Additional guidelines for style and units – Representation of dispersion

The clear, cohesive and correct representation of the results of a research paper is a key component of the characteristics that comprise comprehension, quality and reliability of the scientific publishing process.

However, the direct observation of the manuscripts submitted and the papers published by RBZ enlightens the plurality of the forms of exposure of the indicators of significance and dispersion (measures of uncertainty) of the results presented.

The Editorial Board of RBZ understands that the number of particularities in the form of exposing the results is directly proportional to the number of experimental designs and arrangements, as well as the number of statistical methods utilized.

Nevertheless, standard guidelines should and can be adopted by the authors in order to make the manner of exposure of the results more homogeneous. Thus, the guidelines presented below, which comprise the most common situations, must be followed by the authors for the correct establishment of the publishing style of Revista Brasileira de Zootecnia.

1. About the representation of the descriptive levels of probability for type I error (P-value)

Following the international trend of results exposure in research papers, the authors are recommended to present P-values from the statistical analyses to the readers, regardless of the critical level of probability adopted in the manuscript (α value). Whatever methods have been applied will not alter the discussion content at all. However, this makes the presentation of results more clear and allows the reader to make "judgments" on the results if they have a different view from that presented by the authors. Reference notes for significance (e.g., use of asterisks) should be avoided.

It is mandatory that the P-value be presented with three decimal places. It must not be displayed with 2 decimal places, for it can generate ambiguity of interpretation (e.g., let us suppose that one assumes $\alpha = 0.05$. If two variables tested independently present P-values of 0.049 and 0.051, the rounding off for the two decimal places will make a P-value of 0.05 for both; however, one shows significant effect, whereas the other does not.)

2. About the critical level of probability (the α value) adopted in the manuscript and the significance representation throughout the text

For the right discernment between significance and non-significance in hypothesis testing, according to the Neyman-Pearson school there is the need for establishing a (maximum) critical level of probability acceptable for type I error, from which the differences must be assumed as non-significant, most commonly known as "α value". This must be properly exposed at the end of the description of the statistical procedures, because it is part of the methods set for the research paper.

Example: "...$\alpha = 0.05$."

The choice of the α value must be done during the experimental planning, considering the factors inherent to the environment and the experimental material and the natural variability of the response variables to be assessed at the assay. Although the α value refers nominally to control of type I error, it must be pointed out that the probability of occurrence of type I and II errors commonly manifest antagonistically. Therefore, more strict α values (e.g., 0.01) represent a great control of type I error, but may reduce the level of control of type II error. In this way, it is up to the researcher, after the proper experimental considerations, to define the priorities of control of the statistical errors in their conditions and to adopt the pertinent α level.

If an author chose to make assertions about significance or no significance based on the previous choice of α, the indication of significance must agree with that choice. For instance, let us take a study conducted with $\alpha = 0.05$. In this study, the analysis of variance showed a P-value of 0.019. When presenting this to the reader in the text, the author must utilize: "...a difference was observed ($P<0.05$)."

For expressions in the text, use the letter P (capital letter), not in italic and without spaces. Example: "...intake increased ($P<0.05$), but there was no change in weight gain ($P>0.05$)." Additionally, for an RBZ's convention, the symbols ≤ or ≥ must not be used. Use only < or >. Do not use the form "P=0.XX".

The basic theory of hypothesis testing shows us the fact that there are two, and only two, distinct regions under a distribution of probability when this is utilized in the test: acceptance region of H0 and rejection region of H0 (or region of no rejection of H0 and region of no acceptance of H0, as some areas would rather use).

This leads us to the warning about two common mistakes involving the interpretation of significance: the use of the term "tendency" or "trend" and the qualification of significance (according to the Neyman-Pearson school).
To illustrate the first mistake, let us suppose that an author is conducting a research project in whose planning $\alpha = 0.05$. At the analyses, for one of the variables, a P-value of 0.061 was observed. Due to the proximity of this value to the α value, the researcher presents in their text: "...for the X variable there was tendency for difference..."

Considering the summarized idea of tests and hypotheses presented previously, this type of argument is invalid, since there is no region of "tendency for acceptance of H0" or "tendency for rejection of H0". Thus, the value of the statistics calculated can only be included in the regions of "rejection" or "not rejection" of H0. In this sense, the proximity of the value to α does not matter, contrarily to which region the statistics' calculated value suits.

