Biometrics And Biostatistics Journal

SCIENTIFIC LITERATURE

Special Issue Article "Longitudinal Data Analysis" Review Article

A Primer on Statistics for Researchers Conducting Laboratory Studies

Johannes Ledolter*

Department of Business Analytics, Tippie College of Business at the University of Iowa, and Iowa City VA Center for the Prevention and Treatment of Visual Loss, Iowa City, IA USA

ARTICLE INFO

Received Date: February 03, 2020
Accepted Date: March 01, 2020
Published Date: March 06, 2020

KEYWORDS

Effective visual display of data; Parametric and nonparametric statistical techniques; Tests of normality; Analysis of data from repeated measurement designs; Principles for effective experimental designs; Sample size selection; Meta analysis

ACKNOWLEDGMENT

This research was supported through grant C9251-C from the US Department of Veterans Affairs Office of Rehabilitation Research & Development.

Citation for this article: Johannes Ledolter. A Primer on Statistics for Researchers Conducting Laboratory Studies. Biometrics And Biostatistics Journal. 2020; 2(1):113

Corresponding author:
Johannes Ledolter,
Department of Business Analytics, Tippie College of Business, University of Iowa Center for the Prevention and Treatment of Visual Loss, Iowa City VA Health Care Iowa City, USA,
Email: johannes-ledolter@uiowa.edu

ABSTRACT

Scientific learning is advanced when researchers test their theories empirically. Results of empirical studies will either confirm or refute the theory, but also suggest modifications to the existing theory. All empirical studies need to be carefully and efficiently planned, and the resulting data must be analyzed with appropriate methods. This paper provides a primer on the design and the statistical analysis of laboratory studies that should be useful to scientists carrying out such work.

Several topics are discussed in detail: the display of data; common parametric and nonparametric statistical methods; techniques for checking normality; the appropriate analysis of data from repeated measurement designs; important principles of effective experimental designs and the selection of the sample size; the role of statistical significance and why the size of the estimated effect matters; meta-analysis for combining the results of multiple studies; and the implementation of statistical techniques in commonly used computer packages.

INTRODUCTION

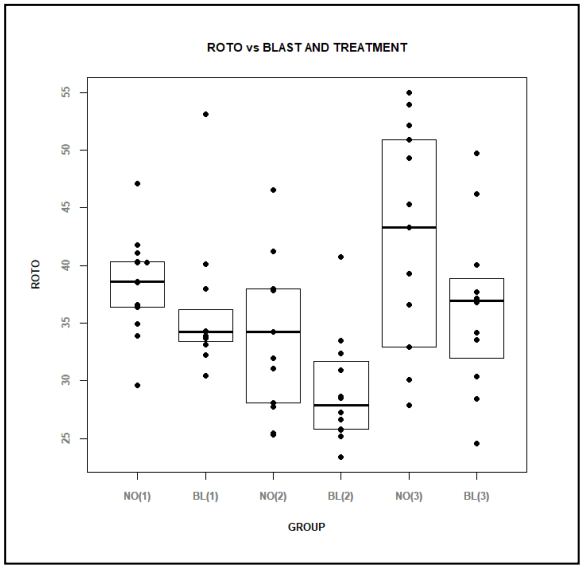
I have been the consulting statistician at the Center for the Prevention and Treatment of Visual Loss at the Iowa City VA Health Care System. Over the last ten years I have been an investigator on more than 30 grants and I have contributed to numerous publications. Many of our projects deal with laboratory studies. Similar questions about the appropriate design of experiments and the correct analysis of the resulting data turn up in every project. Since the same questions are raised repeatedly, I have circulated this brief tutorial to our research team of about 20 investigators to help them avoid the most common mistakes. I believe that this tutorial can also be helpful to the many research groups that work on similar biology and medical science projects. Section 2 of this tutorial discusses how to best display the data. Section 3 discusses common statistical methods, both parametric methods that assume normality of the observations as well as nonparametric methods that relax the assumption of normality. Various methods for checking normality are reviewed. Most of our data comes from repeated measurement designs. There experimental units are assigned to several groups at random, and repeated measurements are taken at various points in time. Chapter 4 discusses how the data from such experiments should be analyzed. Principles of effective experimental designs and a detailed discussion on how to obtain the appropriate sample size that allows investigators to detect scientifically meaningful effects are given in Section 5. Section 6 discusses the role of statistical

significance. Statistical significance is not all that matters; the size of the estimated effect is also critical. Meta-analysis, a statistical procedure for combining the results of multiple studies, is reviewed in Section 7. Individual study results are measured with error. The aim of a meta-analysis is to derive a pooled estimate that is closest to the unknown common truth. Both fixed-effects and random-effects implementations of meta-analysis are reviewed. The primer also explains how the various tools are implemented in the most common statistical software packages such as Minitab, SAS, R Statistical Software, and Graph Pad PRISM.

EFFECTIVE VISUAL DISPLAY OF STUDY RESULTS

Show all observations

For small and moderately-sized studies, our recommendation is to show all individual observations. One can add to the graph of individual observations summary statistics such as the median and draw a box around the first and third quartiles. The plots in (Figure 1) draw attention to obvious outliers, which need to be scrutinized, and also to the shape of the distribution. We do not recommend to visualize the data with a bar chart that only shows the average and its standard error. The standard error of a sample average (calculated as the standard deviation of individual observations divided by the square root of the sample size) reflects the reliability of the sample mean as an estimate of the mean of the population from which the random sample is selected. Such a graph does not visualize the raw data.



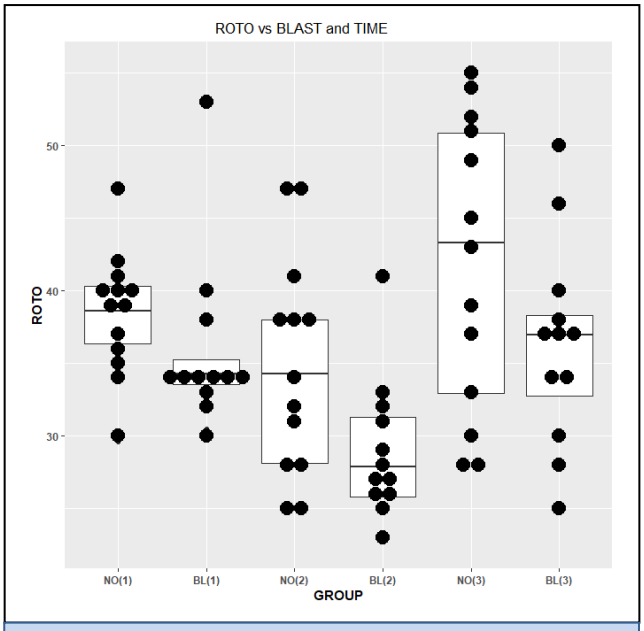


Figure 1: Two displays of the data in our example.

Treatment of outliers

Outliers can be safely omitted if there is clear evidence that something went wrong with a particular measurement or a particular experiment, and if one also knows why this has happened. However, in the absence of any evidence why such an outlier has occurred, the observation needs to be included in the analysis. Conducting two analyses – one with and the other without the questionable measurement – can tell us about the influence of a suspect observation on the conclusion. If the suspect observation does not have an influence on the conclusion, even better – because then there is no issue. But if an observation does and is hugely influential in reaching a certain finding, one should be careful about one's conclusion.

Example

We consider data from an experiment with mice from two different treatment groups. The data was provided by Dr. Matthew Harper from the Iowa City VA Center for the Prevention and Treatment of Visual Loss. One group consists of 13 control mice; the other group represents an experimental group of 12 mice exposed to a blast that is suspected to cause a traumatic brain injury and a deficit in functional performance. The rotarod performance of each mouse is measured at three consecutive time periods (1 = pre blast, 2 = 7 days after blast; 3 = 30 days after blast). The rotarod test is a performance test based on a rotating rod with forced motor

activity being applied. The test measures parameters such as riding time (seconds) or endurance and is typically used to test the effect of experimental drugs or after traumatic brain injury. The data in this example come from a repeated measurement design as the three observations are taken on the very same mouse. Mice are certainly not alike and their responses differ; later on, the mouse effect will be modeled as an additive random effect with its variance expressing the magnitude of that effect. The treatment groups (control and blast) contain different mice; in this design, mice are nested within the treatment group. We will have to say more on how such data should be analyzed. The objective of this particular study is to assess whether there is a treatment effect. That is, is the blast group different from the control group, and does the blast effect change with the time since exposure?

The two graphs shown below (obtained by two different software programs; all data and code for the analysis are available on my website) represent every single measurement that has been taken. The dot plots for each group are displayed vertically, with identical values stacked horizontally (and not "over" plotted as then one would not know how many measurements are actually the same). Box plots summarize the data by showing the first, second (median), and third quartiles of each group. Groups are arranged in an order that facilitates the interpretation. The comparison of the first and second group assesses the blast effect at time 1 (now there shouldn't be any blast effect as time 1 corresponds to pre-blast measurements); the comparison of the third and fourth group assesses the blast effect at time 2 (7 days after blast); and the comparison of the fifth and sixth group assesses the blast effect at time 3 (30 days after blast). The graph shows that blast tends to reduce the rotarod measurements. We can also assess the effect of time: The comparison of groups 1, 3 and 5 assesses the time progression of control mice; we should not see large changes over time for the control mice; any changes seen could be due to day-to-day changes in the experimental set-up and perhaps a learning effect. The comparison of groups 2, 4, and 6 assesses the time progression for blasted mice. For blasted mice, rotarod measurements are smallest at time 2 (7 days after the blast), whereas the measurements at time 3 (30 days after the blast) are back to their pre-blast (time 1) level. For control mice, it is difficult to tell whether the three time groups are different. The dip at time 2, for both control and blasted mice, may be related to changes in the lab conditions at time 2. Of course, one should not over-interpret the data without knowing whether the effects of blast and time and their interaction are statistically significant. And for that we need to take into account how the experiment was carried out. The discussion of the analysis of repeated measurement designs is shown in a later section (Section 4).

