UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Design and Analysis of Experiments

06 - Simple Comparisons

Version 2.12.2018

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

Image by Jorge Bispo: `https://goo.gl/KdqQkd`

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 cutting process with a new method that could potentially improve the performance of the system.

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:

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

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

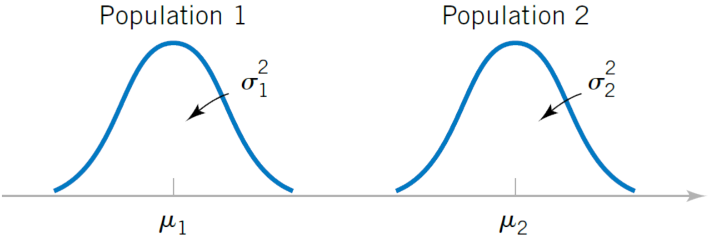


Image: D.C.Montgomery,G.C. Runger, *Applied Statistics and Probability for Engineers*,Wiley 2003.

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 processes. In this case, a possible response variable would be the *absolute error*, e.g., $y = |\ell - \ell_{nominal}|$.

The statistical hypotheses can be stated as:

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

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

Also, lets assume (for the moment) that the variance of the process is unknown but similar for both systems.

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:

$$S_p^2 = \frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{n_1 + n_2 - 2} = wS_1^2 + (1 - w)\, S_2^2$$

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:

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

we have that, under H_0:

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

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

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 relevant effect size (MRES): $\delta^* = 15cm$

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].

[a] Check, for instance, Paul Mathews' *Sample Size Calculations*, MMB, 2010.

Comparison of two means

Sample sizes

For the specific case of approximately equal variances, the optimal sample size ratio is $n_1 = n_2 = n$, with

$$n \approxeq 2 \left(\frac{t_{\alpha/2}^{(2n-2)} + t_{\beta}^{(2n-2)}}{d^*} \right)^2$$

where $d^* = \delta^*/\sigma$ is the (standardized) minimally interesting effect size; and $t_{\alpha/2}^{(2n-2)}$ and $t_{\beta}^{(2n-2)}$ are the $\alpha/2$ and β quantiles of the $t^{(2n-2)}$ distribution. .

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;
- Use a standardized MRES to calculate sample size;
- Perform a pilot study and collect samples to estimate the variance.

Each approach has its own advantages and drawbacks.

Comparison of two means

Sample sizes

For the steel rods experiment, suppose that the engineer uses data available from the system manuals, as well as historical measurements, to estimate a reasonable upper bound for the common standard deviation as $\sigma \approx 15cm$.

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

$$n \approx 2 \left[\left(t_{\alpha/2}^{(2n-2)} + t_{\beta}^{(2n-2)} \right) \frac{\sigma}{\delta*} \right]^2$$

Easy, right?

Comparison of two means
Sample sizes

The last problem we have to solve is that the values of $t_{\alpha/2}^{(2n-2)}$ and $t_{\beta}^{(2n-2)}$ are also dependent of n, which makes the sample size equation transcendental in n.

We can solve that by using an initial estimate of $t_{\kappa}^{(2n-2)} \approx z_{\kappa}$, and iterating until we find the smallest n that satisfies:

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

Image: http://vigilantmeadow.deviantart.com/art/DON-T-PANIC-165415311

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 = "one.sided")

Two-sample t test power calculation
n = 13.09777
delta = 15
sd = 15
sig.level = 0.05
power = 0.8
alternative = one.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 = TRUE)

> t.test(y$Length.error ~ y$Process,
+        alternative = "less",
+        mu          = 0,
+        var.equal   = TRUE,
+        conf.level  = 0.95)

data:  y$Length.error by y$Process
t = -14.312, df = 32, p-value = 9.244e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -7.156884
sample estimates:
mean in group new mean in group old
         7.782353         15.900000
```

Comparison of two means

Example: Length of steel rods

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

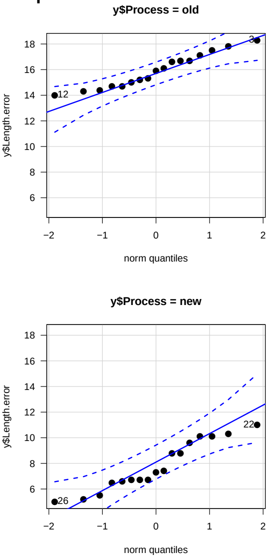
- Normality;
- Equality of variances;
- Independence.

```
> qqPlot(y$Length.error, groups = y$Process,
        cex = 1.5, pch = 16,  las = 1,
        layout = c(2, 1))

> shapiro.test(y$Length.error[y$Process == "new"])
# W = 0.92269, p-value = 0.164

> shapiro.test(y$Length.error[y$Process == "old"])
# W = 0.94971, p-value = 0.4519
```

Reminder: the t-test is quite robust to mild to moderate violations of the normality of the residuals / groups.



Comparison of two means

Example: Length of steel rods

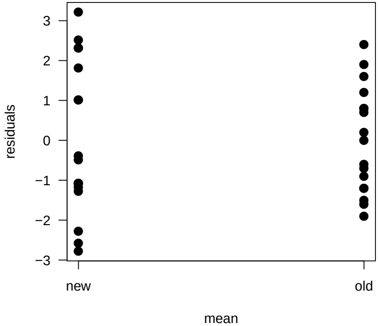
The assumptions of the test must be verified. In this particular case:

- Normality;
- Equality of variances;
- Independence.

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

> resid <- tapply(X = y$Length.error,
         INDEX = y$Process,
         FUN   = function(x){x - mean(x)})

> stripchart(x         = resid,
             vertical = TRUE,
             pch       = 16,
             cex       = 1.5,
             las       = 1,
             xlab      = "mean",
             ylab      = "residuals")
```



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

- Normality;
- Equality of variances;
- Independence.

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 $t^{(\nu)}$ distribution, with:

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

Comparison of two means
Unequal variances

The two-sample Welch t-test for considering unequal variances is usually the first test of choice, since it drops one (often inconvenient) assumption, at a very small cost in terms of power.

Calculating sample sizes for the general case (unequal variances, unequal sample sizes) is not particularly difficult, and can be done for either a *balanced* case (i.e., $n_1 = n_2 = n$) or an optimal, *unbalanced* case (in which $n_1 \neq n_2$).

For the unbalanced case, it is not particularly difficult to prove that the optimal allocation of observation is to keep:

$$\frac{n_1}{n_2} = \frac{\sigma_1}{\sigma_2}$$

.
(if a good estimate of the ratio of variances is available, of course).

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

Please reference this work as:

Felipe Campelo (2018), *Lecture Notes on Design and Analysis of Experiments*.
Online: https://github.com/fcampelo/Design-and-Analysis-of-Experiments
Version 2.12. Creative Commons BY-NC-SA 4.0.

```
@Misc{Campelo2018,
  title={Lecture Notes on Design and Analysis of Experiments},
  author={Felipe Campelo},
  howPublished={\url{https://github.com/fcampelo/Design-and-Analysis-of-Experiments}},
  year={2018},
  note={Version 2.12. Creative Commons BY-NC-SA 4.0.},
}
```