The hormonal regulation of carbamoylphosphate synthetase I expression
Schoneveld, J.L.M.

Citation for published version (APA):
Schoneveld, J. L. M. (2004). The hormonal regulation of carbamoylphosphate synthetase I expression

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 2

“Factor Correction”: an alternative for normalisation and standardisation to remove between-session variation

Jan M. Ruijter, Onard J.L.M. Schoneveld, Wouter H. Lamers
Chapter 2

**ABSTRACT**
In biomedical research, "normalisation" is a common approach to deal with series of measurements with similar proportional differences between experimental conditions, but different absolute values, even though the measurement sessions were carried out under presumably identical circumstances. A major limitation of the normalisation approach is that only a single experimental condition is used to correct for variation between sessions. Another method to remove between-session variation, "standardisation", requires a complete data set to avoid added variation. To remedy these shortcomings, we developed a novel method for the removal of the between-session variation, which performs well with incomplete data sets. This method, dubbed "factor correction" assumes that the between-session variation is due to multiplicative factors working on the respective session. These session factors can be determined by two different approaches: two-way analysis of variance (ANOVA) on log-transformed data, or calculation of a between-session ratio matrix. Both approaches can handle incomplete datasets and use the observed values in all experimental conditions to calculate the session factors. When (part of) the between-session variation is not multiplicative, this variation component remains present after factor correction. Depending on the experimental design and the biological variation the factors resulting from both approaches differ up to 2%, which is negligible compared to biological variation. *

* A computer program that performs factor correction with the ratio approach is available on request: biolab-services@amc.uva.nl; subject: factor correction.
INTRODUCTION

Repeating a series of measurements in biological research under presumably identical circumstances on another day often leads to results that show the same proportional differences between experimental conditions, but clearly different absolute values within the conditions (Figure 1A). This between-session variation results from small random differences in e.g. cell densities, substrate and reagent concentrations, reactions temperatures and exposure times, which all have been shown to proportionally increase or decrease the outcome of a biological measurement. The combined effect of such experimental variables is illustrated in Figure 1, in which the lines connect the measurements resulting from one session.

Figure 1. Comparison of normalisation, standardisation and factor correction. The sample data set shows the activity of 8 different DNA constructs (=conditions) measured in 6 independent sessions (◆ □ ◆ ◄ ■ ◆). A: Original measurements plotted on a logarithmic Y-axis. The parallel lines connecting the results from each session indicate that the variation between sessions is multiplicative. B: Data from A after normalisation, using condition 1 as 'control' (one session did not include condition 1 and had to be dropped). The variation in the control condition is lost (■). C: Data from A after standardisation. Note that a linear transformation (standardised * = 410 + 305 x standardised) was applied to the standardised values to enable this logarithmic plot. D: Data from A after applying factor correction. The minimal remaining distance between the lines indicates that factor correction does a superior job in removing the multiplicative between-session variation.
Chapter 2

The fact that most session lines are parallel in this graph with a logarithmic Y-axis indicates that the between-session variation is indeed the result of multiplicative factors working on each session. The resulting observations can, therefore, be modelled as a mixed multiplicative and additive model (Eq. 4). When such a multi-session experiment is analysed with a two-way analysis of variance (ANOVA) one can assign the respective variance components to the sessions and conditions and test for condition effects. However, when ANOVA is applied directly to the observations, the session factors are treated as additive instead of multiplicative. On the other hand, when the logarithms of the observations are entered into the ANOVA, the condition effects and statistical error are treated as multiplicative effects. The often incomplete and unbalanced design of multi-session experiments, further hinders the direct parametric statistical analysis of such datasets. A nonparametric approach to deal with this between-session variation, that is, replacing the values per session by their ranks, is similar to the Friedman test\(^1\) and requires a complete dataset.