STATISTICAL METHODS

There is a huge variety of statistical methods; the investigator needs to specify the question of interest and also must check whether the assumptions made by each statistical method are satisfied. Conclusions reached from statistical methods when assumptions are violated should not be trusted. Statistical methods can be divided into parametric and nonparametric methods. Common parametric methods, mostly related to the linear model, assume normality of the errors. Parametric methods include the typical t-tests (one-sample t-test for testing a hypothesis about the population mean; two sample t-tests for comparing two population means from independent samples, with and without assumptions on the equality of the two population variances; paired t-test for testing a difference of two population means from paired dependent samples), ANOVA methods for testing effects in models of different complexities, Bartlett tests for comparing the equality of variances, and correlation and regression methods.

Normality should be checked before using statistical methods that assume normality

Normality should be checked, both visually and numerically. Visually, with a *q-q* plot that displays observed values against normally distributed data. Numerically, through one of the numerous significance tests for normality, including the Anderson-Darling normality test, the Shapiro-Francia normality test, the Lilliefors (Kolmogorov-Smirnov) normality test, the Cramer-von Mises normality test, the Pearson chi-square normality test, the Shapiro-Wilk's test for normality, the Jarque-Bera normality test, and the D'Agostino normality test. As one would expect, graphical methods are typically not very useful when the sample size is small. Also, normal probability tests are not that powerful for small samples. Furthermore, the results of the different normality tests are not always the same as not every test for normality is equally sensitive to one or the

other violations of normality – while there is only one normality, there are certainly many different ways of violating normality. For an evaluation of normal probability tests; see Yap BW and Sim CH [1].

The i^{th} largest observation in a sample of size n, $x_{(i)}$, is the observed sample quantile of order $P_i = (i-0.5)/n$. One can calculate the theoretical quantiles of order P_i for the $N(\mu, \sigma^2)$ distribution. This normal quantile is given by $q_i = \mu + \sigma z_i$ where $z_i = \Phi^{-1}(P_i)$ is the quantile of the $N(0,1)$ distribution; $\Phi(z)$ is the S-shaped cumulative distribution function of the standardized normal distribution. The quantiles $z_1, z_2, \ldots, z_n$, for i = 1, 2,..., n,are called the *standardized normal scores* associated with the n ordered observations $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. If the data do in fact come from the $N(\mu, \sigma^2)$ distribution, then $x_{(i)} \cong q_i$ and the points on the scatter plot of the observed quantiles $x_{(i)}$ against the theoretical normal quantiles q_i should about fall on a 45-degree line. Such a plot is called a *quantile-quantile* (or *q-q*) plot, as one plots the quantile of one distribution (empirical) against the corresponding quantile of another (theoretical).

The normal quantiles, q_i, depend on the parameters μ and σ. In practice, we can replace these parameters by their estimates $\bar{x}$ and s. Alternatively, we can plot the observed quantiles $x_{(i)}$ directly against the standardized normal scores z_i. Since $q_i = \bar{x} + sz_i$ and since, under normality, $x_{(i)} \cong q_i$, we find that the points on the scatter plot of $x_{(i)}$ against z_i should lie on a straight line with slope s that goes through the point $(0, \bar{x})$. Deviations from the linear pattern provide evidence that the underlying distribution is not normal. It is also easy to estimate from this plot, at least approximately, the parameters μ and σ of the normal distribution. The slope in the *q-q* plot gives us an estimate of the standard deviation; the value at the intersection of the vertical line at $z_i = 0$ with the straight line through the data provides an estimate of μ.A *q-q* plot is effective because the human eye is quite good at recognizing *linear* tendencies.

(Figure 2) illustrates normal *q-q* plots for the data from our illustrative example. Normality must be checked separately for each of the six groups, as the groups have different means and variances. We find that normal distributions are appropriate for several of these groups; however for some groups, such as the blast group at time 1, an assumption of normality is not reasonable. One could have guessed so from the dot plot of the observations.

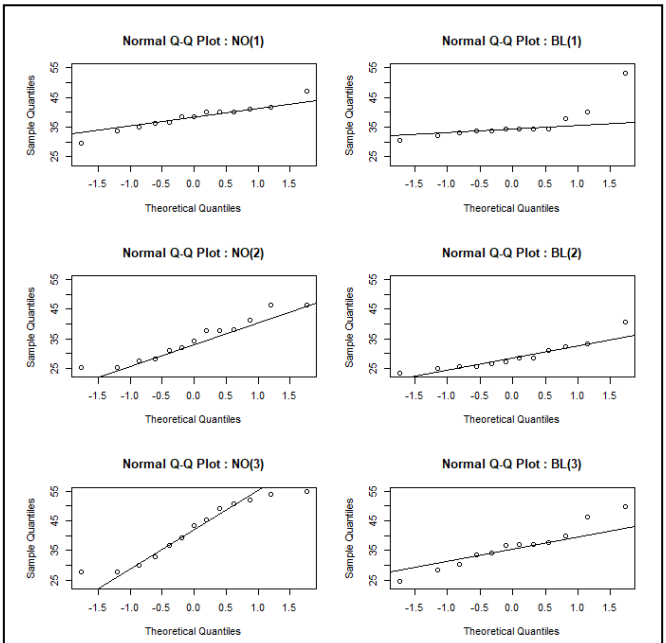


Figure 2: Normal probability plots.

Table 1: Results of the various normal probability tests.

	P-value for NO(1)	P-value for BL(1)
Anderson-Darling normality test	0.5351	0.0002
Shapiro-Francia normality test	0.4835	0.0005
Lilliefors (Kolmogorov-Smirnov)	0.6714	0.0001
Cramer-von Mises normality test	0.5220	0.0003
Pearson chi-square normality test	0.1718	0.0074
Shapiro-Wilk's test for normality	0.8409	0.0006
Jarque--Bera test for normality	0.9625	0.0002
D'Agostino normality test	Sample size < 20	Sample size < 20

We also apply the various normal probability tests mentioned previously to the data for the no-blast and the blast group at time 1 [the NO (1) and the BL (1) groups]. The results are shown in (Table 1). (Figure 2) suggests that the data for NO (1) is normal, while normality should be rejected for BL (1). The null hypothesis in all these tests represents normality; a small probability value in the output of these tests indicates that

normality can be rejected. This is exactly what we see in Table 1: Normality for the data in NO (1), and non-normality for the data in BL (1).

Parametric models can be used if the errors are normally distributed. But what should be done if distributions are non-normal?

Transformable non-normality

Certain aspects of non-normality can be overcome with transformations of the response variable. Box and Cox [2] discuss why and when transformations such as the logarithm, the square root, and the reciprocal can transform a non-normal variable into a normal one. A logarithmic transformation is indicated if the standard deviation is proportional to the level; a square root transformation is indicated if the variance is proportional to the level. Reciprocal transformations are useful if one studies the time from the onset of a disease (or of a treatment) to a certain event such as death. Distributions for time to death tend to be skewed to the right. The distribution of the reciprocal of the time to death, which expresses the rate of dying, can often be better approximated with a normal distribution. The analyst should explore transformations of the data (of the response as well as of the explanatory variables) and check whether histograms and normal-probability plots of the transformed data look (more) normal than those of the original data. For such transformable non-normality a parametric analysis can be applied to the appropriately transformed measurements. But if no reasonable transformation to normality can be found, non-parametric procedures which do not assume normality should be used.

Non-parametric procedures

Nonparametric equivalents are available for most parametric models. The one-sample Wilcoxon signed-rank test is a non-parametric alternative to the one-sample t-test when the data cannot be assumed to be normally distributed. It is used to determine whether the median of the sample is equal to a known standard value. The Wilcoxon signed-rank test is used to compare two paired (matched) samples to assess whether their population mean ranks differ. It can be used as an alternative to the paired Student t-test. The Sign test tests whether matched pair samples are drawn from distributions with equal medians. The Mann–Whitney U-test (also referred to as the Wilcoxon rank sum test) tests whether two *independent* samples selected from populations are having the same distribution. Unlike the two-sample t-test for comparing two means, this test does not require the assumption of normal distributions. It is nearly as efficient as the *t*-test on normal distributions. The Mood median test tests whether two samples are drawn from distributions with equal medians. The Kruskal–Wallis one-way analysis of variance by ranks tests whether 2 or more independent samples are drawn from the same distribution. The Friedman two-way analysis of variance by ranks tests whether k treatments in randomized block designs have identical effects. The squared ranks test is used to test the equality of variances in two or more samples. The Spearman's rank correlation coefficient measures statistical dependence between two variables using a monotonic function.

