Parametric Statistical Inference for Comparing Means and Variances

Johannes Ledolter,[1,2] Oliver W. Gramlich,[2,3] and Randy H. Kardon[2,3]

[1]Department of Business Analytics, Tippie College of Business, University of Iowa, Iowa City, Iowa, United States
[2]Center for the Prevention and Treatment of Visual Loss, Iowa City VA Health Care System, Iowa City, Iowa, United States
[3]Department of Ophthalmology and Visual Sciences, University of Iowa, Iowa City, Iowa, United States

Correspondence: Johannes Ledolter, Department of Business Analytics, Tippie College of Business, University of Iowa, 108 John Pappajohn Business Building, Iowa City, IA 52242, USA; johannes-ledolter@uiowa.edu.

Received: May 3, 2020
Accepted: June 3, 2020
Published: July 21, 2020

Citation: Ledolter J, Gramlich OW, Kardon RH. Parametric statistical inference for comparing means and variances. *Invest Ophthalmol Vis Sci.* 2020;61(8):25. https://doi.org/10.1167/iovs.61.8.25

Purpose. The purpose of this tutorial is to provide visual scientists with various approaches for comparing two or more groups of data using parametric statistical tests, which require that the distribution of data within each group is normal (Gaussian). Non-parametric tests are used for inference when the sample data are not normally distributed or the sample is too small to assess its true distribution.

Methods. Methods are reviewed using retinal thickness, as measured by optical coherence tomography (OCT), as an example for comparing two or more group means. The following parametric statistical approaches are presented for different situations: two-sample t-test, Analysis of Variance (ANOVA), paired t-test, and the analysis of repeated measures data using a linear mixed-effects model approach.

Results. Analyzing differences between means using various approaches is demonstrated, and follow-up procedures to analyze pairwise differences between means when there are more than two comparison groups are discussed. The assumption of equal variance between groups and methods to test for equal variances are examined. Examples of repeated measures analysis for right and left eyes on subjects, across spatial segments within the same eye (e.g. quadrants of each retina), and over time are given.

Conclusions. This tutorial outlines parametric inference tests for comparing means of two or more groups and discusses how to interpret the output from statistical software packages. Critical assumptions made by the tests and ways of checking these assumptions are discussed. Efficient study designs increase the likelihood of detecting differences between groups if such differences exist. Situations commonly encountered by vision scientists involve repeated measures from the same subject over time, measurements on both right and left eyes from the same subject, and measurements from different locations within the same eye. Repeated measurements are usually correlated, and the statistical analysis needs to account for the correlation. Doing this the right way helps to ensure rigor so that the results can be repeated and validated.

Keywords: statistical methods, parametric inference, ANOVA, repeated measurements, variance components

This tutorial deals with statistical parametric tests for inference, such as comparing the means of two or more groups. Parametric tests refer to those that make assumptions about the distribution of the data, most commonly assuming that observations follow normal (Gaussian) distributions or that observations can be mathematically transformed to a normal distribution (e.g., log transformation). Non-parametric tests are used for inference when the sample data are not normally distributed or the sample is too small to assess its true distribution and will be covered in a separate tutorial.

For this tutorial on parametric statistical inference, optical coherence tomography thickness measurements of the inner retinal layers recorded in eyes of control mice and mice with optic neuritis produced by experimental autoimmune encephalitis (EAE) serve as illustration. For brevity, we refer to the measured response as retinal thickness. We have explained the goals of this study in another tutorial on the display of data,[1] and they are summarized here. There are three treatment groups: control mice, diseased mice (EAE) with optic neuritis, and treated diseased mice (EAE + treatment). For the purpose of this tutorial, we consider only mice with measurements made on both eyes. This leaves us with 15, 12, and six subjects (mice) in the three groups, respectively. For the various statistical analyses in this tutorial, the variance (s^2) is defined as the sum of the squared differences of each sample from their sample mean, which is then divided by the number of samples minus 1 (subtracting 1 corrects for the sample bias). The standard deviation is the square root of the variance. The software programs Prism 8 (GraphPad, San Diego, CA, USA) and Minitab (State College, PA, USA) were used to generate the graphs shown in this tutorial.

This tutorial analyzes the average inner retinal thickness of subjects by averaging the measurements on right and left eyes. It also analyzes the inner retinal thickness of eyes, but incorporates the correlation between right and left eye measurements on the same subject.