To remedy this situation, the researcher customarily tries to remove between-session variation by either “normalisation” or “standardisation”\(^2\). In normalisation, a ‘control’ condition is defined and per session all measured values \((Y_{ni})\) are scaled with respect to the control value in the session \((Y_{n1})\) according to Equation 1 (with session \(n\), condition \(i\) and control condition 1).

\[
\text{normalised } Y_{ni} = 100 \times \frac{Y_{ni}}{Y_{n1}} \tag{1}
\]

A weak point of the normalisation approach is that only a single experimental condition is used to correct for the between-session variation. Moreover, this control value is implicitly assumed to be without experimental error. Figure 1B shows the data for each condition when normalisation (using condition 1 as control, which has lead to the loss of one session!) is applied to the sample data set. Normalisation does remove between-session variation but, at the same time, generates a control condition without variation and adds the variation that was present in the control condition to the variation in the other conditions (Figure 2B). Since parametric statistical tests for the comparison of two or more conditions assume an equal variance in all conditions\(^3\), these tests can no longer be used. Also most nonparametric tests are no longer applicable, because they require similar distributions in all conditions\(^1\).
In standardisation, each value per session is transformed into a standard value by subtracting the session mean ($\bar{Y}_n$) and dividing by the session standard deviation ($SD_n$, Equation 2).

$$\text{standardised } Y_n = \frac{Y_n - \bar{Y}_n}{SD_n}$$

Because the session mean after standardisation is zero, standardisation removes between-session variation (Figure 1C). However, when not all conditions are present in every session, the session mean and standard deviation will be biased, which in turn will result in biased standard values and added variability between sessions (triangles and filled diamonds in Figure 1C; Figure 2C). Therefore, standardisation can only be used effectively when the data set is complete, that is, when all conditions are present in every session.

Figure 2. Comparison of normalisation, standardisation and factor correction. Mean and standard error of the original data (A), the data after normalisation (B), after applying standardisation (C), and after applying factor correction (D). Note that normalisation, standardisation, and factor correction reduce the variation within each condition. However, normalisation (B) leads to loss of variance in the control condition and to added variation in the other conditions. With factor correction (D) all conditions retain their statistical variance, which is generally smaller than after normalisation (B) and standardisation (C).
Both correction methods implicitly assume that the between-session variation is due to multiplicative between-session factors. In this paper we demonstrate two approaches for a correction method that is based on the direct estimation of those multiplicative between-session factors. This method, dubbed "factor correction" uses a mixed additive and multiplicative model for the within- and between-session variation. The first approach is to estimate session factors from the output of a two-way analysis of variance (ANOVA) on log-transformed data. The second approach is based on the calculation of a between-session ratio matrix. The difference between the session factors estimated by each of these two approaches depends on the experimental design and the biological variation. However, these differences are negligible compared the effects of biological variation.

METHODS
Sample data
The data in Figure 1A exemplifies a typical data set from a multi-session experiment based on measurements from a larger series of transfections in which the transfection efficiency, the reporter-enzyme assay, and the measurement session are experimental factors that may result in multiplicative between-session variation ². The different DNA constructs in the transfections represent the experimental conditions. Data from groups of transfections on different days make up groups of measurements: the measurement sessions. The multiplicative nature of the between-session variation in this example-data set is apparent from the fact that most session lines connecting the data points run parallel in this graph with a logarithmic Y-axis (Figure 1A).

Mixed additive and multiplicative model
The standard additive model for an experimental design with one measurement session and a number of conditions is given in Equation 3:

\[ Y_i = Y_{\text{mean}} + E_i + \text{error} \]  

This model states that the result of a measurement \( Y \) in condition \( i \) is composed of the population mean (\( Y_{\text{mean}} \)), the effect of condition \( i \) (\( E_i \)), and an experimental error. In this additive model the sum of the condition effects is 0 (\( \sum E_i = 0 \)) and the error is normally distributed with mean 0 and standard deviation \( \sigma \). Note that 'effect' in the sense used here does not represent the difference between a control and an experimental condition, but stands for the effect of each condition relative to the
Removal of between-session variation

population mean, \( Y_{\text{mean}} \). The experimental error reflects the variance within a condition, whereas the condition effects reflect the differences between conditions. In a multi-session experiment with multiplicative between-session variation, this additive model is extended with a multiplicative session-dependent factor (Equation 4).

\[
Y_{ni} = F_n \times (Y_{\text{mean}} + E_i + \text{error})
\]  

Accordingly, for each session \( n \), the measurement results for each condition are multiplied by session factor \( F_n \). In this mixed additive and multiplicative model the product of the session factors equals 1 \( (\prod_{n=1}^{m} F_n = 1) \). This property insures that the overall \( Y_{\text{mean}} \) is not affected by the multiplicative factors. The session factors can be estimated with two different approaches: two-way analysis of variance (ANOVA approach) or calculation of a between-session ratio matrix (Ratio approach). Both approaches will be illustrated with the sample dataset (Figure 1A; panel 1 in Box 1 and Box 2)