ANALYSIS OF DATA FROM REPEATED MEASUREMENT DESIGNS

Most studies, such as the one discussed at the beginning of this paper, involve repeated measurements on the same subject (in this case, the subject is a mouse); for such designs repeated measurements on the same subject can be expected to be dependent. In our illustration, the repeated measurements experiment includes several mice in each of two blast groups (control and blast), and measurements on each mouse are taken at three different times. The observation Y_{ijk} on subject *i*, in blast group *j*, and at time *k* is represented by the model

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

where

- α is an intercept
- β_j are(two) fixed differential blast effects, with $\beta_1 + \beta_2 = 0$. With this restriction, blast effects are expressed as deviations from the average. An equivalent representation sets one of the two coefficients equal to zero; then the parameter for the included group represents the difference between the levels of the included group and the reference group for which the parameter has been omitted.
- $\pi_{i(j)}$ are random subject (mouse) effects, represented by a normal distribution with mean 0 and variance σ_π^2. The subscript notation $i(j)$ expresses that the subject *i* is nested within factor *j*, as each subject is observed under only a single

treatment group. This is different from the crossed design where each subject is studied under both treatment groups.

- γ_k represent fixed time effects with coefficients adding to zero, $\gamma_1 + \gamma_2 + \gamma_3 = 0$.
- $\beta\gamma_{jk}$ represent the interaction effects between the two fixed effects, blast and time, with row and column sums of the array $\beta\gamma_{jk}$ restricted to zero.
- $\varepsilon_{i(j)k}$ are random measurement errors, represented by a normal distribution with mean 0 and variance σ_ε^2.

The model is known as a linear mixed effects model as it involves fixed effects (here blast and time, 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 McCulloch, Searle and Neuhaus [3]. Significance tests for the fixed effects and for any contrasts involving fixed effects that are deemed scientifically meaningful can also be carried out. For example, a contrast for the blast effect at time 1 may be of interest. The mean of the no-blast group at time 1 is $\alpha + \beta_1 + \gamma_1 + \beta\gamma_{11}$; the mean of the blast group at time 1 is $\alpha + \beta_2 + \gamma_1 + \beta\gamma_{21}$; hence the contrast for their difference is given by $\beta_1 - \beta_2 + \beta\gamma_{11} - \beta\gamma_{21}$. This contrast should be zero (insignificant) as time 1 is a pre-blast period.

Estimates of the two error variances come into play when testing the fixed effects. The variability between subjects is used when testing the blast effect; the measurement variability is used in the tests for time and the blast by time interaction. See, for example, Winer, Statistical Principles of Experimental Design, 2nd edition, pages 518 ff [4]. Since two different mean square errors are used for different tests, these experiments are also called split-plot experiments. Computer software for analyzing the data from this repeated measurement design is readily available. Here we describe four commonly-used packages: Minitab, SAS, R, and Graph Pad PRISM. They all lead to the same conclusions. Their general linear mixed effects model allows the user to incorporate random subject effects and repeated measures. An important feature of these software packages is that the 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 data sets would be of little use.

Minitab:

We use the General Linear Model (GLM) function under the Stat > ANOVA tab

MTB > GLM Y = blast mouse(blast) time blast*time;

SUBC> Random 'mouse';

SUBC> Brief 2.

ANOVA for the General Linear Model: Illustrative Example

Factor	Type	Levels	Values
blast	fixed	2	0, 1
mouse(blast)	random	25	5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 23, 24, 25, 1, 3, 4, 15, 16, 18, 19, 20, 21, 22, 26, 28
time	fixed	3	1, 2, 3

Analysis of Variance for roto, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
blast	1	390.74	390.74	390.74	3.89	0.061
mouse(blast)	23	2313.24	2313.24	100.58	4.89	0.000
above MS is used as denominator for the between subject comparison F(blast) = 390.74/100.58 = 3.89						
time	2	690.42	693.37	346.69	16.86	0.000
blast*time	2	42.99	42.99	21.50	1.05	0.360
Error	46	945.66	945.66	20.56		
above MS is used as denominator for the within subject comparisons F(time) = 346.69/20.56 = 16.86; F(blast*time) = 21.50/20.56 = 1.05						
Total	**74**	**4383.05**				

S = 4.53407 R-Sq = 78.42% R-Sq(adj) = 65.29%

The p-value for the blast by time interaction is 0.360; there is negligible interaction between the two fixed effects, blast and time. In the absence of an interaction, the main effects of time and blast can be interpreted on their own. The p-value for time is 0.000; the effect of time is quite significant. The p-value for blast is 0.061; the effect of blast is insignificant at the 0.05 significance level, but significant at the 0.10 significance level. We caution about interpreting probability values too narrowly; a p-value of 0.061 tells us that under the null hypothesis of no difference there is only a 6.1 percent chance of observing such a large effect (or one that is even larger) by pure chance. This small probability certainly puts doubt on the no-difference hypothesis.

SAS:

SAS is another frequently-used software. We use the SAS PROC MIXED package. The input and output for the appropriate analysis are shown below.

```
data example;
 input mouse blast time roto weight;
 datalines;
1       1       1       53.09   29.5
1       1       2       33.48   28.6
1       1       3       46.21   30.2
3       1       1       33.88   27.0
3       1       2       30.95   26.3
…
24      0       3       30.11   28.3
25      0       1       38.54   26.6
25      0       2       31.09   27.1
25      0       3       53.91   27.0
;
proc print data=example;
proc mixed data=example method=REML;
 class mouse blast time;
 model roto=time blast time*blast/solution COVB CORRB;
 random mouse(blast)/type=vc v vcorr;
 contrast 'blast at time 1'
 blast 1 -1 time*blast 1 0 0 -1 0 0/E;
 contrast 'blast at time 2'
 blast 1 -1 time*blast 0 1 0 0 -1 0/E;
 contrast 'blast at time 3'
 blast 1 -1 time*blast 0 0 1 0 0 -1/E;
 contrast 'blast at fixed times'
 blast 1 -1 time*blast 1 0 0 -1 0 0,
 blast 1 -1 time*blast 0 1 0 0 -1 0,
 blast 1 -1 time*blast 0 0 1 0 0 -1/E;
 contrast 'time 1 vs 2 at no blast'
 time 1 -1 0 time*blast 1 -1 0 0 0 0/E;
 contrast 'time 1 vs 3 at no blast'
 time 1 0 -1 time*blast 1 0 -1 0 0 0/E;
 contrast 'time 2 vs 3 at no blast'
 time 0 1 -1 time*blast 0 1 -1 0 0 0/E;
 contrast 'time 1 vs 2 at blast'
 time 1 -1 0 time*blast 0 0 0 1 -1 0/E;
 contrast 'time 1 vs 3 at blast'
 time 1 0 -1 time*blast 0 0 0 1 0 -1/E;
 contrast 'time 2 vs 3 at blast'
 time 0 1 -1 time*blast 0 0 0 0 1 -1/E;
 contrast 'time at no blast'
 time 1 -1 0 time*blast 1 -1 0 0 0 0,
 time 1 0 -1 time*blast 1 0 -1 0 0 0,
 time 0 1 -1 time*blast 0 1 -1 0 0 0/E;
 contrast 'time at blast'
 time 1 -1 0 time*blast 0 0 0 1 -1 0,
 time 1 0 -1 time*blast 0 0 0 1 0 -1,
 time 0 1 -1 time*blast 0 0 0 0 1 -1/E;
 contrast 'times at fixed blast'
 time 1 -1 0 time*blast 1 -1 0 0 0 0,
 time 1 0 -1 time*blast 1 0 -1 0 0 0,
 time 0 1 -1 time*blast 0 1 -1 0 0 0,
 time 1 -1 0 time*blast 0 0 0 1 -1 0,
 time 1 0 -1 time*blast 0 0 0 1 0 -1,
 time 0 1 -1 time*blast 0 0 0 0 1 -1/E;
run;
```

The first table in the output shows the variances of the two random effects, $\hat{\sigma}^2_{\pi} = 26.67$ and $\hat{\sigma}^2_{\varepsilon} = 20.56$. The second table summarizes the tests of the fixed effects: as expected, the results are the same as the ones we have seen earlier for Minitab. The last table illustrates the results when testing for the contrasts of interest. The first three lines list the probability values when testing for a blast effect at each of the three fixed time periods. We refer the reader to the program code on how to ask for these contrasts. As expected, there is no blast effect at time 1 (p-value = 0.382), while there are significant blast effects at time 2 (p-value = 0.043) and time 3 (p-value = 0.049). The insignificance of blast effects at time 1 is reassuring as measurements at time 1 are taken before the blast is administered to the mice in the blast group.

