Experiment Design for Computer Sciences (01CH740)

Topic 04 - Paired Comparison

Claus Aranha

caranha@cs.tsukuba.ac.jp

University of Tsukuba, Department of Computer Sciences

May 22th, 2020

Version 2020.1 (Updated May 21, 2020)

Outline

- Comparison between multiple quantities - Algorithm A or algorithm B?

Report 01

- Evaluation date
- Reproducibility

Questions about Induction

- It's online buy ask questions! - Do you know how to:...

Report 02

- Report 02 description

Researcher

Part I – Two Sample Testing

Statistical Inference for Two Samples

Sometimes we are interested on the comparison between two different populations, based on information from their samples. This type of analysis is frequent when we compare the effect of a technique (**or treatment**) against a *control group*: a placebo, a classical technique, a random search, etc;

The statistics used in this case are actually very similar to the statistics used for the analysis of single populations; and in general the experiment design follow the same principles.

Usual questions involve:

- The comparison of means;
- The comparison of variances;
- The 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.

Statistical Models

A Statistical Model is a useful way to characterize a population from which we obtain some sample. This model describes the possible values from an experiment, and how they are distributed.

For example, when we measure some observed value (y) taken from one of several methods ($i = 1, 2, \ldots$), we understand that the value comes from some distribution with mean μ_i, at it will also have an error (ϵ) away from that mean, which is different for each observation. So we describe the j-th observation taken from the i-th method as

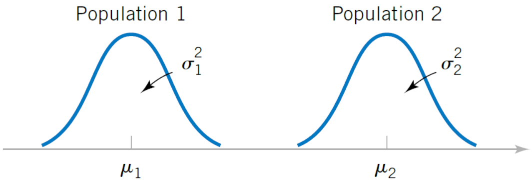
$$y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2 \\ j = 1, \ldots, n_i \end{cases}$$

Statistical Models

Two population Model

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

Using this model for the observed variable (y_{ij}), we assume that the residuals ϵ_{ij} are iid[1] and follow $\mathcal{N}\left(0, \sigma_i^2\right)$. Under these assumptions, the populations of the two samples look like this:



[1] Independent and Identically Distributed

Comparison of two means

Null and Alternate Hypotheses

What should be the observed variable y? The goal of this experiment is to measure if the new method produces steel rods closer to the nominal value. In this case, a possible response variable would be the absolute error, e.g., $y = |\ell - \ell_{nominal}|$.

Keeping in mind our statistical model, we can build the hypothesis around the mean of the absolute error (μ_i). In that case, we can state the null and alternate hypotheses as:

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

Comparison of two means

Calculating the statistic

Lets assume (for the moment) that the variance of the process is unknown but similar for both systems. Since it is unknown, we have to estimate the variance from the sample data. As assume $\sigma_1^2 \approx \sigma_2^2$, 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

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

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, <u>under H_0</u>:

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

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

Comparison of two means

Calculating the Statistic

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

```
> y <- read.table("steelrods.csv", 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
```

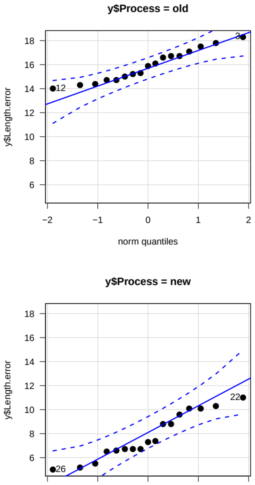
Comparison of two means

Testing the assumptions

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

Testing the assumptions

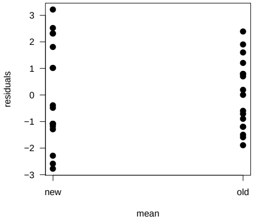
The assumptions of the test must be verified. In this particular case:

- 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")
```



Comparison of two means

Testing the assumptions

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

- Normality;
- Equality of variances;
- Independence;

As mentioned in the last class, 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

To illustrate this technique, let's use the data from the example[2]:

```
> with(y,
+      t.test(Length.error~Process,
+             alternative = "two.sided",
+             mu = 0,
+             var.equal = FALSE,                %% <- We only change this.
+             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
```

[2]Notice that this would not be necessary, since the data collected in the previous example did not violate the equality of variances assumption.

Comparison of two means

Summary

To compare an estimator from samples of two populations that (we assume) follow a normal distribution, we set our statistic and the corresponding hypotheses to be the difference of the target variables.

This technique for comparison testing is equally simple and extremely versatile.

Of course, there are cases where this approach does not apply. Next we will see a relatively common case where using the difference of the target variables would lead to a wrong inferential result.

Part II – Paired Testing

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, which are 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 are the relevant sample sizes for the experiment?

Consider carefully the difference between considering *individual runs* as a population against *individual problem instances* as a population. The important thing is to not mix both!

Paired Experimental Design

The variability of results due to the different test problems is a strong source of spurious variation (noise) 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 *problem differences*;

Paired Design

Statistical Model

Let y_{Aj} and y_{Bj} denote paired observations of average time for methods A and B, for each problem instance j. The *paired difference* of an observation is 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 method on the mean, β_j is the effect of the j-th problem instance, and ε_{ij} is the model residual, then:

$$d_j = \mu + \tau_A + \beta_j + \varepsilon_{Aj} - (\mu + \tau_B + \beta_j + \varepsilon_{Bj})$$

$$\begin{aligned} d_j &= \cancelto{0}{\left(\mu + \beta_j - \mu - \beta_j\right)} + \tau_A - \tau_B + \varepsilon_{Aj} - \varepsilon_{Bj} \\ &= \mu_D + \varepsilon_j \end{aligned}$$

Paired Design

Hypotheses

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);

Paired Design

Considerations

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), which will propagate down to the residual variance.

Paired design

Considerations

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).

Paired Comparison Example

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

- The benchmark set is composed of seven problem instances ($N = 7$);
- 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[1] of each algorithm in each problem, from random initial conditions.

[1] Not that this number is necessarily good, but it is generally an easy alternative if you don't want to keep justifying your choices to less statistically-savvy reviewers.

Executing the Paired Analysis

Step 1: load and precondition the data

```
> # Read data
> data <- read.table("benchmark.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
```

Executing the Paired Analysis

Step 2: analysis

```
> # Perform paired t-test
> t.test(Time ~ Algorithm,
+        paired = TRUE,
+        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
```

Executing the Paired Analysis

Alternatively, we could have done:

```
> difTimes <- aggdata$Time[1:7] - aggdata$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
```

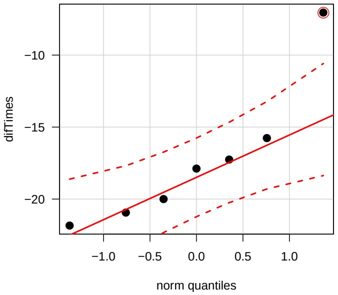
Check your understanding: Why is the paired test on two samples equivalent to the one sample test on the difference vector of the samples?

Verifying the 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
```



Why is Pairing Important?

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
```

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.

Conclusion

Summary

About these Slides

These slides were made by Claus Aranha, 2020. You are welcome to copy, re-use and modify this material.

These slides are a modification of "Design and Analysis of Experiments (2018)" by Felipe Campelo, used with permission.

Individual images in some slides might have been made by other authors. Please see the references in each slide for those cases.

Image Credits I

[Page 9] Steel rod image:
`http://www.shutterstock.com/pic-73207399/`
[Page 11] Two models image from D.C. Montgomery "Applied Statistics and Probability for Engineers", Wiley 2003