**Far East Journal of Mathematical Education**

Volume 11, Number 2, 2013, Pages 205-216

# TEACHING POWER OF THE *t*-TEST AND ITS INTERPRETATION

**David Sotres-Ramos**[1] **and Félix Almendra-Arao**[2]

[1]Colegio de Postgraduados
México

[2]UPIITA del Instituto Politécnico Nacional
México

## Abstract

An experiment is used for teaching the fundamental statistical concept of power of *t*-test. From the results of the experiment, a simple estimator of power of *t*-test is obtained. Within the context of an introductory statistics course, this teaching material gives the students a clear interpretation of this concept.

## 1. Introduction

The last decades have witnessed a progressively increasing use of statistical reasoning in all fields of science and even in humanities, see for instance, [1-5]. Consequently, it is now essential for students in most academic and professional programs to be acquainted with the basic principles and techniques of statistical analysis. This alone justifies research into teaching statistics which can develop appropriate teaching materials. In this article, a simple experiment is used to illustrate the basic ideas of hypothesis testing. This experiment has the advantage that the essential statistical concept of power of *t*-test can be clearly understood. We have

---

Received: August 20, 2013; Accepted: September 21, 2013

Keywords and phrases: hypothesis testing, *t*-test, level of significance.

been used this experiment, with very good results, in introductory statistics courses for students in diverse areas of science.

## 2. Hypothesis Testing in Clinical Research

Suppose that an investigator believes that drug *D* may be effective for reducing the blood cholesterol levels of hypercholesterolemic patients. He/she wants to run a clinical trial to contrast the following hypotheses: $H_0 : \mu_D = \mu_P$ vs $H_a : \mu_D < \mu_P$, where $\mu_D$ is the average cholesterol level for patients taking the drug and $\mu_P$ is the average cholesterol level for patients taking placebo. Figure 1 gives a graphical representation of the alternative hypothesis $H_a : \mu_D < \mu_P$. In this situation, the mean cholesterol levels for the placebo ($P$) and drug ($D$) populations are not the same, that is, $\mu_P = 260\,\mathrm{mg/dl}$ and $\mu_D = 200\,\mathrm{mg/dl}$. Thus, the drug has a positive average effect equals $\mu_P - \mu_D = 60\,\mathrm{mg/dl}$. However, in practice, we always deal with a sample of the population and not the population itself. We never know these population distributions, or the population means $(\mu_P$ and $\mu_D)$ and hence, the objective is to estimate the population parameter $\mu_P - \mu_D$ (drug cholesterol average reduction) based on our sample results.

### *t*-test procedure

The following procedure, known as the *t*-test procedure, is the recommended statistical test for comparing the means of two normal distributions with equal variances and equal sample size in both groups, see for instance, Wayne [4]:

If $t_c > t(2n - 2, \alpha)$, then reject the null hypothesis $(H_0 : \mu_D = \mu_P)$, and otherwise do not reject the null hypothesis, where

$$t_c = (\overline{X}_P - \overline{X}_D)/S_e, \tag{1}$$

$t(2n - 2, \alpha)$ is the $\alpha$ upper quantile of the *t* distribution with $2n - 2$ degrees of freedom, and *n* is the sample size of each of the two samples.

$S_e = \{s^2(2/n)\}^{1/2}$ is an estimate of the standard error of the difference of means,

$s^2 = \{(n-1)s_P^2 + (n-1)s_D^2\}/(2n-2)$ is an estimate of the variance of the means' difference,

$\overline{X}_P,\ \overline{X}_D$ are the sample means, $s_P^2,\ s_D^2$ are the sample variances, and $\alpha$ is the significance level of the test. The set of values of $t_c$ that satisfies the relation $t_c > t(2n-2, \alpha)$, is called the *critical region* and $t_c$ is called the *test statistic*.

**Power of the *t*-test procedure**

The reliability of statistical tests is measured by calculating the probability of two distinct types of error: Type I error = reject the null hypothesis when it is true, and type II error = do not reject the null hypothesis when it is false. In the terminology of hypothesis testing, and using the symbols α and β for the two error probabilities:

$$P\text{ (Type I error)} = \alpha = \text{level of significance;}$$

$$P\text{ (Type II error)} = \beta = 1 - \text{power.}$$

Some authors call a type I error an α *error* and a type II error a β *error*. Notice that power $= 1 - \beta$.