Covariance Parameter Estimates	
Cov Parm	Estimate
mouse(blast)	26.6726
Residual	20.5578

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
blast	1	23	3.89	0.0609
time	2	46	16.86	<.0001
blast*time	2	46	1.05	0.3597

Contrasts				
Label	Num DF	Den DF	F Value	Pr > F
blast at time 1	1	46	0.78	0.3823
blast at time 2	1	46	4.31	0.0435
blast at time 3	1	46	4.10	0.0488
blast at fixed times	3	46	1.99	0.1283
time 1 vs 2 at no blast	1	46	4.20	0.0461
time 1 vs 3 at no blast	1	46	3.82	0.0567
time 2 vs 3 at no blast	1	46	16.04	0.0002
time 1 vs 2 at blast	1	46	14.01	0.0005
time 1 vs 3 at blast	1	46	0.03	0.8568
time 2 vs 3 at blast	1	46	15.40	0.0003
time at no blast	2	46	8.02	0.0010
time at blast	2	46	9.82	0.0003
times at fixed blast	4	46	8.92	<.0001

R Statistical Software:

R is yet another frequently used software package. We use the **lmer** function in the R library **lme4**. Input and output for the appropriate analysis are shown below. Note that R parameterizes effects in a different (but equivalent) way. Instead of expressing blast effects as deviations from the average (with restriction $\beta_1 + \beta_2 = 0$), lmer sets the first coefficient equal to zero ($\beta_1 = 0$) and interprets β_2 as the effect of the second group (here the blast group, coded as 1) as compared to the first group (here the control group, coded as 0).

```
library(lme4)
data=read.table(header=TRUE,text ="mouse blast time roto weight
1 1 1 53.09 29.5
1 1 2 33.48 28.6
1 1 3 46.21 30.2
3 1 1 33.88 27.0
…
24 0 3 30.11 28.3
25 0 1 38.54 26.6
25 0 2 31.09 27.1
25 0 3 53.91 27.0
")
data
data$blast=factor(data$blast)
data$time=factor(data$time)
data$mouse=factor(data$mouse)
```

```
out1 = lmer(roto ~ (1|blast:mouse) + blast + time + time:blast, REML=TRUE, data=data)
out1
```

Linear mixed model fit by REML ['lmerMod']

Formula: roto ~ (1 | blast:mouse) + blast + time + time:blast

Data: data

REML criterion at convergence: 456.0832

Random effects:

Groups	Name	Std.Dev.
blast: mouse	(Intercept)	5.165
Residual		4.534

Number of obs: 75, groups: blast:mouse, 25

Fixed Effects:

(Intercept)	blast1	time2	time3	blast1:time2	blast1:time3
38.405	-2.427	-3.645	3.477	-3.284	-3.141

```
anova(out1)
```

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value
blast	1	79.87	79.87	3.8850
time	2	690.42	345.21	16.7922
blast:time	2	42.99	21.50	1.0457

```
summary(out1)
```

Linear mixed model fit by REML ['lmerMod']

Formula: roto ~ (1 | blast: mouse) + blast + time + time: blast

Data: data

REML criterion at convergence: 456.1

Scaled residuals:

Min	1Q	Median	3Q	Max
-1.83817	-0.48780	-0.01035	0.45855	2.15602

Random effects:

Groups	Name	Variance	Std.Dev.
blast:mouse	(Intercept)	26.67	5.165
Residual		20.56	4.534

Number of obs: 75, groups: blast: mouse, 25

Fixed effects:

	Estimate Std.	Error	t value
(Intercept)	38.405	1.906	20.149
blast1	-2.427	2.751	-0.882
time2	-3.645	1.778	-2.050
time3	3.477	1.778	1.955
blast1:time2	-3.284	2.567	-1.279
blast1:time3	-3.141	2.567	-1.224

Correlation of Fixed Effects:

	(Intr)	blast1	time2	time3	bls1:2
blast1	-0.693				
time2	-0.467	0.323			
time3	-0.467	0.323	0.500		
blast1:tim2	0.323	-0.467	-0.693	-0.346	
blast1:tim3	0.323	-0.467	-0.346	-0.693	0.500

Graph Pad PRISM:

Each column of the data matrix includes the measurements on a mouse, with the repeated measurements at times 1, 2 and 3 given in rows. The first 12 columns come from mice that were exposed to the blast (group A), while the next 13 columns represent the measurements of the control mice (group B). Each row represents a different time point, with matched (repeated) values being stacked into each column. The average response for the blast group is 4.569 units smaller than that of the control group.

Two-way RM ANOVA

Matching: Stacked

Assume sphericity? Yes

Alpha 0.05

Source of Variation	% of total var	P value	P value summary	Significant?
Time x Blast	0.9809	0.3597	ns	No
Time	15.82	<0.0001	****	Yes
Blast	8.915	0.0609	ns	No
Subject	52.78	<0.0001	****	Yes

ANOVA table	SS	DF	MS	F (DFn, DFd)	P value
Time x Blast	42.99	2	21.50	F (2, 46) = 1.046	P=0.3597
Time	693.4	2	346.7	F (2, 46) = 16.86	P<0.0001
Blast	390.7	1	390.7	F (1, 23) = 3.885	P=0.0609
Subject	2313	23	100.6	F(23,46) = 4.892	P<0.0001
Residual		945.7	46	20.56	

Difference between column means	
Mean of Group A	33.78
Mean of Group B	38.35
Difference between means	-4.569
SE of difference	2.318
95% CI of difference	-9.364 to 0.2263

Here we discussed a simple repeated measurement design. Two extensions that are often useful are discussed in Appendix 1.

DESIGN OF EXPERIMENTS

Introduction

Questions about the most effective ways to design experiments and issues of sample size and power come up all the time in medical research. Designing an appropriate experiment is not easy; at the outset of a study there is much uncertainty and not much is known. One could say that "the best time to design an experiment is after the results of the experiment have come in." But knowing this doesn't help the investigator. A thorough knowledge of experimental design principles can improve the efficiency of experiments. Important statistical design principles are replication, randomization, blocking, multi-factor instead of one factor at-a-time experimentation, and a sequential approach to experimentation. The sequential approach to experimentation is important, with the results from initial experiments being used to determine the next experimental steps. Only a portion of the overall budget should be spent on the initial runs. Detailed discussion on these principles can be found in books on the statistical design of experiments, such Box, Hunter and Hunter [5], Ledolter and Swersey [6], Montgomery [7], and of course in R.A. Fisher's seminal contributions. Knowledge about the subject area the experiment addresses, common statistical sense, and a careful implementation of the experimental study plan are also critical. In medical research, investigators run experiments all the time, and evidence-based medicine relies on randomized experiments to confirm which of several treatments are the most effective. The search for effective ways to design experiments and issues of sample size and statistical power are commonplace in scientific experimentation. If experiments are executed poorly, little or even nothing will be learned from the resulting data. While it is true that most experiments increase knowledge (one usually learns "something" through experimentation), the experimenter wants to learn as efficiently as possible. Relatively few experimental runs (observations) are needed in efficient experimental designs to get precise estimates of the factor effects. Sir Ronald Fisher, the eminent statistician and scientist who developed this area, said that "a well-designed experiment may improve the

precision of the results tenfold, for the same cost in time and labor" [8].

Prior to running an experiment one needs to determine the sample size required to identify scientifically meaningful effects. In other words, one must address the question whether a certain sample size is sufficient to detect a specified response effect. If the sample size is too small, observed effects may not be statistically significant and meaningful effects may not be uncovered.

It is very important to know whether the data that can be expected from an experiment have a realistic chance of detecting meaningful effects. Consider, for example, an experiment on patients that studies the effect of an intervention on certain measured health indicators. Typically one knows how large an effect must be to be judged clinically meaningful. Research studies are expensive, and costs increase with the number of subjects that need to be recruited into the study. Prior to running the experiment, one must calculate the statistical power of detecting (practically) meaningful effects. For some planned experiments this may not be possible; many more subjects may be needed to learn about clinically important effects. If one cannot afford the required sample sizes, one must restructure or abandon the problem in favor of problems that can be solved with the budget at hand. If there is little chance that meaningful effects can be detected, the money is better spent elsewhere. While medical grant proposals typically require a section on sample size and power, these sections are usually written in a defensive manner to justify the experimental plan the investigator has settled on a long time ago. Often these sections are written to defend a prior the investigator has, and they rarely assess critically whether the planned research is worth its cost. Many times they represent an intricate "song and dance" to justify why limited funds can be used to study something experimenters want to study. Experimenters need to understand that sample size studies are there to help them; sample size studies are not there to game the system to achieve funding.

Programs for calculating sample size and power

Most statistics packages have this capability. Also, there are programs dedicated to this task exclusively such as the sample size/power programs by Lenth [9]. Lenth's sample size applets (they are free, good and easy to use) cover many situations, such as tests involving:

- Numeric outcome variables (emphasis on means and variances)
- Categorical outcome variables (emphasis on proportions)
- Regression context such as estimating a slope
- Factorial experiments

Theory behind power studies: Testing a hypothesis about the mean of a normal distribution

Here we illustrate the basics behind a power study. We go through this simple example in great detail. The selection of sample sizes and power studies for a few other more complicated testing situations are summarized in Appendix 2.

Assume that you test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_1 : \mu < \mu_0$. You are testing the research hypothesis whether or not an intervention leads to a reduction from the current known mean μ_0.When determining the appropriate sample size, you need to specify values for the four following items:

- $\sigma = \sqrt{Var(Y)}$, the **standard deviation** of the measurement variable *Y*
- The **significance level** (the probability of rejecting a true null hypothesis); usually $\alpha = 0.05$
- The **power** (usually 0.80) to detect a certain specified **detectable difference of interest** $\delta = \mu_1 - \mu_0 < 0$. Note: $\beta = 1 - 0.8 = 0.2$ is the probability of a type II error (that is, accepting the null hypothesis H_0 if the mean has shifted to $\mu_1 = \mu_0 + \delta < \mu_0$)

Result: The required sample size is

$$n = \frac{\left[z_\alpha - z_{1-\beta}\right]^2}{\left[\frac{\mu_1 - \mu_0}{\sigma}\right]^2} = \left[\frac{z_\alpha + z_\beta}{\delta}\right]^2 \sigma^2 ;$$

z_α and z_β are percentiles of the standard normal distribution.

Proof: How does one obtain this result? The proof shown here provides insight into how such questions are solved. The same arguments can be applied to the more elaborate designs covered in Appendix 2.