Analysis of the Average Retinal Thickness of Subjects After Combining Their Measurements on Right and Left Eyes

Comparing Means of Two Treatment Groups: Two-Sample *t*-Test

First we discuss whether there is a difference between the average retinal thickness of control and diseased mice after EAE-induced optic neuritis. We compare the two groups A = control and B = EAE. Measurements in these two groups are independent, as each group contains different mice. The two-sample *t*-test relates the difference of the sample means $\bar{y}_A - \bar{y}_B$ to its estimated standard error, $se_{\bar{y}_A-\bar{y}_B} = \sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}}$. Here, $n_A, \bar{y}_A, s_A$ and $n_B, \bar{y}_B, s_B$ are the sample size, mean, and standard deviation for each of the two groups.

Under the null hypothesis of no difference between the two means, the ratio $\frac{\bar{y}_A-\bar{y}_B}{\sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}}}$ is well-approximated by a *t*-distribution, with its degrees of freedom $\frac{[(s_A^2/n_A)+(s_B^2/n_B)]^2}{(s_A^4/n_A^2(n_A-1))+(s_B^4/n_B^2(n_B-1))}$ given from the Welch approximation.[2] Confidence intervals and probability values can be calculated. Small probability values (smaller than 0.05 or 0.10) indicate that the null hypothesis of no difference between the means can be rejected. Note that, although traditionally a probability of <0.05 has been considered significant, some groups favor an even more stringent criterion, but others feel that a less conservative criterion (e.g., $P < 0.1$) may still be meaningful, depending on the context of the study.

One can also use the standard error that uses the pooled standard deviation, $se_{\bar{y}_A-\bar{y}_B} = s_{pooled}\sqrt{\frac{1}{n_A}+\frac{1}{n_B}} = \sqrt{\frac{(n_A-1)s_A^2+(n_B-1)s_B^2}{n_A+n_B-2}}\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}$, and a *t*-distribution with $n_A - n_B - 2$ degrees of freedom. However, we prefer the first method, where the standard error of each group is calculated separately (not pooled), and the Welch approximation of the degrees of freedom, as it does not require that the two group variances be the same. The pooled version of the test assumes equal variances and can be misleading when they are not.[3] Both *t*-tests are robust to non-normality as long as the sample sizes are reasonably large (sample sizes of 30 or larger; robustness follows from the central limit effect).

The mean retinal thickness of the diseased mice (group B, EAE: mean = 59.81 µm; SD = 3.72 µm) is 6.40 microns smaller than that of the control group (group A, control: mean = 66.21 µm, SD = 3.39 µm). The *P* value (0.0001) shows that this difference is quite significant, leaving little doubt that the disease leads to thinning of the inner retinal layer (Table 1).

Table 1. Subject Average Retinal Thickness (in µm) for Control and Disease Groups: Two-Sample *t*-Test with Welch Correction Comparing Group A (Control) with Group B (EAE)*

Unpaired *t*-test with Welch's correction	
P value	0.0001
Significantly different ($P < 0.05$)?	Yes
One- or two-tailed *P* value?	Two-tailed
Welch-corrected *t* (degrees of freedom)	4.616 (22.62)
How big is the difference?	
Mean of column A	66.21
Mean of column B	59.81
Difference between means (B – A) ± SEM	–6.396 ± 1.385
95% confidence interval	–9.265 to –3.527

* Analysis with GraphPad Prism8.

Comparing Means of Two or More (Independent) Treatment Groups: One-Way ANOVA

The one-way analysis of variance can be used to compare two or more means. Assume that there are k groups (for our illustration, $k = 3$) with observations y_{ij} for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_i$ (number of observations in the ith group). The ANOVA table partitions the sum of squared deviations of the $n = \sum_{i=1}^{k} n_i$ observations from their overall mean, $\bar{y}$, into two components: the between-group (or treatment) sum of squares, $SSB = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$, expressing the variability of the group means $\bar{y}_i$ from the overall mean $\bar{y}$, and the within-group (or residual) sum of squares, $SSW = \sum_{i=1}^{k}\{\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2\} = \sum_{i=1}^{k}(n_i - 1)s_i^2$, adding up all within-group variances, s_i^2. The ratio of the resulting mean squares (where mean squares are obtained by dividing sums of squares by their degrees of freedom), $F = \frac{SSB/(k-1)}{SSW/(n-k)}$, serves as the statistic for testing the null hypothesis that all group means are equal. The probability value for testing this hypothesis can be obtained from the *F*-distribution. Small probability values (smaller than 0.05 or 0.10) indicate that the null hypothesis should be rejected.

The ANOVA assumes that all measurements are independent. This is the case here, as we have different subjects in the three groups. Note that independence could not be assumed if both right and left eyes were included, as right and left eye observations from the same subject are most likely correlated; we will discuss later how to handle this situation.