**Estimation of the session factors with the ANOVA approach (Box 1)**

The estimation of the session factors from the output of a two-way ANOVA is illustrated in Box 1. The process starts with a logarithmic transformation of the data. This transformation converts the multiplicative session factor into an additive component in the model (Equation 5)

\[
\log(Y_{ni}) = \log(F_n) + \log(Y_{\text{mean}} + E_i + \text{error})
\]  

The application of two-way ANOVA without interaction between the factors session and condition then results in estimated marginal means per session (EMM\(_n\)). Note that the condition effects that would result from the two-way ANOVA procedure on the log-transformed data would be multiplicative effects. Therefore, these effects should be ignored and only the results with respect to the session factors should be used. This ANOVA procedure can be carried out with every statistical package. In this paper SPSS (version 11.5.2) was used. In the example dataset several sessions are incomplete.
The General Linear Model ANOVA procedure of SPSS uses substituted values in the calculations of the marginal means to correct for missing conditions in one or more sessions. Such a missing value substitution is implemented in every statistical package. In short, the missing values are assumed to behave similar to the other values in the same session and the same condition. For an experiment with N sessions and I conditions, the substituted value for the missing observation $Y_{in}$ can be calculated with Equation 6.

$$
\hat{Y}_{in} = \left\{ (Y_n + N Y_{i} - Y_i) / \left\{ (I - 1) (N - 1) \right\} \right. 
$$

In this equation $Y_n$ is the sum of non-missing observations for the session with the missing condition and $Y_i$ the sum for the condition with the missing session; $Y_{i}$ is the sum of all non-missing observations. When more values are missing an iterative procedure is followed and new substitutions are calculated until the residual sum of squares is minimized.

The deviation of each EMM$_n$ from the mean of all EMM$_n$ is an estimate of the Log(F$_n$) term in Equation 5. The resulting estimated session factors (Box1, panel 8) are applied to the measured values and corrected values are obtained (panel 9).
Removal of between-session variation

Estimation of the session factors with the Ratio approach (Box 2)

The calculation steps used to estimate the session factors with the Ratio approach are illustrated in Box 2. For each pair of sessions, e.g. session 5 and 6, the ratio between sessions is calculated for each condition that these sessions have in common (Equation 7 and panel 2); e.g. for session 6 and 5 this ratio is:

\[ \text{between-session ratio}_{65} = \frac{Y_{5i}}{Y_{6i}} = \frac{F_6}{F_5} \left( \frac{Y + E_i + \text{error}}{Y + E_i + \text{error}} \right) \]  \hspace{1cm} (7)

In such a between-session ratio, the normally distributed additive parts of the multiplicative model (Equation 4), which have the same mean and standard deviation, lead to a ratio of 1. The error of such a ratio of normally distributed variables has a Cauchy distribution \(^6\), which implies that, strictly speaking, its mean does not exist. However, the Cauchy distribution has a symmetrical clock shape centred on zero, has a median of zero \(^7\) and, with a more general definition of integration, its mean can also be considered to be zero \(^8\). Therefore, on average, the error in the last term of Equation 5 is zero and the term cancels out. Therefore, the between-session ratio is an unbiased estimate of the ratio of between-session factors.

---

**Box 2: estimation of session factors with between-session Ratios**

1. **conditions**
   - 1
   - 2
   - 3
   - 4
   - 5
   - 6
   - 7
   - 8

2. **session**
   - A
   - B
   - C

3. **ratio between sessions 5 and 6**
   - session 1
   - 2
   - 3
   - 4
   - 5
   - 6
   - 7
   - 8
   - observed between-sessions ratios
   - factor

4. **Substitution of missing ratio: row 1 column 6**
   - column 1
   - 2
   - 3
   - 4
   - 5

   - calculate substitutes: \( R_{6,1} = R_{c,1} \times R_{6,4}/R_{c,4} \)
   - (for all rows and columns with \( \bullet \))
   - calculate geometric mean substitution