Starting from the significance level,

$$\alpha = P[\text{reject } H_0 \mid H_0 \text{ is true}] = P[\bar{Y} \le c] = P[\frac{\bar{Y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \le \frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}}]$$

$$= P[Z \le \frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}}],$$

leads to an equation for the percentile of order α, $z_\alpha = \frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}}$.

Starting from the power,

$$Power = P[\text{Reject } H_0 \mid H_1 \text{ is true}] = P[\text{reject } H_0 \mid \mu_1 = \mu_0 + \delta]$$

$$= P[\bar{Y} \le c] = P[\frac{\bar{Y} - \mu_1}{\frac{\sigma}{\sqrt{n}}} \le \frac{c - \mu_1}{\frac{\sigma}{\sqrt{n}}}] = P[Z \le \frac{c - \mu_1}{\frac{\sigma}{\sqrt{n}}}],$$

leads to an equation for the percentile of order $1-\beta$, $z_{Power} = z_{1-\beta} = \frac{c - \mu_1}{\frac{\sigma}{\sqrt{n}}}$.

Solving the two equations $z_\alpha = \frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ and $z_{Power} = z_{1-\beta} = \frac{c - \mu_1}{\frac{\sigma}{\sqrt{n}}}$ for the two unknowns c and *n*, leads to the above result for the required sample size *n*.

Example: $\sigma = \sqrt{Var(Y)} = 1$ and detectable difference $\delta = \mu_1 - \mu_0 = -0.3$; $\alpha = 0.05$ and $z_\alpha == -1.645$; $\beta = 0.20$ (power = 0.80) and $z_{0.20} = -0.8416$.

Then:

$$n = \frac{[z_\alpha + z_\beta]^2}{\left[\frac{\delta}{\sigma}\right]^2} = \frac{(-1.645 - 0.8416)^2}{(0.09)} = (100/9)(2.4816)^2 = 68.4$$

The required sample size is 69.

Facts to remember:

- Sample size increases with power. The more power you want, the larger the sample size.
- Sample size increases with decreasing detectable difference. The smaller the difference you want to detect, the larger the sample size.
- Sample size increases proportionally to the variance. The larger the uncertainty, the larger the sample size must be. The sample size quadruples with a doubling of the standard deviation.
- Two-sided tests require a larger sample size than one-sided tests.

Comment: This result can be applied to the paired (blocked) test with response $D = Y - X$. In this case $\sigma = \sqrt{Var(D)} = \sqrt{Var(Y) + Var(X) - 2\text{cov}(Y,X)} < \sqrt{Var(Y) + Var(X)}$ if blocking has been effective.

STATISTICAL SIGNIFICANCE IS NOT ALL THAT MATTERS: EFFECT SIZE VS STATISTICAL SIGNIFICANCE

In 2005, JPA Ioannidis [10,11] wrote two influential papers that suggest that up to 50% of medical studies are not reproducible. Why is this so? Various explanations can be given to support this claim, among them

- Publication bias. Only statistically-significant results are being published. This excludes studies that have been underpowered from the start, with little chance of detecting meaningful effects.
- Institutional pressure to be funded and published. Trickery to "chase p-values" and to make results significant, even though a fresh view of the evidence may tell you otherwise. This has to do with the treatment of outliers, the selective use of methods and ignoring the violations of assumptions that are critically relevant to the adopted methods, and "looking away" from facts that may contradict what you want to see.

- Statistical significance and probability values is not all that matters. One must also look at the magnitude of the estimated effects. Repeated positive results, even though not statistically significant in each individual study, can combine to very significant results if results are combined through meta-analysis. However, if non-significant studies are not being published, a meta-analysis is compromised as one doesn't have access to studies with statistically insignificant results.
- Cohen's d relates the difference of the two group means to the pooled standard deviation; it is an estimate of the effect size. General "rule of thumb" guidelines consider Cohen's d of 0.2 as a small effect, 0.5 as a medium-sized effect, and 0.8 as a large effect. Cohen's d supplements the results of inferential testing [12].
- Statistical significance does not amount to much if the magnitude of the estimated effect is not scientifically/clinically relevant. One must not confuse statistical significance of estimated effects with the practical significance of estimated effects. Probability values alone do not tell the complete story, as even the smallest effect can be made significant if the sample size is increased.
- Statistical significance does have value as a protection against tampering. One should not change an established protocol on the basis of an estimated effect that has not been found statistically significant. Without statistical significance, the effect may be due to pure chance. But even if an estimated effect is significant, one may not change an established protocol if the size of the effect is scientifically/clinically irrelevant.
- Confidence intervals are preferable to probability values. Confidence intervals tell us about both the magnitude of the estimated effect and the uncertainty of the estimate.
- Probability values are preferable to binary statistical significance findings. Statistical significance (if this is what you are after) is better expressed through the probability value than the associated binary No/Yes statistical significance finding that one gets by comparing the probability value to an arbitrarily-chosen significance level cutoff such as 0.05 or 0.10. A result with probability value 0.105 is not all that different from one with probability value 0.095 or even 0.047, especially if the data set is small and if there is uncertainty whether all assumptions that go into a probability value are actually satisfied.
- In a 2019 special issue of The American Statistician [13], the American Statistical Association recommends against abusive use of probability values; the lead editorial suggests abandoning the use of the term "statistically significant" altogether.

META-ANALYSIS TO COMBINE RESULTS FROM DIFFERENT STUDIES

Meta-analysis is a statistical procedure for combining the results of multiple studies. Individual study results are measured with error. The aim of a meta-analysis is to derive a pooled estimate closest to the unknown common truth. While there are many different methods for meta-analysis (with each version making slightly different assumptions), all existing methods yield a weighted average of the results of the individual studies. The difference is in the way these weights are calculated and the way in which the uncertainty is computed around the weighted point estimate. The common and critical assumption in meta-analysis is that the results of multiple studies are independent. Studies include different subjects, and there are no links between the studies.

For illustration, we use the data from Bjordal et al. [14] shown below (Table 2). It lists the outcomes of 10 individual studies on the effect of non-steroidal anti-inflammatory drugs on osteoarthritic knee pain. Study summary statistics of the effects are shown below. A positive effect indicates that non-steroidal anti-inflammatory drugs work better than the placebo. The standard error of a study effect is obtained by dividing the standard deviation of the effect measurements from individual participants in the study with the square root of the sample size. 95% confidence intervals and one-sample t-ratios are also calculated. Two-sided probability values testing whether or not the intervention is significant (mean different than 0) are also shown. Results for three of the 10 studies are not statistically significant. We use this example to illustrate the basic concepts behind a meta-analysis. When performing a meta-analysis, the investigator must make choices and those choices can affect the results. Tricky issues here are selecting the appropriate studies based on objective criteria, dealing with incomplete data, analyzing the data, and dealing with

how to account for (or choosing not to account for) publication bias.

Table 2: Results of ten studies on the effect of non-steroidal anti-inflammatory drugs on osteoarthritic knee pain.

	nobs	estimated effect = y	se(effect) $\hat{\sigma}$	95% CI lower	95% CI higher	t-ratio	p-value 2-sided	$1/\hat{\sigma}^2$	weights w fixed effects
Dore (1995)	254	0.37	0.133	0.110	0.630	2.789	0.005	56.828	0.065
Fleischmann (1997)	279	0.04	0.128	-0.210	0.290	0.314	0.754	61.466	0.070
Kivitz (2002)	613	0.27	0.084	0.105	0.435	3.207	0.001	141.106	0.162
Lee (1985)	422	0.31	0.102	0.110	0.510	3.038	0.002	96.040	0.110
Lund (1998)	271	0.26	0.122	0.020	0.500	2.123	0.034	66.694	0.076
Schnitzer (1995)	270	0.40	0.133	0.140	0.660	3.015	0.003	56.828	0.065
Scott (2000)	610	0.08	0.082	-0.080	0.240	0.980	0.327	150.062	0.172
Tannenbaum (2004)	1702	0.20	0.069	0.065	0.335	2.904	0.004	210.787	0.242
Uzun (2001)	39	0.53	0.357	-0.170	1.230	1.484	0.138	7.840	0.009
Williams (2001)	104	0.38	0.202	-0.015	0.775	1.886	0.059	24.622	0.028

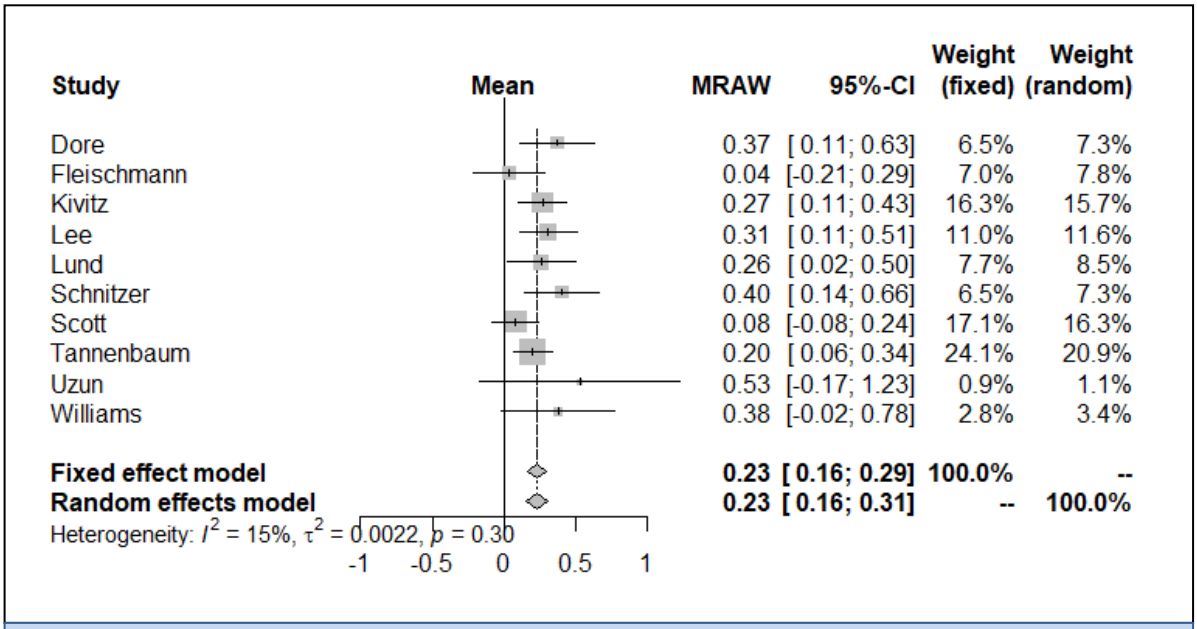


Figure 3: Meta-analysis of the data from Table 2.

