Clínica y Salud

Opinion Article

Procedures for Comparison of Two Means in Independent Groups with R

Alfonso Palmer, Albert Sesé, Berta Cajal, Juan J. Montaño, Rafael Jiménez, and Elena Gervilla

University of the Balearic Islands, Spain

The comparison of two means in independent groups is one of the most used statistical procedures. In the bibliometric study carried out by Sesé and Palmer (2012) on the 623 articles published in 8 impact journals of Clinical and Health Psychology, in 2010, it was found that the parametric comparison of two means in independent groups was the second most used statistical technique (n = 161 times) only surpassed by correlation (n = 207), while non-parametric techniques in general (n = 54), resampling (n = 12), and robust (n = 4) were in descending order. However, only 17.27% of the articles evaluated the assumptions of the procedure used, failing the rest in one of the fundamental aspects of statistics. The objective of this work is to show different procedures that should be used depending on the scenario in which the data analysis is found.

Data Matrix

In this analysis we have a categorical variable, which can be binary or a factor with k = 2 levels and a vector of continuous responses. This information can be entered into R by two data vectors. Below are provided, as an example, the fictitious responses of 12 subjects belonging to each of the two groups:

Y1	6	8	9	12	9	10	8	11	11	12	10	10
Y2	13	13	14	16	15	12	17	14	15	18	17	20

Speaking in generic terms, vectors Y1 and Y2 can correspond to an experimental group and a control group, to two experimental conditions, to responses of men and women, etc. In terms of R, the two response vectors are defined:

```
> Y = c(6, 8, 9, 12, 9, 10, 8, 11, 11, 12, 10, 10)
> Y2 = c(13, 13,14, 16, 15, 12, 17, 14, 15, 18, 17, 20)
A vector, called Y, is generated that joins both vectors Y1 and Y2.
Y = c(Y1,Y2)
> Y
 [1] 6 8 9 12 9 10 8 11 11 12 10 10 13 13 14 16 15 12 17 14 15 18 17 20
```

The dichotomous variable Grupo is generated, which defines which group each response corresponds to and which differentiates the responses of vector Y1, with a value of 0, from the responses of vector Y2, with a value of 1.

```
Grupo = as.factor(c(rep(0,12),rep(1,12)))
> Grupo
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1
Levels: 0 1
```

Finally, a data matrix that will be named Data is created, with two columns: the first is the Y vector and the second is Grupo.
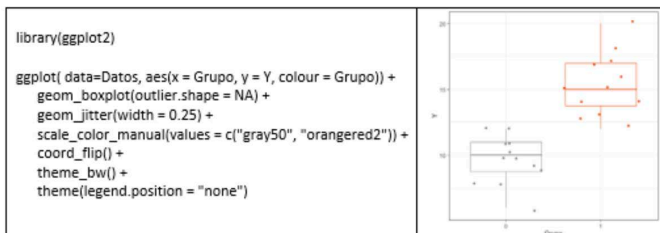
```
> Datos = data.frame(Y, Grupo)
```

The sample means of the two groups are obtained, which allow the comparison of the two population means, as well as the difference of the sample means:

> mean(Y1)	[1] 9.666667
> mean(Y2)	[1] 15.33333
> difm=mean(Y1)-mean(Y2)	[1] -5.66667

Plot

The box plot of the two distributions is displayed using the *ggplot* function of the *ggplot2* library:


```
library(ggplot2)

ggplot( data=Datos, aes(x = Grupo, y = Y, colour = Grupo)) +
    geom_boxplot(outlier.shape = NA) +
    geom_jitter(width = 0.25) +
    scale_color_manual(values = c("gray50", "orangered2")) +
    coord_flip() +
    theme_bw() +
    theme(legend.position = "none")
```

The following is a non-exhaustive overview of the different scenarios in which a researcher may find himself having to make a comparison of two means in independent groups and of the procedures to use, with R, in each scenario. The syntax in R and the most relevant results are provided. A broader description and the formulas on which each test is based can be found in Palmer (2011), while practical applications, with R, can be found in Palmer (2013) and Palmer (2019).