The ANOVA assumes that the variances of the treatment groups are the same. Its conclusions may be misleading if the variances are different. Box[3] showed that the *F*-test is sensitive to violations of the equal variance assumption, especially if the sample sizes in the groups are different. The *F*-test is less affected by unequal variances if the sample sizes are equal. Although the *F*-test assumes normality, it is robust to non-normality as long as the sample sizes are reasonably large (e.g., 30 samples per group).

For only two treatment groups, the ANOVA approach reduces to the two-sample *t*-test that uses the pooled variance. Earlier we had recommended the Welch approximation, which uses a different standard error calculation for the difference of two sample means, as it does not assume equal variances. Useful tests for the equality of variances are discussed later.

If the null hypothesis of equal group means is rejected when there are more than two treatment groups, then follow-up tests are needed to determine which of the treatment groups differ from the others using pairwise comparisons. For three groups, one calculates three pairwise (multiple) comparisons and three confidence intervals for each pairwise difference of two means. The significance level of individual pairwise tests needs to be adjusted for the number of comparisons being made. Under the null

Investigative Ophthalmology & Visual Science

Table 2. Subject Average Retinal Thickness (in µm): One-Way ANOVA with Three Groups (Control, EAE, EAE + Treatment) and Tukey's Multiple Comparison Tests*

Summary	**Control**	**EAE**	**EAE + Treatment**		
Number of values	15	12	6		
Mean	66.21	59.81	61.65		
SD	3.390	3.721	2.780		
ANOVA Table	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	***F*(DFn, DFd)**	***P***
Treatment	287.2	2	143.6	$F(2, 30) = 12.25$	0.0001
Residual	351.8	30	11.73	—	—
Total	639.0	32	—	—	—
Tukey's Multiple Comparisons Test	**Mean Difference**	**95% CI of Difference**	**Significant?**		**Adjusted *P***
Control vs. EAE	6.396	3.126–9.665	Yes	***	0.0001
Control vs. EAE + treatment	4.562	0.4847–8.640	Yes	*	0.0258
EAE vs. EAE + treatment	–1.833	–6.054 to 2.388	No	ns	0.5392

* Analysis with GraphPad Prism8. Dfn, degrees of freedom numerator; Dfd, degrees of freedom denominator.

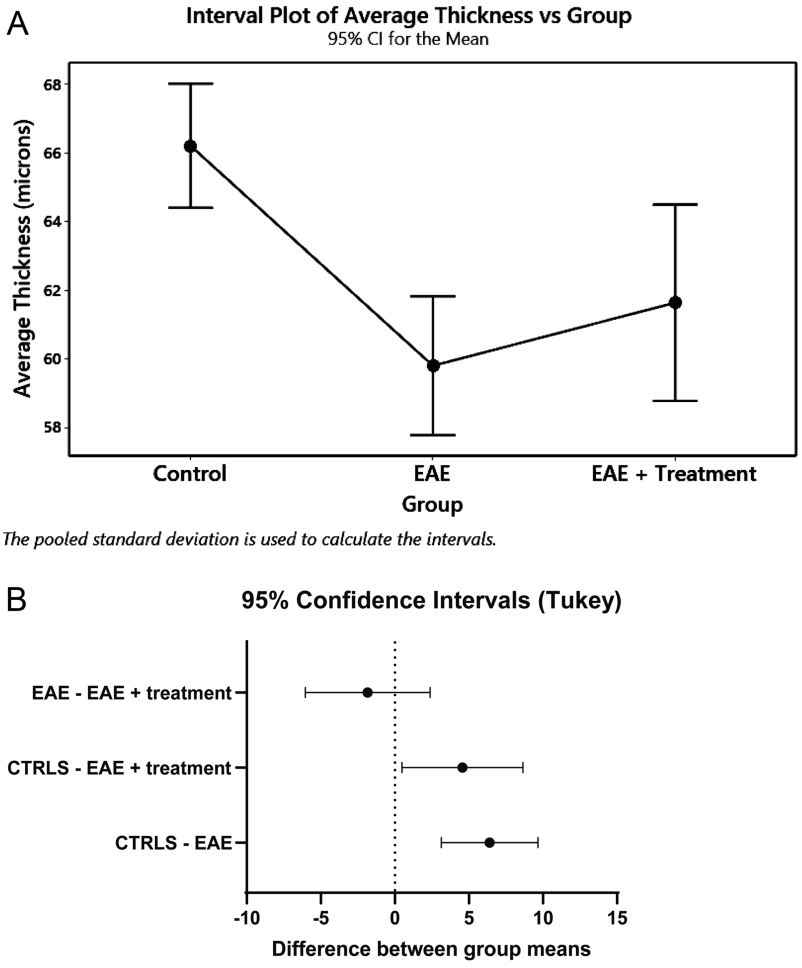

Figure 1. Subject average retinal thickness (in µm). Visualizations of results. (**A**) Plot of group means and their 95% confidence intervals. Confidence intervals are not adjusted for multiple comparisons. Analysis with Minitab. (**B**) Plot of pairwise differences and their Tukey-adjusted confidence intervals. Analysis with GraphPad Prism 8.

