

Design and Analysis of Experiments

06 - Simple Comparisons

Version 2.11

Felipe Campelo
<http://orcslab.cpdee.ufmg.br/>

Graduate Program in Electrical Engineering

Belo Horizonte
March 2015

*“If you get a million dollars for research it
will be very helpful, of course;
but what can really make a difference are
the good ideas, recognizing those questions
which are really important,
which are yet to be answered. ”*

Suzana Herculano-Houzel
1972 –
Brazilian neuroscientist



Simple Comparative Experiments

Statistical inference for two samples

The concepts of comparison between two populations based on information obtained from their samples follow the same principles used for testing hypotheses about a single population;

Inferences for two samples frequently arise when comparing the effect of a technique (treatment) against a *control group*: placebo, classical technique, random search, etc;

Usual questions involve:

- Comparison of means;
- Comparison of variances;
- Comparison of proportions;
- etc.

Comparison of two means

Example: Length of steel rods



One of the critical aspects of manufacturing steel rods is cutting the bars with a precise length, which is expected by the customers.

This process is prone to errors, which result in additional costs for standardizing and reprocessing the rods.

An engineer is interested in comparing the current controller of the cutting scissors with a new method that could potentially improve the performance of the process.

Adapted from D.F. Carvalho Jr.'s course project for the Design and Analysis of Experiments Course, PPGEE-UFMG, June 2012. The data used in this example is not necessarily the original one.

Image: <http://www.shutterstock.com/pic-73207399/>

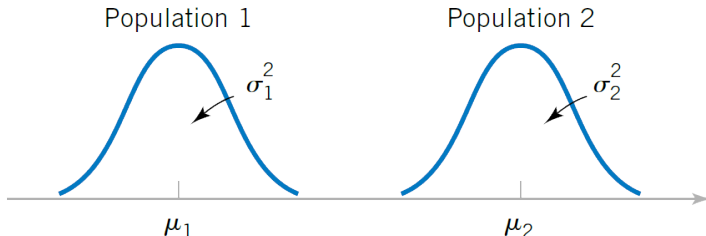
Comparison of two means

Example: Length of steel rods

A possible statistical model for this kind of data would be:

$$y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2 \\ j = 1, \dots, n_i \end{cases}$$

Lets initially assume that the residuals ϵ_{ij} are iid $\mathcal{N}(0, \sigma_i^2)$, which implies:



Comparison of two means

Definitions

What we wish is to perform an inference about the difference in the mean values of constructive deviations for the two controllers. In this case, a reasonable response variable would be the *absolute error*, from which we could test our hypotheses.

The statistical hypotheses can be stated as:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases} \quad \text{or, equivalently,} \quad \begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Suppose a desired significance level $\alpha = 0.05$, and that the engineer is interested in detecting any difference larger than $15mm$ in the mean absolute error with a power $(1 - \beta) = 0.8$.

Also, let's assume that the variance of the process is unknown but similar for both controllers.

Comparison of two means

Definitions

Since the variance is unknown, it will have to be estimated from the data. As we are assuming $\sigma_1^2 \approx \sigma_2^2$, we can use the pooled variance estimator:

$$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2} = w S_1^2 + (1 - w) S_2^2$$

Based on this estimator and the stated assumptions, we have that:

$$T = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1 + n_2 - 2)}$$

Comparison of two means

Rejection threshold

If we recall our working hypotheses:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

we have that, under H_0 :

$$t_0 = \frac{(\bar{y}_1 - \bar{y}_2) - \overset{0}{\cancel{(\mu_1 - \mu_2)}}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{y}_1 - \bar{y}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

We'll reject H_0 at the $(1 - \alpha)$ confidence level if $|t_0| \geq t_{1-\alpha/2; (n_1+n_2-2)}$

Comparison of two means

Sample sizes

Now recall that the process engineer was interested in some very specific characteristics for his test:

- Significance: $\alpha = 0.05$;
- Power: $(1 - \beta) = 0.8$;
- Minimally interesting effect: $\delta^* = 15mm$

From these specifications, we can obtain the required sample sizes. The derivation of the sample size formulas is not particularly difficult, but we'll concentrate only on the results. More details can be easily found in the literature^a.

^aCheck, for instance, Paul Mathews' *Sample Size Calculations*, MMB, 2010.

Comparison of two means

Sample sizes

For the general case of unequal sample sizes, we have:

$$n_1 = \left(1 + \frac{n_1}{n_2}\right) \left(\frac{s_p}{\delta^*}\right)^2 (t_{\alpha/2} + t_{\beta})^2$$

where s_p is the estimated common standard deviation, and $t_{\alpha/2}$ and t_{β} are the $\alpha/2$ and β quantiles of the $t_{(n_1+n_2-2)}$ distribution. The sample size n_2 can be calculated by simply substituting (n_1/n_2) by (n_2/n_1) .