Of course, we would like both α and β to be as small as possible. We might reduce the size of α by including fewer values in the critical region, but then the value of β increases. It is true, in general, that for a fixed sample size, α and β vary inversely. In practice, it is common to choose values of α and power equal to 0.05 and 0.90, respectively. Once α and power have been fixed, it is possible to choose a value of *n*, the sample size, that will guarantee at least this degree of control over these two error probabilities.

It can be proved that an approximation of the power of the *t*-test is given by the following formula:

$$\text{Power}(\sigma, \delta, n) = \Phi[(\delta/\sigma) \times (n^{1/2}/2^{1/2}) - t(2n-2, \alpha)], \qquad (2)$$

where $\Phi$ is the cumulative distribution function of the standard normal random variable, $\sigma$ is the population standard deviation, $\delta = \mu_P - \mu_D$, $n$ is the sample size of each of the two samples and $t(2n-2, \alpha)$ is the $\alpha$ upper quantile of the $t$ distribution with $2n-2$ degrees of freedom, see for instance, Mood et al. [6]. Applying the well known formula for the standard deviation, we obtain that the standard deviation of both distributions in Figure 1 is equal to $\sigma = 58.2\,\text{mg/dl}$, and clearly $\delta = 60\,\text{mg/dl}$. Using formula in (2), with $\alpha = 0.05$, power of $t$-test is calculated for different values of $n$, $\delta$ and $\sigma$, see Table 1.

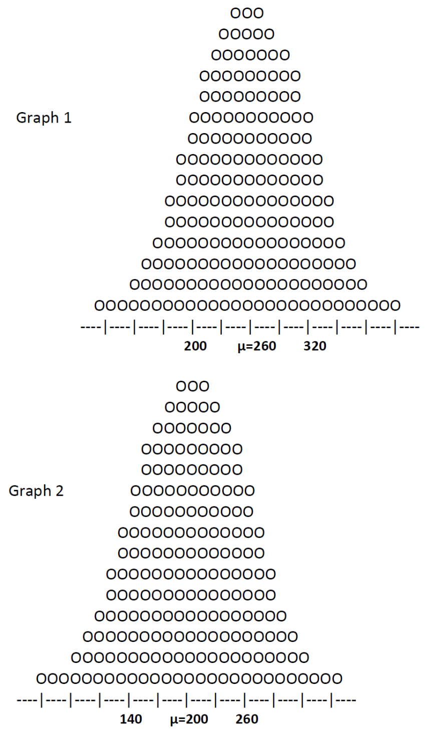

**Figure 1.** Graph 1 shows the distribution of cholesterol level in a population of 200 patients with moderate hypercholesterolemia after 4 weeks of taking placebo that produces an average cholesterol level of 260mg/dl. Graph 2 shows the distribution of cholesterol level in a population of 200 patients after 4 weeks of taking an effective drug that reduces the average cholesterol level to 200mg/dl.

**Table 1.** Power of *t*-test for different values of $n$, δ and σ, using formula in (2)

| $n$ | δ | σ | Power |
|---|---|---|---|
| 8 | 60 | 58.2 | 61.8% |
| 10 | 60 | 58.2 | 71.6% |
| 15 | 60 | 58.2 | 86.9% |

The mathematical derivation of the formula for the power of *t*-test in (2) is out of the scope of this paper. However, the experiment described in Section 3 provides a clear and simple method for the calculation and interpretation of the power of *t*-test.

## 3. The Experiment

In this paper, we present an experiment for explaining the interpretation of the power of the *t*-test defined above. The interpretation of $\alpha = P$ (Type I error) will not be discussed in this paper. A detailed explanation of the interpretation of α can be found in [7]. Thus, in the rest of this paper, we will suppose that the true situation is the one represented in Figure 1 in which the treatment has a positive effect $\delta = \mu_P - \mu_D = 60\,\mathrm{mg/dl}$, that is, the alternative hypothesis $H_a : \mu_D < \mu_P$ is true. However, we will proceed as if we do not know the real situation, and we will simulate the realizations of several trials obtaining the sample data and will calculate the corresponding *t*-test for every trial, to get an interpretation of the power of the *t*-test. The experiment is divided into four steps.

