

Design and Analysis of Experiments

07 - Paired Design

Version 2.11

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

Graduate Program in Electrical Engineering

Belo Horizonte
March 2015

"I am driven by two main philosophies: know more today about the world than I knew yesterday and lessen the suffering of others. You'd be surprised how far that gets you."

Neil deGrasse Tyson
1958 -
American astrophysicist and author.



Comparison of two means

Dependent populations

Suppose the following situation: a young researcher develops an optimization algorithm (A) for a given family of problems, and wants to compare its convergence speed against a method that represents the state-of-the-art (B).

The researcher implements both methods and wants to determine whether the proposed one has a better average performance for problems of that particular family, represented by a given benchmark set.

The measurements are made under homogeneous conditions (same computer, same operational conditions, etc.) and the time is measured in a way that is not sensitive to other processes running in the system.



Comparison of two means

Dependent populations

This problem has some important questions worth considering:

- What is the actual question of interest?
- What is the *population* for which that question is relevant?
- What are the independent observations for that population?
- What is the relevant sample size for the experiment?

Comparison of two means

Paired design

The variability due to the different test problems is a strong source of spurious variation that can and must be controlled;

An elegant solution to eliminate the influence of this nuisance parameter is the *pairing* of the measurements by problem:

- Observations are considered in pairs (A, B) for each problem;
- Hypothesis testing is done on the sample of *differences*;

Comparison of two means

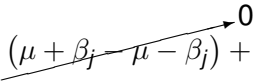
Paired design

Let y_{Aj} and y_{Bj} denote paired observations of average time for methods A and B, for each problem instance j . The *paired differences* of the observations are simply $d_j = y_{Aj} - y_{Bj}$.

If we model our observations as an additive process:

$$y_{ij} = \underbrace{\mu + \tau_i}_{\mu_i} + \beta_j + \varepsilon_{ij}$$

where μ is the grand mean, τ_i is the effect of the i -th algorithm on the mean, β_j is the effect of the j -th problem, and ε_{ij} is the model residual, then:

$$\begin{aligned} d_j &= (\mu + \beta_j - \mu - \beta_j) + \tau_A - \tau_B + \varepsilon_{Aj} - \varepsilon_{Bj} \\ &= \mu_D + \varepsilon_j \end{aligned}$$


Comparison of two means

Paired design

The hypotheses of interest can now be defined in terms of μ_D , e.g.:

$$\begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D \neq 0 \end{cases}$$

which can now be treated as a test of hypotheses for a single sample: the population of interest is the differences in average times until convergence for the problems under investigation. The test statistic is given by:

$$T_0 = \frac{\bar{D}}{S_D/\sqrt{N}}$$

which is distributed under the null hypothesis as a Student-t variable with $N - 1$ degrees of freedom (where N is the number of test problem instances in the experiment);

Comparison of two means

Paired design

Some other important questions worth considering:

- In this example the minimally interesting effect size δ^* must be expressed in terms of *average time gains across problems* (not within individual instances);
- The most important sample size to consider in this situation refers to the *number of problem instances*, and not necessarily to the number of within-problems repeated measures;
- The number of repetitions within each problem will have an impact on the uncertainty associated to each observation (that is, to each value of mean time to convergence for each algorithm on each problem), and should be selected with some care^a.

^aAlternatively, we can set it as arbitrarily large, particularly in cases where the cost of repetitions is small. As much as I hate to admit it, the lazy heuristic of setting it as ≥ 30 should be enough in most algorithmic studies. A more methodologically sound approach to setting this is under development, and will be included in future versions of these lecture notes.

Comparison of two means

Paired design

Some other important questions worth considering:

- Pairing removes the effects of controllable nuisance factors from the analysis.
- Strongly indicated in cases with strong correlations between samples (e.g., heterogeneous experimental conditions).

Comparison of two means

Paired design

Going back to our example, assume the following facts about the desired comparison:

- The benchmark set is composed of seven problems ($N = 7$);
- The researcher is interested in finding differences in mean time to convergence greater than ten seconds ($\delta^* = 10$) with a power of at least $(1 - \beta) = 0.8$, using a significance level $\alpha = 0.05$;
- The researcher performs $n = 30$ repeated runs^b of each algorithm in each problem, from random initial conditions.

^b Not that I necessarily recommend this number, but it is generally an easy alternative if you don't want to keep justifying your choices to less statistically-savvy reviewers.