The fixed-effects model

The fixed-effects model calculates a weighted average of the reported estimated study effects, y_i.The sample variance of a study effect, denoted by $\hat{\sigma}_i^2$, reflects the reliability of the estimate; it can be obtained by squaring the standard error in the fourth column of the table. The inverse of this variance, $\hat{\sigma}_i^{-2}$, is commonly used as the weight for the study effect, so that larger studies contribute more than smaller studies to the weighted average. The weights $\hat{\sigma}_i^{-2}$ and the normalized weights $w_i = \frac{\hat{\sigma}_i^{-2}}{\sum \hat{\sigma}_i^{-2}}$ are shown in the last two columns of the table. The pooled estimate of the common treatment effect calculates the weighted average of the study effects, $effect_{pooled} = \sum_{i=1}^{n} w_i y_i$; the standard error of the pooled estimate is given by $se(effect_{pooled}) = \sqrt{\frac{1}{\sum \hat{\sigma}_i^{-2}}}$.

These results are the ones we get when applying generalized least squares to estimate a common mean effect from *n* study effects with associated variances $\hat{\sigma}_i^2$ (see Abraham and Ledolter [15].

For our example, $effect_{pooled} = 0.228$ and $se(effect_{pooled}) = 0.034$, yielding an approximate 95% confidence interval $0.228 \pm (2)(0.034)$ that extends from 0.16 to 0.30. The probability value for testing whether or not the common effect is zero is less than 0.001.

The fixed-effects model assumes that all included studies are from the same population. This assumption may be unrealistic as studies are heterogeneous. For example, the treatment effects may differ according to local study conditions and dosage levels. This is where the random-effects model shown next comes in as it relaxes this assumption.

The random-effects model

The random-effects model assumes that the treatment effect from the *i*th study, Y_i, is distributed as $Y_i \mid \mu_i \sim N(\mu_i, \sigma_i^2)$ where μ_i is the true underlying treatment effect of the *i*th study and σ_i^2 is the corresponding within-study variance. The variance σ_i^2 is unknown but an estimate $\hat{\sigma}_i^2$ is available from each individual study. The random-effects model further assumes that $\mu_i \sim N(\mu, \tau^2)$, where μ and τ^2 denote the overall treatment effect and the between-study variance, respectively. These two assumptions imply the marginal distribution $Y_i \sim N(\mu, \sigma_i^2 + \tau^2)$.

Random-effects procedures for meta-analysis differ by how τ^2 gets estimated. The procedure suggested by Der Simonian and Laird [16] is the simplest and most common random-effects method. Their approach uses the Q statistic, $Q = \sum_{i=1}^{n} \sigma_i^{-2} (y_i - \bar{y})^2$ where $\bar{y} = \sum_{i=1}^{n} \sigma_i^{-2} y_i \Big/ \sum_{i=1}^{n} \sigma_i^{-2} = \sum_{i=1}^{n} w_i y_i$ is the pooled estimate under the fixed-effects model. Under the assumptions of the random-effects model (where the observed values y_i are generated from the above model for the Y_i) it can be shown that the expectation of Q is $E(Q) = (n-1) + (S_1 - \frac{S_2}{S_1})\tau^2$ where $S_1 = \sum_{i=1}^{n} \sigma_i^{-2}$ and $S_2 = \sum_{i=1}^{n} \sigma_i^{-4}$. Replacing all unknown variances σ_i^2 with their estimates $\hat{\sigma}_i^2$ and solving this equation for τ^2 leads to the Der Simonian and Laird estimate $\hat{\tau}_{DL}^2 = \max(0, \frac{Q-(n-1)}{S_1 - \frac{S_2}{S_1}})$.

The estimate of the common treatment effect is then given by the weighted average

$$effect_{DL} = \frac{\sum_{i=1}^{n} \frac{y_i}{\hat{\sigma}_i^2 + \hat{\tau}_{DL}^2}}{\sum_{i=1}^{n} \frac{1_i}{\hat{\sigma}_i^2 + \hat{\tau}_{DL}^2}} = \sum_{i=n} \tilde{w}_i y_i,$$ with

weights $\tilde{w}_i = \frac{1/(\hat{\sigma}_i^2 + \hat{\tau}_{DL}^2)}{\sum_{i=1}^{n} [1/(\hat{\sigma}_i^2 + \hat{\tau}_{DL}^2)]}$. The standard error of the

estimate is $se(effect_{DL}) = \sqrt{\frac{1}{\sum_{i=1}^{n} \frac{1}{\hat{\sigma}_i^2 + \hat{\tau}_{DL}^2}}}$. Confidence intervals and the probability value can be calculated accordingly.

For our data set $Q = 10.6986$, $S_1 = 872.27$, $S_2 = 111,437$ and $\hat{\tau}_{DL}^2 = 0.002281$. The estimate of the overall treatment effect is $effect_{DL} = 0.234$ and $se(effect_{DL}) = 0.038$.

The difference between a random-effects and a fixed-effects model for meta-analysis is that the random-effects model allows for variability among the study effects, while in the fixed-effects model all studies are assumed to originate from a single common mean. In our example the results of the random-effects and the fixed-effects meta-analysis are quite similar. Larger differences can be expected if the variability among the study effects gets larger.

The R library **meta** can be used to carry out the analysis and visualize the results. The output shows both the fixed-effects and random-effects weights. The forest plot visualizes the information. The forest plot is a graphical display of estimated results from a number of scientific studies addressing the same question, along with the overall meta-analysis results. It plots the estimated effect for each of these studies (with each estimate represented by a square) and displays confidence intervals by horizontal lines. The area of each square is proportional to the weight that is attached to the study by the meta-analysis. The pooled effect from the meta-analysis is commonly plotted as a diamond. A vertical line representing no effect is also plotted. Confidence intervals for individual studies that overlap with this line demonstrate that at the given level of confidence their effect sizes do not differ from zero. The name forest plot refers to the forest of lines produced.

```
> library(meta)
> data=read.table(header=TRUE,text="nobs effect se study
+ 254 0.37 0.133 Dore
+ 279 0.04 0.128 Fleischmann
+ 613 0.27 0.084 Kivitz
+ 422 0.31 0.102 Lee
+ 271 0.26 0.122 Lund
+ 270 0.40 0.133 Schnitzer
+ 610 0.08 0.082 Scott
+ 1702 0.20 0.069 Tannenbaum
+ 39 0.53 0.357 Uzun
+ 104 0.38 0.202 Williams
+ ")
> data$s=data$se*sqrt(data$nobs)
> data
```

	nobs	effect	se	Study	s
1	254	0.37	0.133	Dore	2.119671
2	279	0.04	0.128	Fleischmann	2.138022
3	613	0.27	0.084	Kivitz	2.079742
4	422	0.31	0.102	Lee	2.095349
5	271	0.26	0.122	Lund	2.008373
6	270	0.40	0.133	Schnitzer	2.185413
7	610	0.08	0.082	Scott	2.025251
8	1702	0.20	0.069	Tannenbaum	2.846616
9	39	0.53	0.357	Uzun	2.229464
10	104	0.38	0.202	William	2.060004

```
> mm=metamean(nobs,effect,s,study,data=data)
> print(mm)
```

	mean	95%-CI	%W(fixed)	%W(random)
Dore	0.3700	[0.1093; 0.6307]	6.5	7.3
Fleischmann	0.0400	[-0.2109; 0.2909]	7.0	7.8
Kivitz	0.2700	[0.1054; 0.4346]	16.3	15.7
Lee	0.3100	[0.1101; 0.5099]	1 1.0	11.6
Lund	0.2600	[0.0209; 0.4991]	7.7	8.5
Schnitzer	0.4000	[0.1393; 0.6607]	6.5	7.3
Scott	0.0800	[-0.0807; 0.2407]	17.1	16.3
Tannenbaum	0.2000	[0.0648; 0.3352]	24.1	20.9
Uzun	0.5300	[-0.1697; 1.2297]	0.9	1.1
Williams	0.3800	[-0.0159; 0.7759]	2.8	3.4

Number of studies combined: k = 10

	mean	95%-CI	z p-value
Fixed effect model	0.2285	[0.1621; 0.2950]	-- --
Random effects model	0.2336	[0.1588; 0.3084]	-- --