For equal sample sizes ($n_1 = n_2 = n$) the expression is simplified to:

$$n = 2 \left(\frac{s_p}{\delta^*}\right)^2 (t_{\alpha/2} + t_{\beta})^2$$

Comparison of two means

Sample sizes

These formulas are very convenient, but leave us with a riddle: we need variance estimate in order to calculate the sample size, but we need observations to be able to estimate the variance.

There are a few ways to proceed in this case. The most practical are:

- Use process knowledge or historical data to obtain an (initial) estimate of the variance;
- Perform a pilot study and collect samples to estimate the variance.

The first method is almost always preferable since it does not imply additional costs for the experiment.

Comparison of two means

Sample sizes

If no information is available to estimate the variance, a pilot study must be performed to obtain this value. The sample size required for this pilot study is given by:

$$n_{pilot} \approx 2 \left(\frac{z_{\alpha_n/2}}{e_n} \right)^2$$

where $(1 - \alpha_n)$ is the desired confidence level for the sample size estimate of the main study, and e_n is the maximum relative error allowed for the sample size.

This calculation can yield some scarily large sample sizes for a pilot study (much larger than would be actually required for the main study itself), so use this with caution.

Comparison of two means

Sample sizes

For the steel rods experiment, suppose that the engineer uses data available from the controller manuals, as well as historical measurements, to estimate the common standard deviation for the cutting process as $\hat{\sigma} \cong 15mm$.

Assuming that equal sample sizes are desired, we can simply use the formula:

$$n = 2 \left(\frac{\hat{\sigma}}{\delta^*} \right)^2 (t_{\alpha/2} + t_{\beta})^2$$

Easy, right?

Comparison of two means

Sample sizes

For the steel rods experiment, suppose that the engineer uses data available from the controller manuals, as well as historical measurements, to estimate the common standard deviation for the cutting process as $\hat{\sigma} \cong 15mm$.

Assuming that equal sample sizes are desired, we can simply use the formula:

$$n = 2 \left(\frac{\hat{\sigma}}{\delta^*} \right)^2 (t_{\alpha/2} + t_{\beta})^2$$

Easy, right?



Comparison of two means

Sample sizes

The last problem we have to solve is that the values of $t_{\alpha/2}$ and t_{β} are also dependent of n , which makes the equation

$$n = 2 \left(\frac{\hat{\sigma}}{\delta^*} \right)^2 (t_{\alpha/2} + t_{\beta})^2$$

transcendental in n . We'll have to iterate until we find the smallest n that satisfies:

$$n \geq 2 \left(\frac{\hat{\sigma}}{\delta^*} \right)^2 (t_{\alpha/2} + t_{\beta})^2$$

Usually $t_{\alpha/2} \approx z_{\alpha/2}$ is used for the first iteration. Easy, right?

Comparison of two means

Sample sizes

The last problem we have to solve is that the values of $t_{\alpha/2}$ and t_{β} are also dependent of n , which makes the equation

$$n = 2 \left(\frac{\hat{\sigma}}{\delta^*} \right)^2 (t_{\alpha/2} + t_{\beta})^2$$

transcendental in n . We'll have to iterate until we find the smallest n that satisfies:

$$n \geq 2 \left(\frac{\hat{\sigma}}{\delta^*} \right)^2 (t_{\alpha/2} + t_{\beta})^2$$

Usually $t_{\alpha/2} \approx z_{\alpha/2}$ is used for the first iteration. Easy, right?



Comparison of two means

Example: Length of steel rods

Required sample size:

```
> ss.calc<-power.t.test(delta=15,  
                        sd=15,  
                        sig.level=0.05,  
                        power=0.8,  
                        type="two.sample",  
                        alternative="two.sided")
```

Two-sample t test power calculation

n = 16.71477

delta = 15

sd = 15

sig.level = 0.05

power = 0.8

alternative = two.sided

NOTE: n is number in *each* group

Comparison of two means

Example: Length of steel rods

Computationally, we can perform the t-test for comparing the means of two independent populations by:

```
> y<-read.table("../data files/steelrods.txt",  
+               header=T)  
  
> with(y,  
+      t.test(Length.error~Process,  
+             alternative = "two.sided",  
+             mu = 0,  
+             var.equal = TRUE,  
+             conf.level = 0.95))
```

Two Sample t-test

data: Length.error by Process

t = -14.312, df = 32, p-value = 1.849e-15

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.09272982 -0.06962312

sample estimates:

mean in group new mean in group old

0.07782353

0.15900000

Comparison of two means

Example: Length of steel rods

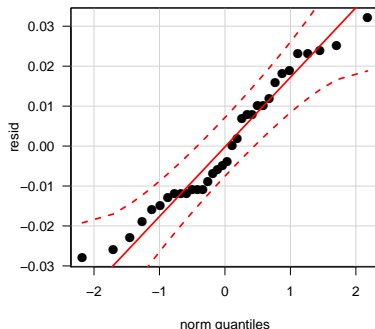
The assumptions of the test must be verified. In this particular case:

