(Y \mid X)$, form a straight line.

A special case occur when the independent variable X is the group identification dummy variable. In this case, the simple linear regression is often referred to as a simple linear regression with a dummy variable.

4. The two-sample t-test from the Perspective of Simple Linear Regression.

In this section, we intend to examine the two-independent samples pooled t-test in light of the simple linear regression. The next theorem shows equivalence of the two procedures under certain conditions.

We first recall the following well-known quantities in introductory statistics.

Lemma 4.1

(i) $\bar{Y}_A = \dfrac{\sum_{i=1}^{n_A} Y_{A_i}}{n_A}$

(ii) $\bar{Y}_B = \dfrac{\sum_{i=1}^{n_B} Y_{B_i}}{n_B}$

(iii) $Sxx' = \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2$

Theorem 4.1. The two-independent samples pooled t-test is exactly simple linear regression with one dummy variable provided that the four conditions in (2.2) are met.

Proof. First we observe that the test statistic in equation (2.2), can be rephrased into a more general form by treating the groups as another variable, say *X*.

$$\text{Let } X = \begin{cases} 0 & \text{when group = A} \\ 1 & \text{when group = B.} \end{cases}$$

Assume that *X* is the independent variable in predicting the dependent variable *Y* in a simple linear regression context. Then by assumption and using Lemma 4.1, we can write

$$\bar{X} = \frac{1}{n_A + n_B}\left(\sum_{i=1}^{n_A} X_{Ai} + \sum_{i=1}^{n_B} X_{Bi}\right) \text{ and}$$

$$Sxx' = \sum_{i=1}^{n_A} X_{Ai}^2 + \sum_{i=1}^{n_B} X_{Bi}^2 - \frac{1}{n_A + n_B}\left(\sum_{i=1}^{n_A} X_{Ai} + \sum_{i=1}^{n_B} X_{Bi}\right)^2 .$$

Since for all *i*, $x_{Ai} = 0$ and $x_{Bi} = 1$, then we have

$$\bar{X} = \frac{n_B}{n_A + n_B} \quad (4.1)$$

$$Sxx' = n_B - \frac{n_B^2}{n_A + n_B} = \frac{n_A n_B}{n_A + n_B} \quad (4.2).$$

Similarly, the pooled standard deviation in equation (2.2) can be rewritten as follows;

$$s_p = \sqrt{\frac{(n_A - 1)s_1^2 + (n_B - 1)s_2^2}{n_A + n_B - 2}}$$

$$= \sqrt{\frac{(n_A - 1)\sum_{i=1}^{n_A} \frac{\left(Y_i - \bar{Y}_A\right)^2}{(n_A - 1)} + (n_B - 1)\sum_{i=1}^{n_B} \frac{\left(Y_i - \bar{Y}_B\right)^2}{(n_B - 1)}}{n_A + n_B - 2}}$$

$$= \sqrt{\frac{\sum_{i=1}^{n_A}\left(Y_i - \bar{Y}_A\right)^2 + \sum_{i=1}^{n_B}\left(Y_i - \bar{Y}_B\right)^2}{n_A + n_B - 2}} .$$

But because each group mean is the predicted value (or the conditional mean) at each *X* value, we can rewrite the above right hand-side of the pooled standard deviation as follows;

$$s_p = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2}{n_A + n_B - 2}}$$

$$= \sqrt{\frac{SS_{error}}{n-2}}$$

$$= \sqrt{MS_{error}} \qquad (4.3).$$

Notice that, when testing the null hypothesis of zero true mean difference, if we divide and multiply the numerator of equation 2.2 by $\frac{n_B n_A}{n_A + n_B}$, the sample mean differences is

$$\bar{Y}_B - \bar{Y}_A = \frac{\frac{n_B n_A}{n_A + n_B}\left(\bar{Y}_B - \bar{Y}_A\right)}{\frac{n_B n_A}{n_A + n_B}} = \frac{\frac{n_B n_A \bar{Y}_B - n_B n_A \bar{Y}_A}{n_A + n_B}}{\frac{n_B n_A}{n_A + n_B}}.$$