### Step # 1. Preparing the tools for the generation of the sample data

The students are asked to do the following activities: Fill urn number one (the placebo urn) with 200 marbles. This placebo urn will represent the cholesterol values for the 200 hypercholesterolemic patients of Graph 1 in Figure 1. Each marble will have written on it a cholesterol value in such a way that 16 marbles will have the value 260, other 16 marbles will have the value 272, 15 with the value 284, etc. Table 2 shows the list of cholesterol values that should be included in the placebo urn and the corresponding frequency of each value.

**Table 2.** List of cholesterol values and its frequency that should be included in the placebo urn

| Choles | 104 | 116 | 128 | 140 | 152 | 164 | 176 | 188 | 200 | 212 | 224 | 236 | 248 | 260 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Freq | 1 | 1 | 1 | 2 | 3 | 4 | 6 | 8 | 10 | 12 | 13 | 15 | 16 | 16 |
| Choles | 272 | 284 | 296 | 308 | 320 | 332 | 344 | 356 | 368 | 380 | 392 | 404 | 416 | Total |
| Freq | 16 | 15 | 13 | 12 | 10 | 8 | 6 | 4 | 3 | 2 | 1 | 1 | 1 | 200 |

Fill urn number two (the drug urn) with 200 marbles. This drug urn will represent the cholesterol values for the 200 hypercholesterolemic patients of Graph 2 in Figure 1. Each marble will have written on it a cholesterol value in such a way that 16 marbles will have the value 200, other 16 marbles will have the value 212, 15 with the value 224, etc. Table 3 shows the list of cholesterol values that should be included in the drug urn and the corresponding frequency of each value.

**Table 3.** List of cholesterol values and its frequency that should be included in the drug urn

| Choles | 44 | 56 | 68 | 80 | 92 | 104 | 116 | 128 | 140 | 152 | 164 | 176 | 188 | 200 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Freq | 1 | 1 | 1 | 2 | 3 | 4 | 6 | 8 | 10 | 12 | 13 | 15 | 16 | 16 |
| Choles | 212 | 224 | 236 | 248 | 260 | 272 | 284 | 296 | 308 | 320 | 332 | 344 | 356 | Total |
| Freq | 16 | 15 | 13 | 12 | 10 | 8 | 6 | 4 | 3 | 2 | 1 | 1 | 1 | 200 |

### Step # 2. Using placebo and drug urns, extract 3 pairs of random samples simulating 3 different clinical trials

A random sample of the cholesterol distribution for the placebo population (Table 2) consists of 8 extractions from the placebo urn. Every time a student extracts a marble he/she writes down the obtained number and returns it to the urn and mixes it in. Similarly, a random sample from the drug population (Table 3) is obtained. These two random samples simulate the first clinical trial. The students are asked to repeat this procedure twice to simulate the second and third clinical trials. The results obtained from these three simulated clinical trials are presented in Table 4.

**Table 4.** Three clinical trials obtained from cholesterol urn and placebo urn

| Clinical Trial # 1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Placebo | 200 | 224 | 236 | 248 | 248 | 260 | 272 | 308 |
| Drug D | 152 | 152 | 164 | 188 | 188 | 188 | 200 | 260 |
| Clinical Trial # 2 | | | | | | | | |
| Placebo | 236 | 248 | 248 | 260 | 260 | 272 | 272 | 284 |
| Drug D | 176 | 176 | 188 | 200 | 212 | 212 | 224 | 236 |
| Clinical Trial # 3 | | | | | | | | |
| Placebo | 188 | 200 | 248 | 248 | 260 | 260 | 272 | 272 |
| Drug D | 200 | 200 | 212 | 224 | 224 | 236 | 248 | 260 |