Quantifying heterogeneity:

tau^2 = 0.0022; H = 1.09 [1.00; 1.52]; I^2 = 15.4% [0.0%; 56.7%]

Test of heterogeneity:

Q d.f.	p-value
10.64	9 0.3013

Details on meta-analytical method:

- Inverse variance method
- DerSimonian-Laird estimator for tau^2
- Untransformed (raw) means

> summary (mm)

Number of studies combined: k = 10

	mean	95%-CI	z p-value
Fixed effect model	0.2285	[0.1621; 0.2950]	-- --
Random effects model	0.2336	[0.1588; 0.3084]	-- --

Quantifying heterogeneity:

tau^2 = 0.0022; H = 1.09 [1.00; 1.52]; I^2 = 15.4% [0.0%; 56.7%]

Test of heterogeneity:

Q d.f.	p-value
10.64	9 0.3013

Details on meta-analytical method:

- Inverse variance method
- Der Simonian-Laird estimator for tau^2
- Untransformed (raw) means

> forest (mm)

REFERENCES

1. Yap BW, Sim CH. (2011). Comparisons of various types of normality tests. Journal of Statistical Computation and Simulation. 81: 2141-2155.
2. Box GEP, Cox DR. (1964). An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological). 26: 211-252.
3. McCulloch CE, Searle SR, Neuhaus JM. (2008). Generalized, Linear, and Mixed Models. 2nd Edition. New York: Wiley& Sons.
4. Winer BJ. (1999). Statistical Principles of Experimental Design. 2nd edition, McGraw-Hill, pages 518 ff.
5. Box GEP, Hunter JS, Hunter WG. (2005). Statistics for Experimenters: Design, Innovation, and Discovery. 2nd Edition. Wiley-Interscience.
6. Ledolter J, Swersey A. (2007). Testing 1 - 2 - 3: Experimental Design with Applications in Marketing and Service Operations. Stanford CA: Stanford University Press.
7. Montgomery D. (2012). Design and Analysis of Experiments. 8th edition. New York: Wiley & Sons.
8. Fisher RA. (1935). The Design of Experiments. 9th edition (Macmillan, 1971). page 217.
9. Lenth RV. (2006). Java Applets for Power and Sample Size [Computer software].
10. Ioannidis JPA. (2005). Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. JAMA. 294: 218-228.
11. Ioannidis JPA. (2005). Why Most Published Research Findings Are False. PLoS Med. 2: 124.
12. Cohen J. (1988). Statistical Power Analysis for the Behavioral sciences. 2nd Edition. NJ: Lawrence Erlbaum.
13. Wasserstein RL, Schirm AL, Lazar NA. (2019). Moving to a World Beyond "p < 0.05". Editorial. The American Statistician. 73: 1-19.
14. Bjordal JM, Ljunggren AE, Klovning A, Slørdal L. (2004). Non-steroidal anti-inflammatory drugs, including cyclo-oxygenase-2 inhibitors, in osteoarthritic knee pain: meta-analysis of randomised placebo controlled trials. BMJ. 329: 1317.
15. Abraham B, Ledolter J. (2006). Introduction to Regression Modeling. Duxbury Press. P: 128.
16. DerSimonian R, Laird N. (1986). Meta-analysis in clinical trials. Control Clin Trials. 7: 177-188.
17. Ledolter J. (2013). Economic Field Experiments: Comments on Design Efficiency, Sample Size and Statistical Power. Journal of Economics and Management. 9: 271-290.

APPENDIX 1: ANALYSIS OF DATA FROM TWO OTHER USEFUL REPEATED MEASUREMENT DESIGNS

Model 1:

Let us generalize the repeated measurement design that we discuss in the manuscript. Let us now assume that there is a third fixed factor, called new, and that this factor is crossed with time. For example, this third factor may express whether or not mice are "blindfolded" when running on the rotarod. Being crossed with time means that at each time a mouse gets exposed to both settings of the new factor, preferably in a randomized order with a coin flip deciding whether the blindfold is applied at the first or at the second rotarod session. The analysis assumes that the resulting 2*3 = 6 measurements on the same mouse are independent. We know that there are level effects due to the different mice, which leads us to also include random mouse effects into the model.

For illustration, suppose that there are 12 mice in the blast group and 13 different mice in the control group. Then the design layout would look like this

	Time=1		**Time=2**		**Time=3**	
	New=1	**New=2**	**New=1**	**New=2**	**New=1**	**New=2**
Blast Yes	G1	G1	G1	G1	G1	G1
Blast NO	G2	G2	G2	G2	G2	G2

For this experimental setup the appropriate model is given by

$$Y_{ijkl} = \alpha + \beta_j + \varepsilon_{i(j)} + \gamma_k + \beta\gamma_{jk} + \delta_l + \beta\delta_{jl} + \gamma\delta_{kl} + \beta\gamma\delta_{jkl} + \varepsilon_{i(j)kl}$$

where *i* stands for subject, *j* for blast, *k* for time, and *l* for the factor new. The presence of a 3-factor interaction between blast, time and the factor new allows the two-factor interactions of any two factors depend on the third factor. The presence of such 3-factor interaction can be tested.

The appropriate ANOVA table for this particular repeated measurement design is shown below.

Source	**DF**
Between subjects	
blast	1
mouse within groups=blast	23
Within subjects	
time	2
blast*time	2
new	1
blast*new	1
time*new	2
blast*time*new	2
error within subjects	115
Total	149

The F-ratio for testing the significance of blast uses the mean square error that comes from the between subjects analysis (the mean square of mouse within blast). The F-ratios for all other fixed effects use the mean square error that comes from the within subjects analysis.

Below we show the input code for MINITAB

```
MTB > GLM roto = blast mouse(blast) time blast*time new blast*new &
CONT> time*new blast*time*new;
SUBC> Random mouse;
SUBC> Brief 2.
```

And SAS

```
data mark;
 input mouse blast time new roto;
 datalines;
1       1       1       1       32.1301
1       1       2       1       30.0854
1       1       3       1       29.7417
1       1       1       2       30.1475
1       1       2       2       31.7020
1       1       3       2       30.8678
.
.
.
28      1       1       2       28.4039
28      1       2       2       30.5219
28      1       3       2       29.8264
proc print data=mark;
proc mixed data=mark;
 class mouse blast time new;
 model roto=blast time blast*time new blast*new time*new blast*time*new /solution;
 random mouse(blast)/type=vc v vcorr;
run;
```

Model 2:

Let us change the experiment in the following way. Assume now that a third factor "type" represents two different genetic mouse strains. Our experiment studies blasted and control (not-blasted) mice of either genetic strain. Blast 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 four groups is then observed at three different times. This design looks like this

		Time			
Blast Yes	Strain1	G1	G1	G1	6 mice
Blast Yes	Strain2	G2	G2	G2	5 mice
Blast NO	Strain1	G3	G3	G3	7 mice
Blast NO	Strain2	G4	G4	G4	4 mice

This is a different repeated measurements design as now subjects are nested within the blast-strain combinations (there are 4 such groups). You certainly can't have the same mouse come from both strain1 and strain2, and we do not allow a mouse to be in both the control and blasted group. Each of the four groups contains different mice.

For this experimental setup the appropriate model is given by

$$Y_{ijkl} = \alpha + \beta_j + \gamma_k + \beta\gamma_{jk} + \varepsilon_{i(jk)} + \delta_l + \beta\delta_{jl} + \gamma\delta_{kl} + \beta\gamma\delta_{jkl} + \varepsilon_{i(jk)l}$$

Where *i* stands for subject, *j* for blast, *k* for strain, and *l* for time. The appropriate ANOVA table for this particular repeated measurement design is shown below.

Source	**DF**
Between subjects	
blast	1
strain	1
blast*strain	1
mouse within groups=blast*strain	18
Within subjects	
time	2
blast*time	2
strain*time	2
blast*strain*time	2
time*subject within groups=blast*strain	36 [2*5+2*4+2*6+2*3]
Total	65

The F-ratios for blast, strain, and the blast*strain interaction use the mean square error that comes from the between subjects analysis (the mean square of mouse within the 4 groups). F-ratios for all other fixed effects use the mean square error from the within-subjects analysis.

Below, we have listed the appropriate code for MINITAB

```
MTB > GLM roto = blast strain blast*strain mouse(blast strain) time &
CONT> blast*time strain*time blast*strain*time;
SUBC> Random mouse;
SUBC> Brief 2.
```

and SAS

```
data mark;
 input mouse blast strain time roto;
 datalines;
1       1       1       1       53.09
1       1       1       2       33.48
1       1       1       3       46.21
3       1       1       1       33.88
.
.
23      0       2       2       37.99
23      0       2       3       55.00
;
proc print data=mark;
proc mixed data=mark;
 class mouse blast strain time;
 model roto=blast strain blast*strain time blast*time strain*time blast*strain*time /solution;
 random mouse(blast*strain)/type=vc v vcorr;
run;
```

The layout for Graph Pad PRISM is similar to the one for the example in the main part of the paper, except here the columns of measurements on each mouse are grouped into four groups, with a group structure for each of the two factors (blast and strain). Each column of the data matrix includes the measurements on a mouse, with repeated measurements stacked in each column.