---

<table>
<thead>
<tr>
<th>conditions</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>session A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>session B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>session C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>condition</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>session 1</td>
<td>108</td>
<td>175</td>
<td>245</td>
<td>246</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 5</td>
<td>72</td>
<td>35</td>
<td>31</td>
<td>92</td>
<td>115</td>
<td>204</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 6</td>
<td>255</td>
<td>256</td>
<td>257</td>
<td>96</td>
<td>405</td>
<td>800</td>
<td>846</td>
<td>742</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>condition</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>session 1</td>
<td>108</td>
<td>175</td>
<td>245</td>
<td>246</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 5</td>
<td>72</td>
<td>35</td>
<td>31</td>
<td>92</td>
<td>115</td>
<td>204</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 6</td>
<td>255</td>
<td>256</td>
<td>257</td>
<td>96</td>
<td>405</td>
<td>800</td>
<td>846</td>
<td>742</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>condition</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>session 1</td>
<td>108</td>
<td>175</td>
<td>245</td>
<td>246</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 5</td>
<td>72</td>
<td>35</td>
<td>31</td>
<td>92</td>
<td>115</td>
<td>204</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>session 6</td>
<td>255</td>
<td>256</td>
<td>257</td>
<td>96</td>
<td>405</td>
<td>800</td>
<td>846</td>
<td>742</td>
</tr>
</tbody>
</table>

51
When two sessions have more than one condition in common, a between-session ratio is calculated for each matching pair of conditions and, because we are dealing with multiplicative effects, the geometric mean of these ratios is used.

The matrix of between-session ratios for all combinations of sessions is shown in panel 3 of Box 2. This matrix can now be used to calculate estimates for the session factors. Equation 8 gives the geometric mean of a column of observed ratios.

\[
\text{geometric mean column}_i = \sqrt[n]{\prod_{j=1}^{n} \left( \frac{F_i}{F_j} \right)^{n}} = \frac{\prod_{j=1}^{n} F_i}{\prod_{j=1}^{n} F_j} = F_i
\]

Because the multiplicative model requires that the product of all session factors in the denominator of the third term of Equation 8 equals 1, the geometric mean of a column of between-session ratios is an estimate of the correction factor for that session (step 5).

In the example dataset, session 1 and session 6 have no conditions in common and, therefore, a between-session ratio cannot be calculated directly for this pair of sessions. The values 20.0 and 0.05 in the grey boxes in panel 3 (Box 2; upper right and lower left corner, respectively) are in fact substituted ratios. If this substitution of missing between-session ratios had not been carried out, the correction factors of session 1 and 6 would have been 50% higher (0.47) and 20% lower (5.15), respectively. This would result in incomplete removal of the session variation for these sessions. To be able to calculate proper session factors without the loss of data sets like sessions 1 and 6, we implemented a procedure to substitute the missing between-session ratio. This substitution procedure is based on the fact that in every row in the matrix of observed between-session ratios, the factor in the denominator is the same, whereas for every column the factor in the numerator is the same. Therefore, it is possible to calculate a substitute for a missing ratio in column \( j \) and row \( i \) (\( R_{ji} \)) from a known ratio in that column (\( R_{i,n} \)) and two other ratios from these two rows in another column (\( R_{k,i} \) and \( R_{k,n} \), respectively). A substitute for the missing ratio \( R_{ji} \) is then calculated as

\[
R_{ji} = \frac{R_{k,i} \times R_{ji,n}}{R_{k,n}}
\]

When this substitute is calculated for all possible \( R_{k,i}, R_{ji,n}, \) and \( R_{k,n} \), then the geometric mean of all substitutes will be the best estimate of the missing ratio \( R_{ji} \). The inverse value can be substituted for \( R_{ji} \). The complete substitution procedure is illustrated in Box 2, panel 4. Note that as many substitutes as possible are used to reach the most accurate estimate for the missing ratio. However, the between-session ratios of 1,
which are on the diagonal of the between-session ratio matrix, are excluded. The substitute values resulting from these ratios also occur in the remainder of the table.

**Factor correction**

The between-session variation in the original data set can now be removed by dividing each measured value by the corresponding session factor obtained with either the Ratio or the ANOVA approach (Equation 10):

\[
\text{corrected } Y_{ni} = \frac{Y_{ni}}{F_n}
\]  

(10)

The resulting corrected dataset for the ANOVA approach is given in panel 9 of Box 1.

**RESULTS**

Results of the application of factor correction on the example dataset.

The session factors estimated with each of the two approaches are given in Table 1. These factors differ up to 1.3% (Table 1). This difference is too small to be visible in the graph of the corrected dataset which is plotted in Figure 1D. The reduced distance between the session lines in Figure 1D, compared to Figure 1A, shows that the multiplicative between-session variation has been successfully removed. This is also shown by the reduced variation of the conditions after factor correction (Figure 2D). The remaining difference between the session lines (Figure 1D) reflects the non-multiplicative component of the variation between the experimental conditions, which is the error component in the additive model (Equation 3).