### Step # 3. Calculating the *t*-test for the 3 pair of samples in Step # 2 and also use simulation

The students are asked to calculate the *t*-test for each of the 3 pair of samples obtained in Step # 2, using the *t*-test procedure in (1). The values of $t_c$ associated with clinical trials numbers 1, 2 and 3 in Table 4 are: 3.75, 5.97 and 1.32, respectively. Note that the data in clinical trial # 3 is associated with a *t* value of 1.32 which leads us to do not reject the null hypothesis, because $t_c = 1.32 < t(2n - 2, \alpha) = t(14, 0.05) = 1.761$. This is one of the key points of the experiment: even that the null hypothesis is false, then it is possible that the conclusion of the *t*-test is to do not reject this null hypothesis which is the wrong answer. If all we had were the data shown in clinical trial # 3 in Table 4, we would conclude that the observations were inconsistent with the hypothesis that the drug had a positive effect and report that it does not reduce the cholesterol level, our conclusion about the drug would be wrong. Therefore, for a given investigation with power $(\sigma = 58.17, \delta = 60, n = 8) = 61.8\%$, it means that if the treatment has a positive effect

$\delta = \mu_P - \mu_D = 60\,\mathrm{mg/dl}$ (that is the alternative hypothesis $H_a : \mu_D < \mu_P$ is true), then 61.8% of the trials will accept the alternative hypothesis and declare a significant effect. This is a clear and correct interpretation of the power of the *t*-test.

At this point, student's attention is called to the *convenience* of having software that is able to perform the same operations than those calculated in Steps # 2 and # 3 more rapidly. Specifically, we want the program to do the following:

1. Do $k$ pair of random samples of size $n$ each from the distributions as in Tables 2 and 3 which have means equal $\mu_P = 260\,\mathrm{mg/dl}$ and $\mu_D = 200\,\mathrm{mg/dl}$, respectively.
2. Calculate the *t*-test for each pair of samples generated above, and obtain the number of trials for which the *t*-test rejects the null hypothesis $(H_0 : \mu_D = \mu_P)$, and call this value *h*.
3. Calculate the ratio $h/k$ which is the percentage of trials that rejects the null hypothesis when this hypothesis is false.

We have written, in the R environment, the program POWER (see Appendix 1) which has these characteristics. Using this program, it is possible to generate, in few seconds, a large number of trials, even for large sample sizes.

### Step # 4. Estimation of power of *t*-test

After we run the program POWER, described in Step # 3 above with $k = 500$, the result obtained was $h = 337$, and thus $h/k = 0.674$, see Table 5 below. This percentage (67.4%) is an estimator of the power of *t*-test, that is, the percentage of trials that rejected the null hypothesis when it is actually false. We explain to the students that if *k* is the total number of trials and we increase the number of trials, then for any fixed value of *n*:

$h/k =$ (# of trials that rejects the null hypothesis)/$k \rightarrow$ power, when $k \rightarrow \infty$.

## 4. The Data Simulation Program

In this section, we use the program POWER presented in the appendix for different values of *n* and *k*. In Table 5, we present several runs of this program for different values of *n* and *k*, in all cases, the samples are drawn from the distributions as in Tables 2 and 3 which have means equal $\mu_P = 260\,\text{mg/dl}$ and $\mu_D = 200\,\text{mg/dl}$, respectively, and $\alpha = 0.05$. For each combination of *n* and *k*, the program gives an estimator of the power of *t*-test.

**Table 5.** Runs of the program POWER, for sampling from the distributions as in Tables 2 and 3 which have means equal $\mu_P = 260\,\text{mg/dl}$ and $\mu_D = 200\,\text{mg/dl}$, respectively, and $\alpha = 0.05$, for different sample sizes (*n*), different number of trials (*k*), using the *t*-test for the results of each trial, and obtaining an estimator of the power of *t*-test, see Step # 4 above

| *n* | *k* | Power *t*-test | *n* | *k* | Power *t*-test | *n* | *k* | Power *t*-test |
|---|---|---|---|---|---|---|---|---|
| 8 | 100 | 0.610 | 25 | 100 | 0.970 | 50 | 100 | 1 |
| 8 | 200 | 0.620 | 25 | 200 | 0.975 | 50 | 200 | 1 |
| 8 | 300 | 0.653 | 25 | 300 | 0.973 | 50 | 300 | 1 |
| 8 | 500 | 0.674 | 25 | 500 | 0.972 | 50 | 500 | 1 |
| 8 | 2000 | 0.633 | 25 | 2000 | 0.981 | 50 | 2000 | 1 |

## 5. Conclusions

We have proposed an experiment to be used during class for teaching didactically the concept of power of *t*-test. From the results of the experiment, a simple estimator of the power of *t*-test is obtained, and also gives the students a clear interpretation of the power of *t*-test which is: “the percentage of trials that rejected the null hypothesis when it is actually false”. This material has been used in introductory statistics courses for students in diverse areas of knowledge. Feedback obtained from the students revealed that the material is an effective pedagogical tool for the significant understanding of this concept.