APPENDIX 2: MORE RESULTS ON SAMPLE SIZE DETERMINATION AND POWER STUDIES

Case 1: Testing a hypothesis about the difference of the means from two independent normal distributions

Assume that you want to test $H_0 : \mu_2 - \mu_1 = 0$ against $H_1 : \mu_2 - \mu_1 < 0$. As before, we need to specify values for the following (now five) quantities:

- σ_1 and σ_2 : **two standard deviations** that need not be equal
- **significance level**; usually $\alpha = 0.05$
- **power** (usually 0.80) to detect a given **detectable difference of interest** $\delta = \mu_2 - \mu_1 < 0$

The function of the data relevant to test the above hypothesis is $\frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\frac{(\sigma_2)^2}{n_2} + \frac{(\sigma_1)^2}{n_1}}}$. The proof of the results that are shown below can be found, for example, in Ledolter [17].

Result: The required total sample size (for groups 1 and 2 together) is given by

$$N = \left[\frac{z_\alpha + z_\beta}{\delta}\right]^2 \left[\sigma_1 + \sigma_2\right]^2.$$

The sample sizes in the two groups, n_1 and n_2, should be selected proportional to the standard deviations such that $\frac{n_1}{n_2} = \frac{\sigma_1}{\sigma_2}$.

That is, $n_1 = \frac{\sigma_1}{\sigma_1 + \sigma_2} N$ and $n_2 = \frac{\sigma_2}{\sigma_1 + \sigma_2} N$.

Result: Assuming that the standard deviations are the same ($\sigma_1 = \sigma_2 = \sigma$), the optimal sample size in either of the two groups is $n = 2\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2 \sigma^2$, for a combined sample size of $N = 2n = 4\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2 \sigma^2$.

Facts:

- Sample sizes should be proportional to the standard deviations. Equal sample sizes should be selected for $\sigma_1 = \sigma_2$.
- Sample size increases with power. The more power you want, the larger the sample size
- Sample size increases with decreasing detectable difference. The smaller the difference you want to detect, the larger the sample size.
- Sample size increases proportionally to the variances. The larger the uncertainty, the larger the sample size must be.
- Two-sided tests require a larger sample size than 1-sided tests.

Example: $\sigma_2 = 3$ and $\sigma_1 = 1$; detectable difference $\delta = \mu_2 - \mu_1 = -0.5$; $\alpha = 0.05$ and $z_\alpha = -1.645$; $\beta = 0.20$ (power = 0.80) and $z_{0.20} = -0.8416$.

Then: $N = \left[\frac{z_\alpha + z_\beta}{\delta}\right]^2 \left[\sigma_2 + \sigma_1\right]^2 = \left[\frac{-1.645 - 0.8416}{-0.5}\right]^2 (3+1)^2 = 394$, for a total sample size of about 400. We should put 300 subjects into the group with standard deviation $\sigma_2 = 3$ and 100 subjects into the group with standard deviation $\sigma_1 = 1$.

Case 2: Sample sizes when comparing two independent lognormal distributions

Often the response variable Y follows lognormal distributions, which implies that $X = \log Y$ has a normal distribution with mean μ and standard deviation σ. The mean of the lognormal distribution is given by $E(Y) = \exp(\mu + 0.5\sigma^2)$, and the variance by $Var(Y) = \left[E(Y)\right]^2 \left[\exp(\sigma^2) - 1\right]$.

Usually we are given the coefficient of variation for variable *Y*. Using results about the mean and variance of a lognormal distribution, the coefficient of variation is given by $c = \frac{\sqrt{Var(Y)}}{E(Y)} = \sqrt{\exp(\sigma^2) - 1}$. We can solve this equation for σ, the standard deviation of the log-transformed observations $X = \log Y$. It is $\sigma = \sqrt{\log(1 + c^2)}$.

Furthermore, typically we are given the (proportionate) effect in the levels of the *Y*-observations that we want to detect. This means that

$E(Y_1) = E(Y_0)(1 + f)$ or $\exp(\mu_1 + 0.5\sigma^2) = (1 + f)\exp(\mu_0 + 0.5\sigma^2)$.

f = 0.2 means that we want to detect a 20 percent increase in the level; *f* = 0.25 means that we want to detect a 25% increase. We have assumed here that σ is the same in both groups and that the change is only in μ. This implies that the coefficients of variation under the null and alternative hypothesis are assumed the same. Under these assumptions, the difference in the means of transformed log-observations is $\delta = \mu_1 - \mu_0 = \log(1 + f)$.

Hence, for the power calculations, we transform the data to logs, $X = \log Y$ with $\sigma = \sqrt{\log(1 + c^2)}$, and apply the result for case 1. We want to detect the difference $\delta = \mu_1 - \mu_0 = \log(1 + f)$. With one-sided significance α and power $1 - \beta$, the number of observations needed in each group is

$$n = 2\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2 \sigma^2 = 2\left[\frac{z_\alpha + z_\beta}{\log(1 + f)}\right]^2 \log(1 + c^2)$$

Example: Assume, for illustration, coefficient of variation $c = 0.15$. Then $\sigma = \sqrt{\log(1 + (0.15)^2)} = 0.149$ .Assume that we want to detect a 20 percent change in outcome ($f = 0.20$); then $\delta = \log(1 + 0.2) = 0.182$. Hence

$$n = 2\left[\frac{z_\alpha + z_\beta}{\log(1 + f)}\right]^2 \log(1 + c^2) = 2\left[\frac{-1.645 - 0.8416}{0.182}\right]^2 (0.149)^2 = 8.29$$

That is, we need 9 subjects in each of the two groups.

Case 3: Cluster designs. Your sample sizes may need to be larger than what you think

Assume that we study two groups, with equal variances. The earlier result (case 1) shows that the sample size for each of the two groups is $n = n_1 = n_2 = 2\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2 \sigma^2$. The derivation assumes that the two treatments are assigned to the experimental units (subjects, rats, mice) at random.

Sometimes the randomization is carried out on **clusters** that consist of groupings of the experimental units. Clusters could be communities, and experimental units could be people. The randomization is at the cluster level; the treatment groups (experimental and control, such as absence and presence of a certain incentive) are assigned to **clusters** at random, and each of *m* experimental

units in a cluster is then assigned to the same treatment. While the data of interest comes from the experimental units in the two experimental groups, the randomization is carried out on the clusters. Usually subjects from the same cluster tend to be more alike. Since the observations in the same cluster are most likely correlated, with intra cluster correlation coefficient $\rho > 0$, the *m* observations in a cluster don't carry the same weight as *m* independent observations. Hence, in the presence of large intra cluster correlation it is important to randomize over many small clusters so as to maximize the efficiency of the experiment. Taking more and more replicates within a rather small number of clusters may not provide the desired power.

Here is a theoretical justification. The variability of an experimental unit is the sum of two variances, $\sigma^2 = \sigma_C^2 + \sigma_\varepsilon^2$; a cluster variance σ_C^2 and a unit-specific variance σ_ε^2. The intra cluster correlation coefficient is $\rho = \dfrac{\sigma_C^2}{\sigma_C^2 + \sigma_\varepsilon^2}$. Assume that each cluster contains *m* experimental units. The cluster average has variance $\sigma_C^2 + \dfrac{\sigma_\varepsilon^2}{m}$. The required number of clusters in each treatment group (for specified significance level and specified power at given detectable difference δ) is

$$k = 2\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2\left[\sigma_C^2 + \frac{\sigma_\varepsilon^2}{m}\right].$$

Hence the required number of observations *n* (number of clusters, *k*, times number of observations in each cluster, *m*) in each treatment group is

$$n = 2\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2\left[\sigma_\varepsilon^2 + m\sigma_C^2\right] = 2\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2\sigma^2\left[\frac{\sigma_\varepsilon^2 + m\sigma_C^2}{\sigma_C^2 + \sigma_\varepsilon^2}\right]$$

$$= 2\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2\sigma^2\left[1 + (m-1)\frac{\sigma_C^2}{\sigma_C^2 + \sigma_\varepsilon^2}\right] = 2\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2\sigma^2\left[1 + (m-1)\rho\right].$$

The intra cluster correlation inflates the sample size that we obtain under complete random sampling, $2\left[\dfrac{z_\alpha + z_\beta}{\delta}\right]^2\sigma^2$, by the factor $\left[1 + (m-1)\rho\right]$. For $\rho = 0$, we are back at our earlier result. For $\rho = 1$, we multiply the sample size that we obtain under complete random sampling by the number of experimental units in the cluster (*m*). Each experimental unit in a cluster is a carbon-copy of the other units in that cluster. The *m* experimental units in the cluster basically count as one unit (and not as *m*).Hence, in the presence of large intra cluster correlation, it is important to randomize over many, small clusters so as to maximize the efficiency of the experiment. Taking more and more replicates within the cluster doesn't increase the power of the experiment. Taking more clusters does.

Case 4: Sample size in the linear regression context

The least squares estimate of the slope in the linear regression model, $y = \alpha + \beta x + \varepsilon$, has standard error $se(\hat{\beta}) = \left[\dfrac{\sigma_\varepsilon}{\sigma_x}\right]\Big/\sqrt{n}$. One wants to select the values of *x* such that the standard error of the slope estimate is as small as possible. One usually knows the experimental region for the *x*'s, but one needs to come up with a design for the placement of the

x's. The standard error of the slope estimate is minimized if we put $n/2$ observations on either end of the experimental region. This is the most efficient (optimal) design, but leaves no room for model checking. In order to allow for checking curvature (quadratic effect), one wants to put at least some x's into the middle of the experimental region.