Analysis of Paired Data

Jeffrey Michael Franc, MD, MS(Stats), MSc (DM), FCFP(EM), D Sport Med

Associate Professor, Department of Emergency Medicine, University of Alberta
Visiting Professor in Disaster Medicine, Università del Piemonte Orientale
Adjunct Faculty, Harvard/BIDMC Disaster Medicine Fellowship

Correspondence:
Jeffrey Michael Franc, MD, MS (Stats)
MSc (DM), FCFP(EM), Dip Sport Med
Research Director
Department of Emergency Medicine
University of Alberta, Alberta, Canada
736c University Terrace, 8203-112 Street NW
Edmonton, AB, Canada, T6G 2T4
E-mail: jeffrey.franc@ualberta.ca

Conflicts of interest: JMF is the CEO/Founder of STAT59 and the Editor-in-Chief of *Prehospital and Disaster Medicine*.

Keywords: paired data; statistics; t-Test

Abbreviation:
PPE: personal protective equipment

Received: March 20, 2025
Accepted: March 21, 2025

doi:10.1017/S1049023X25001517

Abstract
A common and unfortunate error in statistical analysis is the failure to account for dependencies in the data. In many studies, there is a set of individual participants or experimental objects where two observations are made on each individual or object. This leads to a natural pairing of data. This editorial discusses common situations where paired data arises and gives guidance on selecting the correct analysis plan to avoid statistical errors.

Franc JM. Analysis of paired data. ***Prehosp Disaster Med.*** **2025;40(2):61–63.**

Introduction
A common and unfortunate error in statistical analysis is the failure to account for dependencies in the data. Dependencies develop when data points are not independent but instead are more closely related to some data points than to others. For instance, in a study that analyzes the time from ambulance dispatch to scene for 1,000 ambulance transportations in five different countries, we would expect the observations in one country would be more like other observations in that country than observations in another country. Any statistical analysis would need to accommodate these dependencies.

One of the more common dependencies - that is often overlooked - is paired data.

What is Paired Data?
In many studies, there is a set of individual participants or experimental objects where two observations are made on each individual or object. This leads to a natural pairing of data.

Xu, et al elegantly summarize the issue: "The reason for this confusion revolves around whether we should regard two samples as independent (marginally) or not. If not, what's the reason for correlation?"[1]

The Video Teaching Experiment
For the sake of an example, we will use an imaginary experiment. Assume a researcher is interested in the effect of a video training module on the ability to correctly don personal protective equipment (PPE) among first responders. Her researcher's hypothesis is that a short video lesson will improve prehospital provider's score on a standardized PPE scoring checklist.

For the sake of this experiment, we will assume that the standardized checklist is scored as a percentage from 0 to 100, has been previously validated, and that the scores are known to be normally distributed. She will consider P values of $< .05$ to be significant.

The Two Sample t-Test
One way the researcher could approach this problem would be to randomize participants to either receive the video lesson or not. Then, each person is given a test using the standardized checklist and the scores are recorded. She chooses 20 people and randomizes 10 to each group. Here, the statistical null hypothesis is that the score is equal in the video and control group. The alternative hypothesis is that the scores are not equal.

H0: $\mu_{video} = \mu_{control}$ (Mean score is the same in both groups).
HA: $\mu_{video} \neq \mu_{control}$ (Mean score is not the same in both groups).

A sample of what the data may look like is presented in Table 1. The number of observations (n) – also known as the sample size – for this experiment is 20. Each row of the table is fully independent. Statistically, we consider that the video group and the control group are completely independent, and we can use a t-test.

This is commonly called the two-sample t-test. The two sample t-test requires that both populations are normal, so that $X_1, X_2, \ldots, X_M$ is a random sample from a normal distribution and so is $Y_1, \ldots, Y_n$ (with the X's and Y's independent of one another).[2]

Name	Group	Score
Tyler	Control	50
Henrietta	Control	55
Davina	Control	58
Laena	Control	61
Casper	Control	61
Anges	Control	46
Bertie	Control	49
Findlay	Control	37
Rosa	Control	73
Charles	Control	73
Aiden	Video	64
Darcy	Video	77
Lilian	Video	80
Abdullah	Video	64
Casey	Video	76
Daewoo	Video	94
William	Video	59
Janice	Video	72
Beka	Video	71
Ava	Video	45

Table 1. Video Training Experiment Unpaired Groups

We can quickly analyze the data using any statistics software. For example, using the R programming language:

```
>    t.test(score~group,    data=unpaired,
var.equal=TRUE);
    Two Sample t-test
data: score by group
t = 2.5093, df = 18, p-value = 0.02188
alternative hypothesis: true difference in
means between group control and group video is
not equal to 0
95 percent confidence interval:
 25.5379 2.244027
sample estimates:
mean in group control    mean in group video
                 56.3                   70.2
```

The output shows the calculated degrees of freedom (df) is 18 - based on the total sample size of 20 and two groups. We see that the mean in the control group (56.3) was lower than the mean of the video group (70.2). In addition, the P value of .022 shows a significant difference between the groups. Our researcher writes the following in her paper:

> *There was a significant difference between the mean score in the control group (56.3%) and the video group (70.2%); (95% CI for difference 2.2 to 25.4; P = .022).*

The Paired t-Test

On the other hand, the researcher could have chosen a different study design. In this case, she invites 10 participants. First each participant completes a pre-course test using the standardized checklist. Then, each person attends the video lesson and takes the same evaluation as a post-test. The primary outcome measure is

Name	Pre	Post
Ethan	50	64
Axle	55	77
Faizen	58	80
Michael	61	64
Marius	61	76
Anna	46	94
Curtis	49	59
Jordana	37	72
Nettie	71	73
Ievan	73	45

Table 2. Video Training Experiment Paired

difference between the pre- and post-test score for each participant. Here the sample size (n) is 10. The statistical hypothesis is that the true difference between pre- and post- scores is zero. The alternative hypothesis is that the true difference is not zero.