hypothesis of no treatment effects, we set the error that one or more of these multiple pairwise comparisons are falsely significant at a given significance level, such as $\alpha = 0.05$. To achieve this, one must lengthen individual confidence intervals and increase individual probability values. This is exactly what the Tukey multiple comparison procedure[4] does (Table 2, Fig. 1). Many other multiple comparison procedures are available (Bonferroni, Scheffe, Sidak, Holm, Dunnett, Benjamini–Hochberg), but their discussion would go beyond this introduction. For a discussion

Table 3. Subject Average Thickness (in µm): Bartlett and Brown–Forsythe Tests for Equality of Group Variances*

Test	Result
Brown–Forsythe test	
F (DFn, DFd)	0.7118 (2, 30)
P value	0.4989
P value summary	Not significant
Are SDs significantly different ($P < 0.05$)?	No
Bartlett's test	
Bartlett's statistic (corrected)	0.5149
P value	0.7730
P value summary	Not significant
Are SDs significantly different ($P < 0.05$)?	No

* Analysis with GraphPad Prism8.

of the general statistical theory of multiple comparisons, see Hsu.[5]

The ANOVA results in Table 2 show that mean retinal thickness differs significantly across the three treatment groups ($P = 0.0001$). Tukey pairwise comparisons show differences between the group means of thickness for control and EAE and for control and EAE + treatment. The means of EAE and EAE + treatment are not significantly different.

Comparing Variances of Two or More (Independent) Treatment Groups: Bartlett, Levine, and Brown–Forsythe Tests

As stated above, ANOVA testing assumes that the group variances are equal. How does one test for equal variances? Bartlett's test[6] (see Snedecor and Cochran[7]) is employed for testing if two or more samples are from populations with equal variances. Equal variances across populations are referred to as homoscedasticity or homogeneity of variances. The Bartlett test compares each group variance with the pooled variance and is sensitive to departures from normality. The tests by Levene[8] and Brown and Forsythe[9] are good alternatives that are less sensitive to departures from normality. These tests make use of the results of a one-way ANOVA on the absolute value of the difference between measurements and their respective group mean (Levine test) or their group median (for the Brown–Forsythe test).

We apply these tests to the average retinal thickness data. We cannot reject the hypothesis that all three variances are the same, so we can be more confident in our interpretation of the ANOVA results, as the variances of the groups appear to be similar (Table 3). If one of the tests shows unequal variance but the other test does not, then one needs to evaluate how significant the *P* value was in rejecting the null hypothesis of equal variance. If a fair amount of uncertainty remains, then alternative approaches are discussed in the next section.

Approaches to Take When Variances Are Different

A finding of unequal variances is not just a nuisance (because it puts into question the results from the ANOVA on means) but it also provides an opportunity to learn something more about the data. Discovering that particular groups have different variances gives valuable insights.

Transforming measurements usually helps to satisfy the requirement that variances are equal. Box and Cox[10] discussed transformations that stabilize the variability so that the variances in the groups are the same. A logarithmic transformation is indicated when the standard deviation in a group is proportional to the group mean; a square root transformation is indicated when the variance is proportional to the mean. Reciprocal transformations are useful if one studies the time from the onset of a disease (or of a treatment) to a certain failure event such as death or blindness. The reciprocal of time to death, which expresses the rate of dying, often stabilizes group variances. For details, see Box et al.[11]

If one cannot find a variance-stabilizing transformation, one can proceed with the Welch approximation of pairwise two-sample comparisons. For nearly equal and moderately large sample sizes, the assumption of equal standard deviations is not a crucial assumption, and moderate violations of equal variances can be ignored. Another alternative would be to use nonparametric procedures (they are covered in a different tutorial).

Analysis of Retinal Thickness Using Both Right and Left Eye Measurements of Each Subject

Comparing Means of Two Repeated Measurements: Paired *t*-Test

In the earlier two-sample comparison, different subjects were assigned to each of two treatment groups. Often it is more efficient to design the experiment such that a treatment (or induction of a disease phenotype, as in this example) is applied to the same subject. For our example, each mouse could be observed both under its initial healthy condition and after having been exposed to a multiple sclerosis phenotype EAE protocol. Measurements are then available on the same mouse under both conditions, and one can control for (remove) the subject effect that exists. A within-subject comparison of the effectiveness of a treatment or drug is subject to fewer interfering variables than a comparison across subjects. The same is true for the comparison of right and left eyes when both measurements come from the same subject and only one eye is treated, with the other eye acting as a within-subject control. The large subject effect that affects both eyes in a similar way can be removed, resulting in an increase of the precision of the comparison, potentially making it more sensitive to detecting an effect, if one exists.