Table 1. Comparison of the estimated session factors obtained with the ANOVA and the Ratio approach for the data of the example dataset. The last column gives the ratio between the two estimated factors per session.
Variation and differences of estimated session factors

To determine the effect of the biological error and the absence of conditions in a measurement session on the session factors estimated by each of the two approaches, a series of simulations was carried out. A number of datasets with known session factors but increasing relative error (coefficient of variation, CV, from 0.01 to 0.25), different numbers of observations per session and condition and an increasing number of missing conditions per session was simulated. Session factors were then estimated with both the ANOVA and the Ratio approach. The results of the simulations of a dataset with 7 conditions and 7 sessions are summarized in Figure 3. When the number of observations per condition and session was lowered from 5 to 2, the deviation of the factors from the expected value increased (Figure 3C). A further increase was observed when the coefficient of variation was increased from 0.1 to 0.25 and when the number of conditions per session was decreased from 7 to 3. Only when the condition–session matrix was incomplete, a difference in estimated session factors occurred and reached 3% and 3.5% of the estimated factors at 4 and 3 conditions per session, respectively. The variation of the estimated factors around the expected factors was symmetrical in all simulations and both approaches. An example of such a distribution is given in Figure 3A for the simulation of 3 conditions per session. The factors estimated by the ANOVA approach showed a high correlation ($r=0.984$) with those of the Ratio approach as illustrated by the scatter plot of the logarithms of the deviations of the estimated factors from the expected factors (Figure 3B). The deviations from the expected factor were found to be independent of the magnitude of the expected factor (Figure 3B). When only two conditions are present per session, each condition providing an overlap with one of the other sessions, both factor-estimation approaches resulted in the same factors. However, because of the low number of observations included in this design, the resulting factor estimates deviate strongly from the expected factors (Figure 3C).
Figure 3. Results of the application of the two methods for estimation of session factors on a series of simulated datasets. The datasets used in this figure consisted of 7 conditions (effects: -50, -20, -10, 0, 10, 20, and 50) and 7 sessions (factors: 0.1, 0.2, 0.5, 1, 2, 5, and 10). The number of observations per condition-session combination (n), the coefficient of variation (CV) and the number of conditions per session (C per S) were systematically varied in the simulations. For each input 25 datasets were simulated. A: line plot of the estimated factors for the Ration approach (left) and the ANOVA approach (right) for a simulation with 3 conditions per session. Note that all estimated factors cluster symmetrically about the expected (=input) factor. The error bars on both sides of the graph give the expected factor and the range of the estimated factors. B: Scatter plot of the log-deviation of each point in panel A from its expected value. The markers are given next to panel A. The inset gives the similar graph for a simulation with 5 conditions per session. C: Box and whisker plots (boxes: 25th and 75th percentile, whiskers 10th and 90th percentile) of average the deviation of the estimated factors from the expected factor for each of the simulations. The graphs marked with D show the difference between the factors obtained from the two approaches. Note the increasing deviation when the number of observations per condition-session decreases, when the CV increases and when the number of condition per session (C per S) decreases. Only when not all conditions are present in each session, the factors estimated with two approaches show a difference.
Chapter 2

Application of the factor correction for the correction of transfection data

To characterise the regulatory regions of a gene, different sequences can be cloned upstream of a promoter and a reporter gene in a plasmid. The regulatory effects of such a DNA construct can then be measured by transfection of these constructs into cells. These cells have to be divided into several batches of cells to allow different treatments. Ideally, all constructs that have to be compared are transfected into the same batch of cells within the same transfection session. However, practical restrictions, e.g. the availability of cultured cells and the duration of the experiment, limit the number of transfections that can be conducted at the same day. This results in a two-level multi-session design: measurement sessions on different days, with multiple transfection sessions per measurement (Figure 4). Biological differences between batches of cells (e.g. cell passage number) cause variation between the different measurement sessions. In addition, transient transfection of DNA constructs into cells is subject to large variation in efficiency between transfections. Without correction for these sources of variation the use of these data in statistical tests would result in an increased type II error (false negatives). It is therefore essential to estimate these contributions of both session levels on the variation and remove these from the observed data. According to a mixed multiplicative and additive model (Equation 4), the measured value of a DNA construct can be described by the following equation:

\[
X_{\text{tmch}} = F_m (F_t (X_{\text{mean}} + E_{\text{ch}} + \text{error})) \tag{11}
\]