H_O: $\mu_D = 0$ (True mean difference is equal to zero).
H_A: $\mu_D \neq 0$ (True mean difference is not equal to zero).

Where D = Post-test score – Pre-test score.

A sample of how the data would look is presented in Table 2. Of note, the numbers are exactly the same as in Table 1, just arranged into a different format. Here we see that there are only 10 independent samples - each pair of pre- and post-test is not independent as they are performed by the same person. Statistically, we cannot use the two-sample t-test as we did above as the requirements for the two-sample t-test explicitly state that X's and Y's must be independent.

We can, however, use the paired t-test. The paired t-test requires that the data consists of (n) independently selected pairs $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$, so that the difference between pairs (D), where $D_n = X_n - Y_n$, is normally distributed.[2] We can quickly analyze the data using any statistics software. For example, using the R programming language:

```
>    t.test(x=paired$pre,    y=paired$post,
paired=TRUE);
    Paired t-test
data: paired$pre and paired$post
t = 2.1162, df = 9, p-value = 0.06343
alternative hypothesis: true mean difference
is not equal to 0
95 percent confidence interval:
 -0.9586592 28.7586592
sample estimates:
mean difference
           13.9
```

The output shows a lower degree of freedom (df = 9) than the two-sample test above. This is calculated from the sample size of 10 observations in one group. While the mean difference of scores was 13.9, the P value of .063 indicates no significant difference between the pre- and post-test scores. The confidence interval for the mean difference is wide (-0.96 to 28.8) and contains zero within it. This wide confidence interval is caused by larger error

https://doi.org/10.1017/S1049023X25001517 Published online by Cambridge University Press

terms secondary to the smaller effective sample size. Our researcher writes the following in her paper:

> *There was no significant difference between pre- and post-test scores (Mean Difference = 13.9; 95 % CI -0.96 to 28.8; P = .063).*

What's the Difference?

What makes this distinction between the two-sample t-test and the paired t-test so important? Firstly, as we saw above, choosing the correct statistical analysis will change the results. We saw above that when analyzed by one method, the result was statistically significant, while using the other method on the same numbers gives a non-significant result.

This is also extremely important for sample size planning and power calculations. We saw that for paired tests, the total number of observations (n) is the number of participants in the study, not the total number of recorded numbers. This is a common error among researchers. For example, where the sample size calculation requires a sample size of (for instance) 20 observations, and the researcher mistakenly plans for 10 individuals but two measurements from each. Statistically, this is not at all the same.

Non-Parametric Tests

While the examples above used parametric testing (t-test), the same distinction is also required for non-parametric testing. For instance, a non-parametric analog to the two-sample t-test would be the Wilcoxen rank-sum test (Mann-Whitney test).

For paired samples, the Wilcoxen Signed-Rank test (with the null hypothesis being a median of 0) can be used to analyze the differences between pre- and post-tests for each participant.

Dependencies Just Sneak In

While the example above featured a pre- and post-test design, paired data can also sneak into many other study designs – often without the research noting it initially. Before and after treatment studies are a common design. Crossover studies, where participants are assigned to one of two treatment arms, and then after a period, cross over to the other treatment arm, also produce paired data. Paired data can also arise from quasi-experimental designs, for instance, comparing individual prehospital provider score on a standardized burnout scale before and after a terrorist event.

While paired designs are a very common source of dependencies, other dependencies also appear frequently in disaster medicine and prehospital medicine research. For instance, a common scenario to test a triage method is to have each study participant triage several simulated patients. If the study recruits 20 participants, each triaging 10 simulated patients, this is not a sample size of 200. Statistically, the 10 triages performed by one participant are not independent of each other. The statistical methodology must account for this mathematically.

Planning for Paired Data

For all researchers, developing a statistical plan in advance to get the correct analysis is crucial. It is a very frustrating experience to perform a study, analyze the results, and write the manuscript only to find out that the wrong analysis was performed. If the revised analysis reveals different statistical conclusions, rewriting the entire paper may be required.

Fortunately, analysis of paired data is usually simple if planned for in advance. As always, researchers who have any doubt on how to correctly analyze their data should consult a statistician early – ideally in the planning phase of the project before data are collected.

Author Contribution

JMF performed the conception, writing, and final approval of the manuscript.

Use of AI Technology: No AI technology was used.

References

1. Xu M, Fralick D, Zheng JZ, Wang B, Tu XM, Feng C. The differences and similarities between two-sample t-test and paired t-test. *Shanghai Arch Psychiatry*. 2017;29(3): 184–188.
2. Devore JL. *Probability and Statistics for Engineering and the Life Sciences*. Seventh Edition. Belmont, California USA: Thomson Brooks/Cole; 2008:p336–346.

https://doi.org/10.1017/S1049023X25001517 Published online by Cambridge University Press