Using Lemma 4.1(iii) and rearranging terms, we have

$$\bar{Y}_B - \bar{Y}_A = \frac{-\frac{n_B}{n_A + n_B}\sum_{i=1}^{n_A} Y_{A_i} + \frac{n_A}{n_A + n_B}\sum_{i=1}^{n_B} Y_{B_i}}{S_{xx'}}$$

$$= \frac{\left(0 - \frac{n_B}{n_A + n_B}\right)\sum_{i=1}^{n_A} Y_{A_i} + \left(1 - \frac{n_B}{n_A + n_B}\right)\sum_{i=1}^{n_B} Y_{B_i}}{S_{xx'}}$$

$$= \frac{\sum_{g=A}^{B}\sum_{i=1}^{n_g}\left(X_{g_i} - \frac{n_B}{n_A + n_B}\right)Y_{g_i}}{S_{xx'}} \text{ (by 4.1 and definition of grouping variable)}$$

$$= \frac{\sum_{g=A}^{B}\sum_{i=1}^{n_g}\left(X_{g_i} - \bar{X}\right)\left(Y_{g_i} - \bar{Y}\right)}{S_{xx'}}$$

which can further be rewritten as follows;

$$\bar{Y}_B - \bar{Y}_A = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{S_{xx'}} = B_1 \qquad (4.4).$$

Equation 4.4 is the least square estimate of the slope of the simple linear regression model. Now, putting equations 4.2 through 4.4 together in equation 2.2 we have that

$$t = \frac{b_1 - 0}{\sqrt{MS_{error}}/\sqrt{S_{xx}}}$$

which is exactly the equation 3.5. Thus, this proves the theorem.

An alternative proof can also be taken. Namely, by noting that the independent sample t-test is a special case of Analysis of Variance (ANOVA) and drawing its equivalence to linear regression with dummy coding. However, this line of proof cannot be presented to beginning students as they would not have been introduced to such material at the introductory stage.

On the other hand, it is clear that the first three assumptions of the simple linear regression are the same as those of the pooled 2-sample t-test. Also, note that the last assumption that the means of Y given $X, E(Y \mid X)$, form a straight line is also met since there are only two independent groups forming the X independent dummy variable which ensures that the conditional group means will always form a straight line.

In particular, the following results provide detailed equivalence between these two methods.

Lemma 4.2

(i) $\bar{Y}_B - \bar{Y}_A = B_1$

(ii) $s_p = \sqrt{MS_{Error}}$

(iii) $n_A + n_B - 2 = n - 2$

(iv) $\dfrac{\bar{Y}_A - \bar{Y}_B - 0}{s_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} = \dfrac{B_1 - 0}{\sqrt{MSE}/\sqrt{S_{xx}}}$

(v) $\bar{Y}_A = B_o$

(vi) $1 - \dfrac{(n_A + n_B - 2)s_p^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = R^2.$

The proof of parts (i) through (iv) is clearly given in the previous proof, for part (v), note that $\bar{Y}_A = \dfrac{n_A\ \bar{Y}_A + n_B\ \bar{Y}_B}{n_A + n_B} - \left(\bar{Y}_B - \bar{Y}_A\right)\dfrac{n_B}{n_A + n_B} = \bar{Y} - B_1\bar{X} = B_{o.}$

Similarly, for part (vi), we have

$$1 - \frac{(n_A + n_B - 2)s_p^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = 1 - \frac{\left((n_A - 1)s_1^2 + (n_B - 1)s_2^2\right)}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = 1 - \frac{SS_{Error}}{SS_{total}} = R^2.$$

Because the equivalence of the sampling distribution of the left and right handside of Lemma 4.2 (iv), the *P*-value for significance of the regression slope is also the same as the *P*-value for the significant mean differences in the 2-independent samples pooled t-test.

5. Example.

Here we illustrate the two procedures by example, and show how their equivalence. For this, consider the following data generated from two independent normal distributions.