Parametric Tests

Assumptions. Parametric tests require the fulfillment of some assumptions, that in the case of a comparison of two means in independent groups refer to normality and equality of variances (homoscedasticity) in the original populations of both groups. These conditions of application are evaluated below:

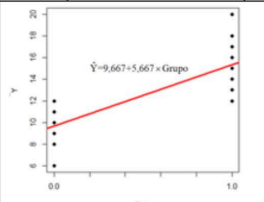| Normality		Homoscedasticity
> shapiro.test(Y1)	> shapiro.test(Y2)	> library(car) > leveneTest(Y,Grupo,center=median)
W = 0.94666, p-value = 0.5888	W = 0.96262, p-value = 0.8205	``` Df F value Pr(>F) group 1 0.9083 0.3509 22 ```

Thus, the Shapiro-Wilk test allows us to conclude the null hypothesis of normality in the original populations of both groups, while the robust Levene test maintains the null hypothesis of equality of population variances.

It should be remembered that the evaluation of the assumptions must always be carried out and, based on its results, a correct decision can be made about the procedure to follow.

***t* test.** The parametric test to use when both assumptions are met, is the Student's *t*-test. In this case, in addition to providing the test value and its degree of significance (*p*-value), the APA recommends also providing the confidence interval, generally 95%, (Palmer y Sesé, 2013).

> t.test(Y1,Y2,var.equal=TRUE)	Two Sample t-test t = -6.6679, df = 22, p-value = 1.054e-06 IC(95%): [-7.429120 ; -3.904213]

Linear regression model. Equivalently to Student's *t*, the linear regression model can also be applied, but in this case using the *lm* function:

```
> Modelo=lm(Y~Grupo)                      > for(i in 1:24) { if(Grupo[i]=="0")
> summary(Modelo)                          {Grup[i]=0} else {Grup[i]=1}}
> confint(Modelo)                          > plot(Grup,Y,pch=16)
> anova(Modelo)                            > abline(Modelo,col="red",lwd=3)

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.6667     0.6009  16.086 1.19e-13
Grupo        5.6667     0.8498   6.668 1.05e-06

              2.5 %   97.5 %
Grupo       3.904213  7.42912

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
Grupo      1 192.667 192.667  44.462 1.054e-06
Residuals 22  95.333   4.333
```

The Grupo coefficient is the difference of means. In this case, when moving the Grupo from 0 to 1, the mean of Y increases 5.667, so it appears with a positive sign. For this reason, the CI (95%) appears with a change in sign to that of the *t*-test, but they are equivalent.

Model comparison. In this case, the model comparison consists in comparing the predicted value of Y with (extended model) or without (reduced model) the Grupo variable. The comparison of models is applicable to the parametric way and its results are equivalent to the *t*-test.

The following table shows the application to the parametric way:

# Parametric > Model0=lm(Y~1) > Model1=lm(Y~Grupo) > anova(Model0,Model1)	Analysis of Variance Table ``` Res.Df RSS Df Sum of Sq F Pr(>F) 1 23 288.000 2 22 95.333 1 192.67 44.462 1.054e-06 ```

Welch test. If the homogeneity of variance is not satisfied, then the Welch test is used, which is a Student's *t*-test with a modification in the degrees of freedom:

> t.test(Y1,Y2,var.equal=FALSE)	Welch Two Sample t-test t = -6.6679, df = 20.477, p-value = 1.524e-06 IC(95%): [-7.436752 ; -3.896582]

Weighted linear model. Under non-equality of variances, the weighted linear model can be used, where the weight of each observation is the inverse of the variance of the group to which it belongs. The weight affects the value of the standard errors.

> peso=c(rep(var(Y1),12),rep(var(Y2),12)) > Modelow=lm(Y~Grupo,weights=peso) > summary(Modelow)	``` Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 9.6667 0.7304 13.235 5.93e-12 Grupo 5.6667 0.9156 6.189 3.14e-06 ```

Sample size, effect size and power. Some important elements in a hypothesis test have to do with the sample sizes of each group, the size of the effect and the Power of the statistical test. All of these elements are related to each other, together with the alpha risk used (generally the value .05) and the type of test used (two-tailed or one-tailed).