The value \(X\) of a construct \(c\) with hormone treatment \(h\) in transfection \(t\) on day \(m\) is composed of the population mean \((X_{\text{mean}})\), a combined condition effect \((E_{\text{ch}})\), an experimental error, and two multiplicative session factors: the transfection-efficiency factor \((F_t)\) and the measurement-session factor \((F_m)\). By determining the transfection-efficiency factors within a measurement-session, we can remove the variation in transfection-efficiency without interfering with the measurement-session variation (Figure 4). To enable the removal of the variation in transfection-efficiency, an internal-control plasmid, with its own reporter signal, has to be co-transfected with the test-DNA construct. Because this control plasmid can be assumed not to be affected by the test-DNA construct, the variance in the internal-control reporter signal should be the result of differences in the transfection efficiency between sessions and an experimental or biological error.
Removal of between-session variation

Figure 4. Schematic representation of the two-level multi-session transfection experiment and the application of a two-step factor correction of transfection data. Different samples from a hepatoma FTO-2B cell line are used on different days (the measurement sessions) with different transfections per day (transfection sessions). To allow statistical comparison of transfection data, the multiplicative variation components resulting from different transfection efficiencies and from the different measurement sessions need to be removed. The transfection-session factor ($F_t$) can be estimated from the signal of the co-transfected internal control. After applying these factors to the test construct data, the measurement-session factor ($F_m$) can be determined and applied. The resulting corrected data are now representative for the whole initial cell population. –Dex: no hormone treatment; +Dex: glucocorticoid treatment. Dotted lines represent data input into the session-factor calculation and the application of the session-factors.

The measured value of the internal control ($Y_t$) can therefore be described by the following equation:

$$ Y_t = F_t \cdot (Y_{\text{mean}} + \text{error}) \quad (12) $$

This model states that the control value $Y$ in transfection $t$ is composed of the population mean ($Y_{\text{mean}}$), an experimental or biological error and a transfection-efficiency factor $F_t$. Factor correction can be used to estimate these factors for the different transfection sessions. Because the internal-control and the test construct are co-transfected, the transfection-efficiency correction factor can also be applied to correct the corresponding test-construct values. This procedure is complicated by the fact that in the experiments the effects of hormones on the expression of our
Chapter 2

DNA constructs are tested. Hormones like glucocorticoids affect the expression of many genes, and they may therefore also influence the expression of the internal-control construct. Using the values of untreated as well as hormone-induced cells originating from the same transfection event, would therefore lead to incorrect correction factors. To avoid this, only the values of untreated cells are used to calculate the transfection-session factors.

To allow comparison of data sets from different days, a second correction step is required to remove the effect of the measurement sessions on different days. Because the transfection-efficiency factor that accounts for differences in transfection efficiency ($F_t$) has already been determined and removed in the first correction step, equation 11 can be simplified to:

$$X_{mch} = F_m (X_{mean} + E_{ch} + error)$$  \hspace{1cm} (13)

By combining the parameters that define the construct identity and its treatment into one condition (e.g. construct_2 no_hormone; construct_2 hormone), all data are included in the factor-correction calculations to determine the session factors $F_m$. The data obtained after this second correction step can be used for further statistical analysis (Figure 5).

![Graph](image)

**Figure 5. Removal of variation from transfection data.** Three different DNA constructs were co-transfected with an internal control-reporter construct to FTO-2B hepatoma cells and split into two equal parts. After 24 hours culturing with and without glucocorticoids, reporter-gene activity of the test-constructs and the internal control was measured. These transfections were repeated on different days. A: uncorrected data from these experiments; B: the corrected data according to the two-step procedure described in the text. The Kruskal Wallis analysis (reference 3) of these data did not reveal any statistical difference between the wild type construct C6 and the other constructs in the uncorrected data, while significant differences were found between construct C6 and construct C3 after removal of multiplicative variation components.
Removal of between-session variation

Note that the combination of construct identity and hormone treatment into one condition variable only serves to distinguish all different conditions in the session-factor estimation. In the statistical analysis of the corrected data, construct and hormone treatment have to be treated as two independent factors in a two-way ANOVA test.