Otherwise, to illustrate the second mistake, let us take a research paper in whose planning $\alpha = 0.05$. In this case, two variables presented at ANOVA, P-values of 0.035 and 0.002. Some may state that the first result is taken as significant, and the second as "highly" significant, which characterizes qualification. Again, there is the warning: the proximity between the values of P and α does not matter. Hence, there are no "little", "very", "highly" or "poorly" significant results, but only significant or non-significant.

However, there is an increasing tendency among authors worldwide to commingle the Fisher school with the Neyman-Pearson school, i.e., to present significance level and compromise statistical precision with body of evidence in rejecting or not rejecting the null hypothesis. The Fisher school is based on body or strength of evidence, which means that the lower the P-value, the stronger the evidence. By body of evidence we mean that for some reason, such as some experimental conditions that could be controlled but were not, or some variable or variables that are known to interfere on treatment effects but were not dealt with for some particular reason (cost, rain, drought, etc.), a researcher is not forced to conclude in favor of the maintenance of the status quo simply because he (she) found P=0.058. Therefore, we strongly suggest the presentation of the confidence intervals because they combine the magnitude of a treatment effect with the statistical precision and, as such, it circumvents the accept-reject dichotomy of the null hypothesis. Confidence intervals move us away from that dichotomy (Stang et al., 2010)[1].

The probability that a continuous random variable equals any one value is ZERO. That's why confidence intervals are built, because instead of making inference about the true value of a parameter, we are now interested in inferring that the true value of the parameter lies within some interval, i.e., the confidence interval. For all practical applications this means that estimates have to be given as the estimate of the mean plus or minus a certain amount (Mood et al., 1974)[2]. Therefore,

$$P\left[\overline{x} - t_{1-\alpha/2}\sqrt{s^2/n} < \mu < \overline{x} + t_{1-\alpha/2}\sqrt{s^2/n}\right] = 0.95$$

means that the probability that the random interval $\left(\overline{x} - t_{1-\alpha/2}\sqrt{s^2/n}, \overline{x} + t_{1-\alpha/2}\sqrt{s^2/n}\right)$ covers the unknown true mean μ equals 0.95. The length of the interval is $2t_{1-\alpha/2}\sqrt{s^2/n}$ and is dependent on sample size (n) and sample variance (s^2). The statistics $t_{1-\alpha/2}$ is some statistics that could be computed from data and on the prior establishment of the significance level(α). Therefore, if authors want to present confidence intervals, they must previously define them. As possible examples we list:
".... the means were presented as $\overline{x}\left(\overline{x} - t_{1-\alpha/2}\sqrt{s^2/n}, \overline{x} + t_{1-\alpha/2}\sqrt{s^2/n}\right)$.";

"... and confidence intervals for the means presented as $\overline{x} + t_{1-\alpha/2}\sqrt{s^2/n}$ ..."

There are statistical softwares that present confidence intervals as outputs, and in such cases, the length of the intervals presented can be calculated as the *upper* minus the *lower* limits of the confidence interval. Therefore, provided that the assumption about the distribution of errors holds true, for a given statistics computed from the data, $t_{1-\alpha/2}\sqrt{s^2/n} = (upper - lower)/2$. For all cases reported above, s^2 = RMS, in which RMS is the residual mean square.

3. Suggestions of styles for the representation of P-values and dispersion indicators in Tables for the most common experimental designs and arrangements[3]

Balanced experiments with qualitative treatments, conducted without the adoption of experimental arrangements, and considering homogeneous variances among treatments

[1] Stang, A.; Poole, C. and Kuss, O. 2010. The ongoing tyranny of statistical significance testing in biomedical research. European Journal of Epidemiology 25:225-230.

[2] Mood, A. M.; Graybill, F. A. and Boes, D. C. 1974. Intoduction to the theory of statistics. McGraw-Hill Kogakusha, LTD., Tokyo.

[3] All the examples herein described are hypothetical. None of them was taken from real experimental situations.