Effect size is a standardized measure of the difference between means. A very popular measure is Cohen's (1988) *d*, which can be obtained using the *cohen.d* function from the *effsize* library:

> library(effsize) > cohen.d(Y~Grupo)	Cohen's d d estimate: -2.722179 (large)

Cohen differentiates between small ($d = 0.20$), medium ($d = 0.50$) and large ($d = 0.80$) effect sizes, although it is better to interpret it in terms of percentiles (Palmer, 2011). With a Cohen's *d* of 2, 97.7% of the "treatment" group will be above the mean of the "control" group (Cohen's U3), 31.7% of the two groups will overlap, and there is a 92.1% probability that a person chosen randomly from the treatment group will score higher than a person chosen randomly from the control group (superiority probability).

In the following link, https://rpsychologist.com/cohend/, entering the value of the index *d* obtained (the maximum is 2), the values and the above mentioned interpretations are provided.

The power of the statistical test is the probability of rejecting the null hypothesis when it is false. The APA recommends always providing this term (American Psychological Association [APA, 2010]) Its value can be obtained using the *pwr.t2n.test* function from the *pwr* library.

> library(pwr)   > pwr.t2n.test(n1=12,n2=12,sig.level=0.05,d=2.722,alt="two.sided",power=)   power = 0.9999945

If the alternative hypothesis is one-sided, it will be enough to change the parameter alt = "less" or alt = "greater". The power

obtained is very high (0.99). If a low power is obtained, for example 0.22 when obtaining a $d = 0.593$ and the power of the test is to be increased to a certain value, the minimum sample size to use should be found. This minimum value of n, for each group, can be obtained with the *pwr.t.test* function, entering, among others, the value of the desired power, for example 0.80:

```
> library(pwr)
> pwr.t.test(sig.level=0.05,d=0.593,type="two.sample",power=0.80,n=)
n = 45.62085
```

Therefore, to have a power of 80%, the size of each group should be at least $n = 46$.

Nonparametric Tests

When the assumption of normality is not fulfilled, a non-parametric test can be used. In these tests, the original data is transformed into order numbers.

Wilcoxon-Mann-Whitney test. One of the most common tests is the so-called Mann-Whitney U, also known as Wilcoxon-Mann-Whitney. It can be obtained through the *wilcox.test* function:

> wilcox.test(Y1,Y2)	Wilcoxon rank sum test with continuity correction W = 1, p-value = 4.433e-05

Linear regression model. The linear regression model can be applied, equivalently to the U test, in the non-parametric way, transforming the original values by their order numbers and applying the *lm* function:

```
> Yr=rank(Y)           Coefficients:
> Modelor=lm(Yr~Grupo)             Estimate Std. Error t value Pr(>|t|)
> summary(Modelor)     (Intercept)    6.583      1.070   6.152 3.42e-06
> anova(Modelor)       Grupo         11.833      1.513   7.819 8.61e-08

                       Response: Yr
                                 Df Sum Sq Mean Sq F value     Pr(>F)
                       Grupo      1 840.17  840.17  61.137 8.613e-08
                       Residuals 22 302.33   13.74
```

Model comparison. The comparison of models is applicable to the non-parametric route, its results being equivalent to the Mann-Whitney U test. To apply it in the nonparametric way, the vector Y must be transformed into order numbers, which is achieved with the *rank* function:

```
# Nonparametric        Analysis of Variance Table
> Yr=rank(Y)
> Model0=lm(Yr~1)        Res.Df     RSS Df Sum of Sq      F    Pr(>F)
> Model1=lm(Yr~Grupo)  1     23 1142.50
> anova(Model0,Model1) 2     22  302.33  1    840.17 61.137 8.613e-08
```

Robust test. When normality is not fulfilled and, above all, when there are "outlier" values, a robust test should be used, which uses a robust statistic instead of the mean (see Palmer, 1999).

Yuen test. The Yuen test uses the trimmed mean. This test can be obtained with the yuen.t.test function from the PairedData library which by default cuts 20%:

```
> library(PairedData)  Two-sample Yuen test, trim=0.2
> yuen.t.test(Y1,Y2,trim=0.2)  t = -5.831, df = 12.848, p-value = 6.149e-05
                       IC(95%): [-7.368814 ; -3.381186]
```

You can also get a bootstrap version of the Yuen test, with the *yuenb* function from the *WRS2* library:

```
> library(WRS2)                    $ci  [-7.3708 ; -3.3792]
> yuenbt(Y~Grupo,tr=0.20,nboot=599,side=T)  $test.stat   -5.3714
                                   $p.value     0
```

Compare M-estimators. A comparison of two M-estimators can be made instead of comparing two means using the *pb2gen* function of the *WRS2* library.