The paired *t*-test considers treatment differences, d, on n different subjects and compares the sample mean ($\bar{d}$) to its standard error, $se_{\bar{d}} = s_d/\sqrt{n}$. Under the null hypothesis of no difference, the ratio (test statistic) $\bar{d}/se_{\bar{d}}$ has a *t*-distribution with $n - 1$ degrees of freedom, and confidence intervals and probability values can be calculated. Small probability values (usually smaller than 0.05 or 0.10) would indicate that the null hypothesis should be rejected.

For illustration, we use the right eye (OD) and left eye (OS) retinal thickness measurements from the 15 mice of the control group. Figure 2 demonstrates considerable between-subject variability; the intercepts of the lines that connect measurements from the same subject differ considerably. Pairing the observations and working with changes on the same subject removes the subject variability and makes the analysis more precise. Table 4 indicates that there is no difference in the average retinal thickness of right and left eyes. We had expected this result, as neither eye was treated. However, if one wanted to test a treatment that is given to

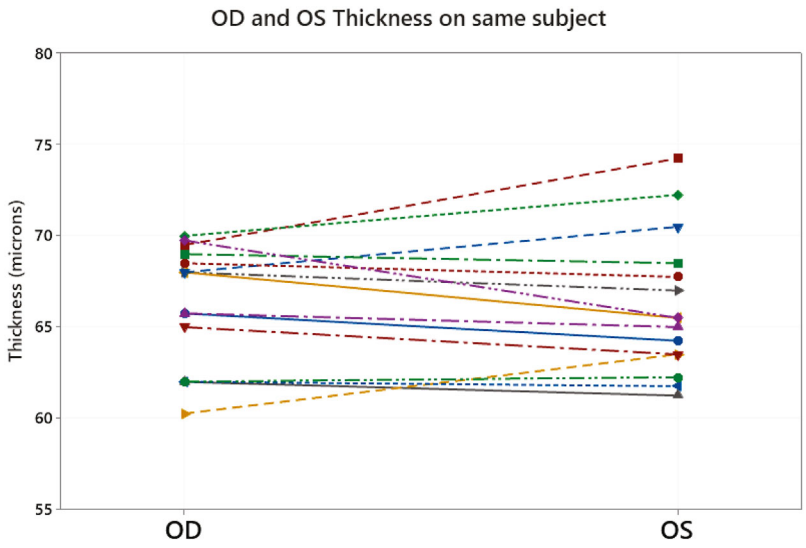

Figure 2. Retinal thickness (in µm) of OD and OS eyes in the control group (15 mice). OD and OS measurements from the same subject are connected. Analysis with Minitab.

Table 4. Retinal Thickness (in µm) of OD and OS Eyes in the Control Group (15 Mice): Paired *t*-Test*

Variable	Value
Estimation for paired difference = OD – OS	
Mean	0.050
SD	2.334
SEM	0.603
95% CI for mean difference	–1.243 to 1.343
Test H_0: $\mu_{Diff} = 0$ vs. H_1: $\mu_{Diff} \neq 0$	
Test statistic	0.08
P value	0.935

* Analysis with GraphPad Prism8.

just one eye without affecting the other, such a paired treatment comparison between the two eyes would be a desirable analysis plan.

Correlation Between Repeated Measurements on the Same Subject

The two-sample *t*-test in Table 1 and the ANOVA in Table 2 used subject averages of the thickness of the right and left eyes. Switching to eyes as the unit of observation, it is tempting to run the same tests with twice the number of observations in each group, as now each subject provides two observations. But, if eyes on the same subject are correlated (in our illustration with 33 subjects, the correlation between OD and OS retinal thickness is very large: $r = 0.90$), this amounts to "cheating," as correlated observations carry less information than independent ones. By artificially inflating the number of observations and inappropriately reducing standard errors, the probability values appear more significant than they actually are.

Suppose that measurements on the right and left eye are perfectly correlated. Adding perfect replicates does not change the group means and the standard deviations that we obtained from the analysis of subject averages; however, with perfect replicates, the earlier standard error of the difference of the two group means gets divided by $\sqrt{2}$, which increases the test statistic and makes the difference appear more significant than it actually is. The earlier ANOVA is equally affected. Adding replicates increases the between-group mean square by a factor of 2 but does not affect the within-group mean square, thus increasing the *F*-test statistic. This shows that a strategy of adding more and more perfect replicates to each observation makes even the smallest difference significant. One cannot ignore the correlation among measurements on the same subject! The following two sections show how this correlation can be incorporated into the analysis.