## Acknowledgement

The second author gratefully acknowledges the partial support of SNI-CONACYT, COFAA-IPN, EDI-IPN and Project SIP-IPN 20130948.

## Appendix 1

Program POWER for calculating an estimator of the power of *t*-test, for given values of $\alpha$, sample size $n_P$ (for placebo distribution), sample size $n_D$ (for drug distribution), where the samples are drawn from the distributions as in Tables 2 and 3 which have means equal $\mu_P = 260\,\mathrm{mg/dl}$ and $\mu_D = 200\,\mathrm{mg/dl}$, respectively, and doing $k$ trials for testing the null hypothesis $(H_0 : \mu_D = \mu_P)$ against the alternative $(H_a : \mu_D < \mu_P)$.

```
#
# Estimation of power for t test
#
# Variables
#
# alfa: level of significance
# n: sample size for placebo and drug population.
# k: number of experiments to do
#
POWER < –function (alfa, n, k)
{
placebo <–
c(104,  116,  128,  rep(140,  2),  rep(152,  3),  rep(164,  4),
rep(176, 6), rep(188, 8), rep(200, 10), rep(212, 12),
rep(224,  13),  rep(236,  15),  rep(248,  16),  rep(260,  16),
rep(272,  16),  rep(284,  15),  rep(296,  13),  rep(308,  12),
rep(320, 10), rep(332, 8), rep(344, 6), rep(356, 4),
rep(368, 3), rep(380, 2), 392, 404, 416)
```

```
drug <–
c(44, 56, 68, rep(80, 2), rep(92, 3), rep(104, 4), rep(116, 6),
rep(128, 8), rep(140, 10), rep(152, 12), rep(164, 13),
rep(176,  15),  rep(188,  16),  rep(200,  16),  rep(212,  16),
rep(224,  15),  rep(236,  13),  rep(248,  12),  rep(260,  10),
rep(272, 8), rep(284, 6), rep(296, 4), rep(308, 3),
rep(320, 2), 332, 344, 356)
h <– 0
t.crit <– –qt(alfa,2*n-2)
for (i in 1:k)
{
sample.p <– rep(0,n)
sample.d <– rep(0,n)
sampleu.p <– floor(runif(n, 1, 200))
sampleu.d <- floor(runif(n,1,200))
for (m in 1:n)
{
sample.p[m] <– placebo[sampleu.p[m]]
sample.d[m] <– drug[sampleu.d[m]]
}
mean.p <– mean(sample.p)
mean.d <– mean(sample.d)
var.p <– var(sample.p)
var.d <– var(sample.d)
var.sample <– ((n-1)*(var.p+var.d))/(2*n-2)
stand.error <– sqrt(var.sample*(2/n))
t.calc <– (mean.p-mean.d)/stand.error
if ( t.calc > t.crit) h <– h+1
}
print(h/k)
```

## References

[1] J. F. Healey, Statistics: A Tool for Social Research, 8th ed., Wadsworth Cengage Learning, 2008.

[2] D. Holmes, P. Moody and D. Dine, Research Methods for the Biosciences, 1st ed., Oxford University Press, 2006.

[3] M. J. Kiemele and S. R. Schmidt, Basic Statistics: Tool for Continuous Improvement, 4th ed., Air Academy Press, 1997.

[4] W. W. Wayne, Biostatistics: A Foundation for Analysis in the Health Sciences, 9th ed., John Wiley and Sons, Inc., 2010.

[5] D. Sotres-Ramos and F. Almendra-Arao, Using a simple experiment for teaching hypothesis testing, Far East J. Math. Edu. 6(2) (2011), 149-165.

[6] A. M. Mood, F. A. Graybill and D. C. Boes, Introduction to the Theory of Statistics, 3rd ed., McGraw-Hill Series in Probability and Statistics, 1974.

[7] F. Almendra-Arao and D. Sotres-Ramos, Some activities designed for teaching the type I error concept in a constructivist environment using simulation, Far East J. Math. Edu. 5(1) (2010), 1-15.