**DISCUSSION**

The factor correction procedure proposed in this paper is based on a mixed additive and multiplicative model for the variation observed in multi-session experiments (Equation 4). The results of the sample data set and the application in the transfection experiment show that factor correction effectively removes between-session variation in an incomplete data set. The corrected data set can be used for statistical testing of differences between conditions, because the experimental error is not affected by the factor correction. The only assumption of the method is that the between-session variation is the result of a multiplicative factor working on each session. If this assumption is incorrect and all between-session variation is random, application of factor correction will lead to session factors that are all close to 1 and correction will not affect the data. Also, if part of the between-session variation is non-multiplicative, this variation component remains present after factor correction. This is illustrated by the data of session 4 of the example dataset: the lines connecting the values of session 4 (open diamonds) cross over the other lines, both before (Figure 1A) and after factor correction (Figure 1D).

The two approaches that can be used to estimate the session factors result in factors that differ depending on the sample size, biological variation and experimental design. The normal distribution of the statistical error in the original measurements as well as the Cauchy distribution of the deviates of the ratios, should still insure that in both approaches the substitution procedure is free of bias. The estimated factors in the simulated datasets indeed show a symmetrical distribution around the expected factor values. This distribution becomes wider with increasing biological variation reflecting the effect of the random error on both factor estimates. The 10th and 90th percentile values deviate up to 17% of the expected factors. Compared to this variation, the difference between the two approaches, which starts to occur when missing values and/or missing between-session ratios have to be substituted, are negligible. Even in the very incomplete design of only 3 conditions in each of the seven sessions, the 90th percentile of the difference is only 3.5% of the estimated
factor value. This difference is most probably due to the difference in dealing with missing values in the condition–session matrix. In both approaches this substitution procedure is based on the assumption that the missing value or ratio would have behaved similarly to the observed data. However, because of the limited number of observations, sampling errors will negatively affect the accuracy of this assumption. Noteworthy, the difference between the approaches disappears when only two conditions per session are present in a fully balanced design. However, such a design is not recommended because of the large variation in the factor estimates. Moreover, the ANOVA approach uses substitutes for each missing value in the condition–session matrix whereas the Ratio approach only substitutes missing ratios in the between-session ratio matrix. The latter will occur less often than the former. To improve the factor estimates, it is recommended to avoid missing values and to spread conditions as much as possible over sessions. It should be emphasised that neither approach replaces missing values in the dataset; both only use substitutions for the unbiased estimation of session factors.

After normalisation, standardisation, and factor correction, the pattern of between-condition differences is very similar (Figure 2). However, in normalisation, the control condition has lost its variance and the variance of all other conditions is larger than when factor correction is applied (Figure 1B and 1D). In other words: the variation that is lost in the control condition has been added to the other conditions. In standardisation, each value per session is transformed into a standard value by subtraction of the session mean and division by the session standard deviation. As can be derived from Equation 4, the magnitude of the session standard deviation will be proportional to the session factor and, therefore, division by this session standard deviation will remove between-session variation. In case of an incomplete data set, not all conditions are present in every session, and therefore, both the session mean and the session standard deviation will be biased, which leads to added variability within conditions compared to factor correction (Figure 1C and 1D). In factor correction, both approaches for the estimation of session factors allow the usage of incomplete data sets, and, therefore, factor correction is the preferred method for the removal of between-session variation. Some statistical packages enable the use of estimated parameters in macro programs and with those packages the factors estimated from the marginal means per session can be directly applied to the dataset. However, in most packages the output of the ANOVA procedure cannot be automatically imported into a data transformation procedure. The ANOVA approach is therefore often difficult to use. This is especially the case in biological experiments.
Removal of between-session variation

in which a combination of experimental procedures leads to an accumulation of session effects as in the two-level multi-session transfection experiment. The implementation of the Ratio approach in a program that reads and writes Microsoft Excel datasheets facilitates the removal of session effects from such datasets.

Removal of the multiplicative variation component from the dataset enables the statistical test for the additive condition effect with the standard one-way ANOVA. When one suspects that an additive part of the session variation is still present, the session can be included in a two way ANOVA. Both tests are no longer hindered by the mixed nature of the variation in the original observations.

We propose the use of "factor correction" as an alternative to the normalisation or standardisation procedures for the removal of multiplicative between-session variation. Factor correction performs well with incomplete data sets and does not affect the biological error. It is, therefore, the preferred method for getting rid of a between-session variation resulting from multiplicative factors working on each measurement session.
Chapter 2

REFERENCES