In these situations, this form of table is recommended:

Table 1 - Voluntary intake of animals fed a diet with different energetic sources

Item	Energetic source[1]			P-value	CV (%)
	Alpha	Beta	Gamma		
		kg d^{-1}			
Dry matter	6.301a	5.302b	5.892ab	0.036	5.3
...					
		g/kg of body weight			
Neutral detergent fiber	12.5a	10.4b	11.2b	0.045	4.8

[1] Means in the same row followed by different letters are different by the Tukey test (P<0.05).

In this example, the coefficient of variation (CV) is calculated as:

$$CV\ (\%) = \frac{\sqrt{RMS}}{\overline{Y}} \times 100$$

in which: RMS = residual mean square; and $\overline{Y}$ = overall mean obtained from all the observations.

Although CV is widely adopted in Brazil, there is a trend for its replacement in the international journals by the standard error of the mean. This also shows as reality for the users of PROC MIXED of SAS, which does not compute CV values for ANOVA. If this is the option for the authors, the tables can be put together as:

Table 2 - Total digestibility coefficients (g g^{-1}) of animals fed diets containing different energetic sources

Item	Energetic source[1]			P-value	SEM
	Alpha	Beta	Gamma		
Dry matter	0.605b	0.612b	0.669a	0.0172	0.035
...					

[1] Means in the same row followed by different letters are different by the Tukey test (P<0.05).

The standard error of the mean must be expressed with the same number of decimal places applied to the means, and can be represented in the table by the acronym "SEM" or by the notation $S_{\overline{X}}$. For the specific case of this example, SEM is calculated as:

$$S_{\overline{X}} = \frac{\sqrt{RMS}}{\sqrt{n}}$$

in which: RMS = residual mean square; and n = number of observations in each treatment.

It is important to emphasize that in case of supposition of homogeneous variances among treatments, only one indicator of variance must be presented; the indication of different standard errors to the different treatments is inconsistent with the presuppositions of the analyses.

Balanced experiments balanced with qualitative treatments, conducted without the adoption of experimental arrangements and considering heterogeneous variances among treatments

This type of experimental interpretation has become common with the evolution of the statistical software, especially with the utilization of PROC MIXED, from SAS. In this case, as different variances will be assumed among treatments, each treatment must be followed by its respective indicator of dispersion; in this case, the standard error may be used. Another possibility is to present the associated confidence intervals for treatment means.

Table 3 - Characteristics of the metabolism of nitrogen compounds in animals fed different protein sources

Item	Protein source[1]			P-value
	Omega	Pi	Kapa	
Serum urea nitrogen (mg dL^{-1})	12.35±1.36b	17.18±1.75a	18.54±0.98a	0.023
...				

[1] Means in the same row followed by different letters are different by the Tukey-Kramer test (P<0.05).

We stress that the indicator of dispersion presented in Table 1 is inherent to the treatment's mean (thence the association by the symbol ±). In this case, the standard error is mandatory (standard deviation must not be used). The presentation of the confidence intervals may offer a rather comprehensive data description.

Balanced experiments with quantitative treatments, conducted without the adoption of experimental arrangements and considering homogeneous variances among treatments

The differences between quantitative treatments must not be interpreted by means of conventional tests of multiple comparisons (e.g., Tukey, LSD, Duncan, SNK, Dunnett). Utilize appropriate tests of multiple comparisons (e.g., The Williams test) or utilize regression models (linear or nonlinear).

A common and usually efficient form to interpret can be achieved by performing orthogonal decomposition of the sum of squares for treatments in contrasts associated with the different order effects (e.g., linear, quadratic, cubic, etc.). This decomposition can be done through the adjustment of equation of linear regression corresponding to the highest significant order effect[4].

[4] When fitting the linear regression models, use the notation "r^2" (lowercase) for functions with a single independent variable (e.g., simple linear) and "R^2" (capital letter) for the functions with more than one independent variable or for polynomial models (e.g., quadratic).

In the case of orthogonal decomposition, it must be emphasized that experiments carried out with "p" levels (in the case above, four levels of additive in the diet; p = 4) provide evaluation of "p-1" order effects (in the example, p – 1 = 3; linear, quadratic and cubic).

The adoption of the maxim "models of cubic or superior order do not make sense" must be careful, and in some cases, this can distort the presentation and interpretation of results.