Analysis of Repeated Measures Data

Many studies involve repeated measurements on each subject. Here we have 15 healthy control mice, 12 diseased mice (EAE), and 6 treated diseased mice (EAE + treatment), and we have repeated measurements on each subject: measurements on the left and right eye. But, repeated measurements may also reflect measurements over time or across spatial segments (e.g., quadrants of each retina). The objective is to study the effects of the two factors, treatment and eye. Repeated measurements on the same subject can be expected to be dependent, as a subject that measures high on one eye tends to also measure high on the other. The correlation must be incorporated into the analysis. This makes the analysis different from that of a completely randomized two-factor experiment where all observations are assumed independent.

The model for data from such a repeated measures experiment represents the observation Y_{ijk} on subject i in treatment group j and eye k according to

$$Y_{ijk} = \alpha + \beta_j + \pi_{i(j)} + \gamma_k + \beta\gamma_{jk} + \varepsilon_{i(j)k}$$

where

- α is an intercept.
- β_j in this example represents (three) fixed differential treatment effects, with $\beta_1 + \beta_2 + \beta_3 = 0$. With this restriction, treatment effects are expressed as devia-

Table 5. Retinal Thickness (in µm) of OD and OS Eyes*

Two-Way Repeated Measures ANOVA

Matching	Stacked				
Assume sphericity?	Yes				
Alpha	0.05				
ANOVA Table	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	***F*(DFn, DFd)**	***P***
Between subject					
Treatment	574.5	2	287.2	$F(2, 30) = 12.25$	**0.0001**
Subject	703.5	30	**23.45**	$F(30, 30) = 10.70$	**<0.0001**
Within subject					
Eye	1.351	1	1.351	$F(1, 30) = 0.6164$	0.4385
Eye × treatment	0.5959	2	0.298	$F(2, 30) = 0.1360$	0.8734
Residual	65.75	30	**2.192**	—	—
Total	1345.31	65	—	—	—

Shown is the GraphPad Prism8 ANOVA output of the two-factor repeated measures experiment with three treatment groups and the repeated factor eye. Sphericity assumes that variances of differences between all possible pairs of within-subject conditions are equal.

tions from the average. An equivalent representation sets one of the three coefficients equal to zero, then the parameter of each included group represents the difference between the averages of the included group and the reference group for which the parameter has been omitted.

- $\pi_{i(j)}$ represents random subject effects, represented by a normal distribution with mean 0 and variance σ_π^2. The subscript notation $i(j)$ expresses the fact that subject i is nested within factor j; that is, subject 1 in treatment group 1 is a different subject than subject 1 in treatment group 2. Each subject is observed under only a single treatment group. This is different from the "crossed" design where each subject is studied under all treatment groups.
- γ_k represents fixed eye (OD, OS) effects with coefficients adding to zero: $\gamma_1 + \gamma_2 = 0$.
- $\beta\gamma_{jk}$ represents the interaction effects between the two fixed effects, treatment and eye, with row and column sums of the array $\beta\gamma_{jk}$ restricted to zero. There is no interaction when all $\beta\gamma_{jk}$ are zero; this makes effects easier to interpret, as the effects of one factor do not depend on the level of the other.
- $\varepsilon_{i(j)k}$ represents random measurement errors, with a normal distribution, mean = 0, and variance = σ_ε^2. Measurement errors reflect the eye by subject (within treatment) interaction.

This model is known as a linear mixed-effects model as it involves fixed effects (here, treatment and eye and their interaction) and random effects (here, the subject effects and the measurement errors). Maximum likelihood or, preferably, restricted maximum likelihood methods are commonly used to obtain estimates of the fixed effects and the variances of the random effects; standard errors of the fixed effects can be calculated, as well. For detailed discussion, see Diggle et al.[12] and McCulloch et al.[13]

Computer software for analyzing the data from such repeated measurement design is readily available. Minitab, SAS (SAS Institute, Cary, NC, USA), R (The R Foundation for Statistical Computing, Vienna, Austria), and GraphPad Prism all have tools for fitting the appropriate models. An important feature of these software packages is that they can handle missing data. It would be quite unusual if a study would not have any missing observations, and software that can handle only balanced datasets would be of little use. Without missing data (as is the case here), the computer output includes the repeated measures ANOVA table. The output from the mixed-effects analysis (which is used if observations are missing) is similar. Computer software also allows for very general correlation structure among repeated measures. The random subject representation discussed here implies compound symmetry with equal correlations among all repeated measures. With time as the repeated factor, other useful models include conditional autoregressive specifications that model the correlation of repeated measurements as a geometrically decreasing function of their time.