Comparison of two means

Paired design

Step 1: load and precondition the data.

```
> # Read data
> data<-read.table("../data files/soltimes.csv",
+                  header=T)

# "Problem" is a categorical variable, not a continuous one
> data$Problem<-as.factor(data$Problem)

# Summarize within-problem observations by mean
> aggdata<-aggregate(Time~Problem:Algorithm,
+                    data=data,
+                    FUN=mean)
> summary(aggdata)
```

Problem	Algorithm	Time
1:2	A:7	Min. : 37.63
2:2	B:7	1st Qu.:109.45
3:2		Median :178.73
4:2		Mean :175.48
5:2		3rd Qu.:245.25
6:2		Max. :296.79
7:2		

Comparison of two means

Paired design

Step 2: analysis

```
> # Perform paired t-test  
> t.test(Time~Algorithm,  
+         paired=T,  
+         data=aggdata)
```

Paired t-test

data: Time by Algorithm

t = -9.1585, df = 6, p-value = 9.54e-05

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

95 percent confidence interval:

-21.85862 -12.64118

sample estimates:

mean of the differences

-17.2499

Comparison of two means

Paired design

Alternatively, we could have done:

```
> difTimes<-with(aggdata,  
+               Time[1:7]-Time[8:14])  
> t.test(difTimes)
```

One Sample t-test

data: difTimes

t = -9.1585, df = 6, p-value = 9.54e-05

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

-21.85862 -12.64118

sample estimates:

mean of x

-17.2499

Comparison of two means

Paired design

Verify assumptions:

```
> shapiro.test(difTimes)
```

Shapiro-Wilk normality test

data: difTimes

W = 0.8387, p-value = 0.09655

```
# Redo test without outlier
```

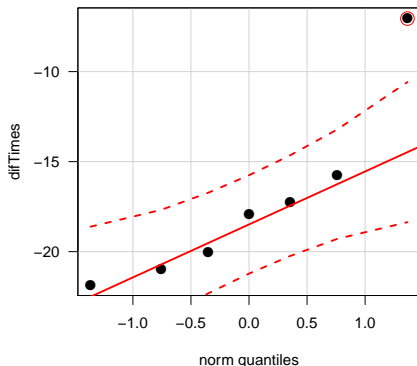
```
> indx<-which(difTimes==max(difTimes))
```

```
> t.test(difTimes[-indx])$p.value
```

```
[1] 6.179743e-06
```

```
> t.test(difTimes[-indx])$conf.int
```

```
[1] -21.41856 -16.48037
```



Comparison of two means

Paired design

What happens if we fail to consider the problem effects?

```
> t.test(Time~Algorithm,data=aggdata)
```

Welch Two Sample t-test

data: Time by Algorithm

t = -0.3609, df = 11.993, p-value = 0.7245

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

95 percent confidence interval:

-121.40320 86.90341

sample estimates:

mean in group A mean in group B

166.8527 184.1026

Comparison of two means

Paired design

Paired designs can require smaller sample sizes for equivalent power in cases where the between-units (in our example, the between-problems) variation is relatively high;

More specifically, if the within-level variation is given by σ_ϵ and the between-units variation is σ_u , we have that, for large enough N (e.g., $N \geq 10$),

$$\frac{N_{\text{unpaired}}}{N_{\text{paired}}} \approx \sqrt{2 \left[\left(\frac{\sigma_u}{\sigma_\epsilon} \right)^2 + 1 \right]}$$

Failure to consider inter-unit variability can result in the masking of relevant effects by the nuisance factor.

Similarly, failure in recognizing the dependence structure of within-unit measurements yields tests with artificially inflated degrees of freedom, which results in the inflation of the effective value of α .

Bibliography

Required reading

- 1 D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, Ch. 10. 5th ed., Wiley, 2010.
- 2 M.J. Crawley, *The R Book*, Ch. 8. 1st ed., Wiley, 2007;

Recommended reading

- 1 L. Lehe and V. Powell, *Simpson's Paradox* - <http://vudlab.com/simpsons/>
- 2 J.P. Simmons, L.D. Nelson, and U. Simonsohn, *False-Positive Psychology : Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant*, Psychological Science 22(11):1359-1366, 2011 - <http://goo.gl/9e0cdw>

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 7; 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 7; Creative Commons BY-NC-SA 4.0.},  
}
```