Example:

Table 4 - Performance characteristics of animals fed diets containing different levels of additive

Item	Additive (g kg^{-1} of dry matter)				CV (%)	P-value[1]		
	0	3	6	9		L	Q	C
Intake (g)[2]	125	135	147	152	3.8	0.015	0.225	0.567
...								

[1] L, Q and C - linear, quadratic and cubic effects, concerning the inclusion of additive in the diet.

[2] $\hat{Y} = 125.8 + 3.10 \times X$ ($r^2 = 0.976$).

In some cases where high-degree effects are not significant, one can proceed to its grouping in the interpretation of the experiment as "lack of fit", which can reduce the number of columns in the tables.

Example:

Table 5 - Performance characteristics of animals fed diets containing different levels of additive

Item	Additive (g kg^{-1} of dry matter)					CV (%)	P-value[1,2]		
	0	3	6	9	12		L	Q	LF
Intake (g)[3]	125	135	147	152	161	4.1	0.032	0.359	0.603
...									

[1] L and Q - effects of linear and quadratic order concerning the inclusion of additive in the diet.

[2] LF - lack of fit.

[3] $\hat{Y} = 126.2 + 2{,}966 \times X$ ($r^2 = 0.985$).

One example is shown in Figure 1, which simulates the interpretation of the concentration of rumen ammonia nitrogen as a function of the time after feeding. Observing the points equivalent to the average concentrations obtained in each period, it can be easily seen that the concentration of ammonia nitrogen rises up to the point of highest concentration more intensely than it declines after this point. So, at the interval evaluated, the elevation and reduction of the concentration of ammoniacal nitrogen are asymmetric in relation to the point of maximum concentration. The interpretation of this by a model of second degree (quadratic) implicitly assumes that elevation and reduction happen with the same intensity, i.e., symmetrically in relation to the point of maximum concentration (which ends up distorting the location of the maximum point). In this case, as can be seen in Figure 1, the description is more coherent and logically done by function of the third degree (asymmetric in relation to the maximum point).

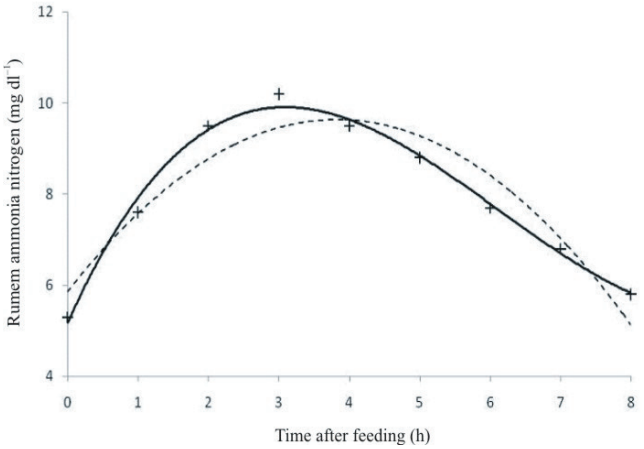

Figure 1 - Concentration of ruminal ammonia nitrogen as a function of the time after feeding (dashed line indicates quadratic function; continuous line indicates cubic function).

Balanced experiments with qualitative treatments, conducted with the adoption of experimental arrangements and considering homogeneous variances among treatments

The adoption of experimental arrangements (e.g., factorial, split plot) is common in experiments in the animal science area, and the information from their application must be adequately exposed to the reader.

As an example, in factorial arrangements the treatments are defined by the combination of the different levels (quantitative or qualitative) of the factors studied. They start to build the aim of studies in terms of their possible interaction or their direct (independent) effects, should they not interact with themselves, on the response variables. Hence, this piece of information (interaction and/or independent effects) must be presented coherently to the reader.

Example:

Table 6 - Voluntary intake in ruminants fed low-quality forage supplemented with nitrogen compounds and/or starch

Item	WN		N		SEM	P-value[1]		
	WS	S	WS	S		N	S	N × S
	g kg^{-1} of body weight							
NDFap	11.2	10.5	12.8	12.0	1.1	0.003	0.046	0.485
...								

WN - without nitrogen compounds; N - with nitrogen compounds; WS -without starch; S - with starch; NDFap - neutral detergent fiber corrected for ash and protein.

[1] N, S and N × S - effects of supplementation with nitrogen compounds, supplementation with starch and their interaction, respectively.