Results of the two-way repeated measures ANOVA for the thickness data are shown in Table 5. Estimates of the two error variances come into play differently when testing fixed effects. The variability between subjects is used when testing the treatment effect; the measurement (residual) variability is used in all tests that involve within-subject factors. See, for example, Winer.[14] These variabilities are estimated by the two mean square (MS) errors that are shown in Table 5 with bold-face type.

In Table 5, MS(Subject) = 23.45 is used to test the effect of treatment: $F(\text{Treatment}) = 287.2/23.45 = 12.25$. The treatment effect is significant at $P = 0.0001$. MS(Residual) = 2.192 is used in the test for subject effects and in tests of the main effect of eye and the eye × treatment interaction: $F(\text{Subject}) = 23.45/2.192 = 10.70$ (significant; $P < 0.0001$); $F(\text{Eye}) = 1.351/2.192 = 0.6164$ (not significant, $P = 0.4385$) and $F(\text{Eye} \times \text{Treatment}) = 0.298/2.192 = 0.1360$ (not significant, $P = 0.8734$). In summary, the mean retinal thickness differs among the control, EAE, and EAE + treatment groups. Thickness varies widely among subjects, but difference in means between right and left eyes are not significant.

Assume that we ignore the correlation of repeated measures on the same subject and run a one-way ANOVA (with our three treatment groups) on individual eye measurements. The mean square error in that analysis is $(1345.31 - 574.5)/(65 - 2) = 12.23$, increasing the *F*-statistic to $F = 287.2/12.23 = 23.47$ which is highly significant. However, such incorrect analysis that does not account for the high correlation between measurements on right and left eyes leads to wrong probability values and wrong conclusions. It makes the treatment effect appear even more significant than it really is. In this example, the conclusions about the factors are not changed, but that is not true in general for all cases.

Table 6. Retinal Thickness (in μm) of OD and OS Eyes: MINITAB Output of the Repeated Measures Experiment with Three Factors: Treatment Group, Eye, and Quadrant*

	OS				OD			
Design	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**
Control	G1	G1	G1	G1	G1	G1	G1	G1
EAE	G2	G2	G2	G2	G2	G2	G2	G2
EAE + treatment	G3	G3	G3	G3	G3	G3	G3	G3
ANOVA Table	**Degrees of Freedom**	**Adjusted Sum of Squares**	**Adjusted Mean Square**	***F***	***P***			
Between subject								
Treatment	2	2297.86	1148.93	12.25	0.000			
Subject	30	2814.15	**93.80**	7.41	0.000			
Within subject								
Eye	1	5.40	5.40	0.43	0.514			
Treatment × eye	2	2.38	1.19	0.09	0.910			
Quadrant	3	36.97	12.32	0.97	0.406			
Treatment × quadrant	6	68.45	11.41	0.90	0.495			
Eye × quadrant	3	17.51	5.84	0.46	0.710			
Treatment × eye × quadrant	6	92.61	15.43	1.22	0.298			
Residual	210	2659.15	**12.66**	—	—			
Total	263	8001.26	—	—	—			

The residual sum of squares pools the interaction sums of squares between subjects and the effects of eye, quadrant, and eye by quadrant interaction. The three-factor ANOVA in GraphPad Prism8 is quite limited (two of the three factors can only have two levels) and could not be used.

A standard two-way ANOVA on treatment (with three levels) and eye (with two levels) that does not account for repeated measurements also leads to incorrect results, as such analysis assumes that observations in the six groups are independent. This is not so, as observations in different groups come from the same subject.

More Complicated Repeated Measures Designs

Extensions of repeated measures designs are certainly possible. Here are two different illustrations for a potential third factor.

In the first model, the third factor is the (spatial) quadrant of the retina in which the measurement is taken. Measurements on the superior, inferior, nasal, and temporal quadrants are taken on each eye. The model includes random subject effects for the different mice in each of the three treatment groups (G1, G2, G3), with each mouse studied under all eight eye/quadrant combinations. The design layout is shown in Table 6.

Data for the 15, 12, and six mice from the three treatment groups are analyzed. A total of 33 subjects × 8 regions (four quadrants for the right eye and four for the left eye) = 264 measurements is used to estimate this repeated measures model. Results are shown in Table 6. MS(Subject) = 93.80 is used for testing the treatment effect, F(Treatment) = 1148.93/93.80 = 12.25. MS(Residual) = 12.66 is used in all other tests (subject effects, main and interaction effects of eye and quadrant, and all of their interactions with treatment). Treatment and subject effects are highly significant, but all effects of eye and quadrant are insignificant, meaning that eyes and quadrants had no effect on retinal thickness.