- Normality of the residuals;
- Equality of variance of the residuals;
- Independence of the residuals.

```
> resid<-y$Length.error - rep(means[2:1],  
+                             each=n)
```

```
> shapiro.test(resid)  
Shapiro-Wilk normality test  
data:  resid  
W = 0.9552, p-value = 0.176
```

```
> library(car)  
> qqPlot(resid,  
+        pch=16,  
+        cex=1.5,  
+        las=1)
```



Comparison of two means

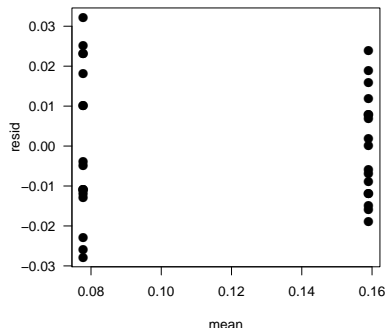
Example: Length of steel rods

The assumptions of the test must be verified. In this particular case:

- Normality of the residuals;
- Equality of variance of the residuals;
- Independence of the residuals.

```
> with(y,  
+      fligner.test(Length.error~Process))  
Fligner-Killeen test of homogeneity of variances  
data:  Length.error by Process  
Fligner-Killeen:med chi-squared = 1.6837,  
df = 1, p-value = 0.1944
```

```
> plot(x=rep(means[2:1],  
+           each=n),  
+      y=resid,  
+      xlab="mean",  
+      pch=16,  
+      cex=1.5,  
+      las=1)
```



Comparison of two means

Example: Length of steel rods

The assumptions of the test must be verified. In this particular case:

- Normality of the residuals;
- Equality of variance of the residuals;
- **Independence of the residuals.**

As mentioned in an earlier lecture, there is no general test for the independence assumption, and it has to be guaranteed in the design phase.

One can at most test for serial autocorrelation in the residuals using Durbin-Watson's test, but this test is absolutely dependent on the ordering of the observations - very useful to detect ordering-related trends in the residuals, but not much more than that.

Comparison of two means

Unequal variances

Suppose now a more general case, in which the variances of the two populations are unknown and cannot be assumed equal.

For this cases, a modification on the t-test called *Welch's t test* is usually employed. The Welch statistic can be calculated as:

$$t_0^* = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Under the null hypothesis t_0^* is distributed approximately as a t_ν distribution, with:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

Comparison of two means

Unequal variances

To illustrate the use of this technique, we can use the same data from the previous example^b:

```
> with(y,
+       t.test(Length.error~Process,
+             alternative = "two.sided",
+             mu = 0,
+             var.equal = FALSE,
+             conf.level = 0.95))
Welch Two Sample t-test
data:  Length.error by Process
t = -14.312, df = 28.386, p-value = 1.645e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09278780 -0.06956515
sample estimates:
mean in group new mean in group old
      0.07782353      0.15900000
```

^bNotice that this would not be necessary, since the data collected in the previous example did not violate the equality of variances assumption.

Bibliography

Required reading

- 1 D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, Ch. 10. 5th ed., Wiley, 2010.; **OR**
- 2 D.C. Montgomery, *Design and Analysis of Experiments*, Ch. 2. 5th ed., Wiley, 2005;
- 3 R. Nuzzo, *Scientific method: Statistical errors*, Nature 506(7487) - <http://goo.gl/Kbq6Rc>

Recommended reading

- 1 P. Mathews, *Sample Size Calculations: Practical Methods for Engineers and Scientists*, Ch. 1-2, 1st ed., MMB, 2010.
- 2 Radiolab (podcast): <http://radiolab.org>

About this material

Conditions of use and referencing

This work is licensed under the Creative Commons CC BY-NC-SA 4.0 license (Attribution Non-Commercial Share Alike International License version 4.0).

<http://creativecommons.org/licenses/by-nc-sa/4.0/>

Please reference this work as:

Felipe Campelo (2015), *Lecture Notes on Design and Analysis of Experiments*.

Online: <https://github.com/fcampelo/Design-and-Analysis-of-Experiments>

Version 2.11, Chapter 6; Creative Commons BY-NC-SA 4.0.

```
@Misc{Campelo2015-01,  
  title={Lecture Notes on Design and Analysis of Experiments},  
  author={Felipe Campelo},  
  howPublished={\url{https://github.com/fcampelo/Design-and-Analysis-of-Experiments}},  
  year={2015},  
  note={Version 2.11, Chapter 6; Creative Commons BY-NC-SA 4.0.},  
}
```



SOME RIGHTS RESERVED