Experiment	Control
60.6798	62.2146
47.8421	59.1447
55.9217	33.1350
51.3782	21.0975
39.7180	55.7910
61.7317	38.9156
68.7972	50.7410
65.0855	23.2715
41.0055	50.2861
59.3898	33.2222
64.4013	48.7167
59.6482	20.3333
37.9622	60.4277
68.6570	73.0460
61.7293	

Analyses were conducted with the MINITAB statistical software. Note that the population variances can be assumed equal for this data since the *P*-value for testing this equivariance assumption is 0.081. The analyses results for both the 2-independent samples pooled t-test procedure and the simple linear regression procedure on this data are shown in Table 1. In addition, equivalence between the two procedures previously discussed is highlighted in the accompanying footnotes.

Although there appear to be some differences in the results highlighted above, these are only dissimilar due to some rounding errors. This routinely happens as statistical software packages such as MINITAB currently do not strive to uphold the same level of rounding between procedures.

It is also somewhat important to note that depending on which group is treated as the control group, the regression intercept and slope may produce results in the opposite sign from the t-test procedure. This is not a real difference since it can easily be addressed by adopting the same consistency throughout analyses.

```
Two-Sample T-Test and CI: Experiment, Control

Two-sample T for Experiment vs Control

            N      Mean     StDev   SE Mean
Experime   15      56.3      10.3       2.7
Control    14      45.0[a]   16.8       4.5

Difference = mu Experiment - mu Control
Estimate for difference:  11.24[b]
95% CI for difference: (0.69, 21.79)
T-Test of difference = 0 (vs not =): T-Value = 2.19[d]  P-Value = 0.038[e]  DF = 27[c]
Both use Pooled StDev = 13.8[c]


Regression Analysis: Y versus X(grp)

The regression equation is
Y = 45.0 + 11.2 X(grp)

Predictor         Coef     SE Coef        T       P
Constant        45.025[a]    3.699    12.17   0.000
X(grp)          11.239[b]    5.143     2.19[d] 0.038[e]

S = 13.84[f]      R-Sq = 15.0%     R-Sq(adj) = 11.9%

Analysis of Variance

Source            DF         SS        MS        F       P
Regression         1      914.6     914.6     4.77   0.038[e]
Residual Error    27[c]  5172.2     191.6
Total             28     6086.9

Unusual Observations
Obs     X(grp)         Y        Fit      SE Fit     Residual     St Resid
 29       0.00     73.05      45.02        3.70        28.02        2.10R

R denotes an observation with a large standardized residual
```

Notes:
a Control group mean = regression intercept
b Group mean difference (Experimental – Control) = regression slope
c pooled t-test degree of freedom = error degree of freedom in regression
d t statistic for mean difference = t statistic for nonzero significant slope
e p-value for mean diff t-statistic = p-value for nonzero slope t-test
f pooled standard deviation = regression standard error of estimate

Figure 1. Illustration of the equivalence between the pooled t-test and the simple linear regression procedure.

6. Concluding remark.

In this note, we have shown the common grounds between the pooled t-test procedure and the simple linear regression with a dummy variable. Furthermore in the proof, we have utilized only usual material covered in an introductory statistics course. Although other alternative proofs such as which shows equivalence between ANOVA (where independent samples t-test is a special case) and linear regression with dummy variable can be used, this will require more advanced curriculum than necessary to establish the equivalence. It is hoped that the highlighted equivalence between procedures will help provide insights into a more unified approach to teaching statistics concepts in introductory statistics courses at the undergraduate level.

Acknowledgements

The authors wish to acknowledge the excellent research services of King Fahd University of Petroleum & Minerals

Reference

D. F. Groebner, P.W. Shannon, P.C. Fry, & K.D. Smith (2005). *Business Statistics: A Decision-Making Approach* (6th ed). Pearson Prentice-Hall: New Jersey.

MINITAB Inc. (2000). MINITAB Statistical Software Release 13.20.