In the second model, a third factor, type, represents two different genetic mouse strains. The experiment studies the effect of treatment on mice from either of two genetic strains (type 1 and type 2 below). Treatment and strain are crossed fixed effects, as every level of one factor is combined with every level of the other. Each mouse taken from one of the six groups has a measurement made at four different quadrants in one eye. This is a different repeated measures design, as now the mice are nested within the treatment–strain combinations. The design looks as follows:

	Quadrant			
	Q1	**Q2**	**Q3**	**Q4**
Control type 1	G1	G1	G1	G1
Control type 2	G2	G2	G2	G2
EAE type 1	G3	G3	G3	G3
EAE type 2	G4	G4	G4	G4
EAE + treatment type 1	G5	G5	G5	G5
EAE + treatment type 2	G6	G6	G6	G6

The variability between subjects is used for testing main and interaction effects of treatment and strain. The measurement (residual) variability is used in the test for subject effects and the tests for the main effect of quadrant and its interactions with treatment and strain.

Conclusions

This tutorial outlines parametric inference tests for comparing means of two or more groups and how to interpret the output from statistical software packages. Critical assumptions made by the tests and ways of checking these assumptions are discussed.

Efficient study designs increase the likelihood of detecting differences among groups if such differences exist. Situations commonly encountered by vision scientists involve repeated measures from the same subject over time, on both right and left eyes from the same subject, and from different locations within the same eye. Repeated measures are usually correlated, and the statistical analysis must account

for the correlation. Doing this the right way helps to ensure rigor so that the results can be repeated and validated with time. The data used in this review (in both Excel and Prism 8 format) are available in the Supplementary Materials.

Two Excel data files can be found under the Supplementary Materials: Supplementary Data S1 contains measurements on each eye as well as on each subject, whereas Supplementary Data S2 contains measurements for each quadrant of the retina. The two GraphPad Prism8 files under the Supplementary Materials illustrate the data analysis: Supplementary Material S3 on the analysis of subject averages, and Supplementary Material S4 on the analysis of individual eyes.

Acknowledgments

Supported by a VA merit grant (C2978-R); by the Center for the Prevention and Treatment of Visual Loss, Iowa City VA Health Care Center (RR&D C9251-C, RX003002); and by an endowment from the Pomerantz Family Chair in Ophthalmology (RHK).

Disclosure: **J. Ledolter**, None; **O.W. Gramlich**, None; **R.H. Kardon**, None

References

1. Ledolter J, Gramlich OW, Kardon RH. Focus on data: display of data. *Invest Ophthalmol Vis Sci*. 2020;0:30023, https://doi.org/10.1167/iovs.0.0.30023.
2. Welch BL. The generalization of "Student's" problem when several different population variances are involved. *Biometrika*. 1947;34:28–35.
3. Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problem. I. Effect of inequality of variance in one-way classification. *Ann Math Stat*. 1954;25:290–302.
4. Tukey JW. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley; 1977.
5. Hsu JC. *Multiple Comparisons: Theory and Methods*. London: Chapman & Hall, 1996.
6. Bartlett MS. Properties of sufficiency and statistical tests. *Proc R Soc Lond A Math Phys Sci*. 1937;160:268–282.
7. Snedecor GW, Cochran WG. *Statistical Methods*. 8th ed. Ames, IA: Iowa State University Press; 1989.
8. Levene H. Robust tests for equality of variances. In: Olkin I, Ghurye SG, Hoeffding W, Madow WG, Mann HB, eds. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Palo Alto, CA: Stanford University Press; 1960:278–292.
9. Brown MB, Forsythe AB. Robust tests for the equality of variances. *J Am Stat Assoc*. 1974;69:364–367.
10. Box GEP, Cox DR. An analysis of transformations. *J R Stat Soc B Methodol*. 1964;26:211–243.
11. Box GEP, Hunter S, Hunter WG. *Statistics for Experimenters: Design, Innovation, and Discovery*. 2nd ed. New York: John Wiley & Sons; 2005.
12. Diggle PJ, Liang KY, Zeger SL. *Analysis of Longitudinal Data*. New York: Oxford University Press; 1994.
13. McCulloch CE, Searle SR, Neuhaus JM. *Generalized, Linear, and Mixed Models*. 2nd ed. New York: John Wiley & Sons; 2008.
14. Winer BJ. *Statistical Principles of Experimental Design*. 2nd ed. New York: McGraw-Hill; 1999.