The final values of the Huber M-estimators are:

```
> mest(Y1)   [1] 9.766414
> mest(Y2)   [1] 15.25413
```

```
> library(WRS2)                    Test statistic: -5.5252, p-value = 0.04715
> set.seed(123)                    IC(95%): [-7.2386 ; -2.9343]
> pb2gen(Y~Grupo,est="onestep",nboot=5599)
```

In this case, this function uses the value of the first iteration of the M-estimator

Exact test. With small samples or when population distributions are unknown, resampling techniques, such as bootstrap or permutations procedures, can be used.

For example, with the *boot.ci* function of the *boot* library, different formats of the CI (95%) can be obtained, while with the *oneway_test* function of the *coin* library, the exact Fisher-Pitman permutations test can be obtained:

```
> library(coin)                    Exact Two-Sample Fisher-Pitman
> Grupo=as.factor(Grupo)           Permutation Test
> oneway_test(Y~Grupo,distribution="exact")  Z = -3.9226, p-value = 2.219e-06
```

Summary

Below are the results of the *p*-value for each of the statistical tests used. As can be seen, all the tests reach the same conclusion: The alternative hypothesis (*p*-value < .05) according to which there are significant differences between the two analyzed means can be accepted (see Table 1).

Also, this can be verified by seeing that the confidence intervals do not contain the value zero. The most important thing is to use the best procedure in each case based on the circumstances that arise. In the example presented, when the application conditions are met, the most powerful procedure to choose would be the parametric route. As there is normality, the non-parametric way is not required. As there are no outliers, the Robust way is not required, and although

Table 1. *p* value for Each of the Statistical Test Used

	t-test	Model Linear Param.	Welch	Mann-Whitney	Model Linear Nonpar	Model Comp Param.	Model Comp Nonpar	Yuen	Yuen boot	M-estim	Fisher-Pitman
p-value	1.054 e-06	1.054 e-06	1.524 e-06	4.433 e-05	8.613 e-08	1.054 e-06	8.613 e-08	6.149 e-05	0	0.04715	2.219 e-06

the exact way can be used, as it is small samples, in this case the parametric way is more powerful.

In this example, all the procedures provide similar results since the conditions of application are met which will not happen when any of the assumptions are not fulfilled.

The diagram in the following page provides an overview of the content explained in this article (Figure 1).
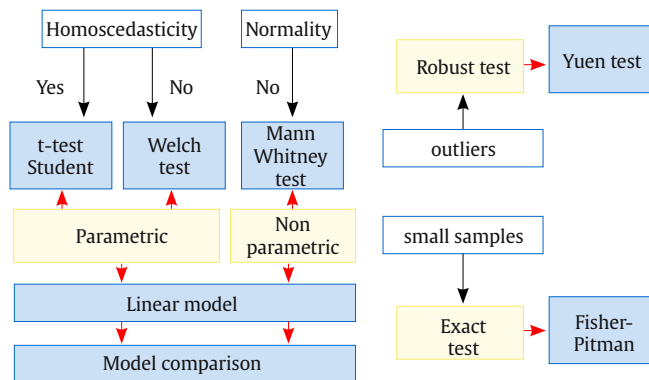
Figure 1. Comparison of two Means in Independent Groups.

References

American Psychological Association (APA, 2010). *Publication manual of the American Psychological Association.* American Psychological Association.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Palmer, A. (1999). *Análisis de datos. Etapa exploratoria.* Ediciones Pirámide.

Palmer, A. (2011). *Procedimientos estadísticos con SPSS y R para la comparación de dos medias.* Edicions UIB.

Palmer, A. (2013). *Estadística aplicada con R. 1. Prácticas de comparación de medias.* Edicions UIB.

Palmer, A. (2019). *Estadística aplicada con R. 7. Análisis de la comparación de dos medias con el Package P2Means.* Edicions UIB.

Palmer, A. & Sesé, A. (2013). Recommendations for the use of statistics in clinical and health psychology. *Clínica y Salud, 24*(1), 47-54. https://doi.org/10.5093/cl2013a6

R Core Team. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Sesé, A., & Palmer, A. (2012). The current use of statistics in clinical and health psychology under review. *Clínica y Salud, 23*(1), 97-108. https://doi.org/10.5093/cl2012v23n1a