Draft date: 2019.3.29

Statistics Required by International Journals: Use of Their Checklists

<Material provided by>
AMED “International Journals Project”

Contents

Introduction

Nature: Statistical Description Checklist

New England Journal of Medicine: Statistical Description Checklist

Science: Statistical Description Checklist

The EMBO Journal: Statistical Description Checklist

JAMA: Statistical Description Checklist

Websites for Reference

Introduction

This module extracts and paraphrases the points of the statistical checklists that the following international journals recommend to authors of research papers.

[1] Nature
[2] New England Journal of Medicine
[3] Science
[4] The EMBO Journal
[5] JAMA

These checklists cover basic to advanced statistical issues and contain many technical terms that may be unfamiliar to you. You will learn the basics in the subsequent modules. In this module, you are going to grasp the overall picture of the points that each international journal requires.

Learning Objective

Your goal in this module is:

- To understand the overall picture of the descriptions of statistical matters required by international journals.

Nature: Statistical Description Checklist

- Comparisons of interest are clearly defined
- Name of tests applied are clearly stated
- All statistical methods identified unambiguously
- Justification for use of test is given
- Data meet all assumptions of tests applied (with particular attention paid to non-normal data sets or small sample sizes, which should be identified in the text as such)
- Adjustments made for multiple testing is explained
- *n* is reported at the start of the study and for each analysis thereafter
- Sample size calculation (or justification) is given
- Unit of analysis is given for all comparisons

- Alpha level is given for all statistical tests (e.g., less than 5%)
- Tests are clearly identified as one or two-tailed
- Randomization procedures or other ways to eliminate bias in sampling (in particular for experiments involving animals) described
- Actual *P* values are given for primary analyses
- *n* for each data set is clearly stated
- A clearly labelled measure of centre (e.g. mean or median) is given
- A clearly labelled measure of variability (e.g. standard deviation or range) is given
- All numbers following a ± sign are identified as standard errors (s.e.m.) or standard deviations (s.d.)
- Any unusual or complex statistical methods are clearly defined and explained for Nature's wide readership. (Authors are encouraged to use Supplementary Information for long explanations.)
- Any data exclusions are stated and explained
- Any discrepancies in the value of *n* between analyses are clearly explained and justified
- Any method of treatment assignment (randomization, etc.) is explained and justified
- Any data transformations are clearly described and justified
- Error bars are present on all graphs, where applicable.
- All error bars are clearly labelled

New England Journal of Medicine: Statistical Description Checklist

- When comparing outcomes in two or more groups, investigators should use testing procedures that control overall type I error that may be caused in multiple testing. The t-tests are appropriate when data are normally distributed, and the test results are firmly established when a sufficiently large data set is used. If earlier study suggests skewness in a small data set, non-parametric tests should be used.
- Results should be presented with only as much precision as is of scientific value. For example, measures of association, such as odds ratios, should ordinarily be reported to two significant digits.
- Measures of uncertainty of data, such as confidence intervals, should be used consistently. The same applies to figures that illustrate the results.
- Except when one-sided tests are required by study design, such as in noninferiority trials, all reported P values should be two-sided. In general, P values larger than 0.01 should be reported to two decimal places, and those between 0.01 and 0.001 to three decimal places; P values smaller than 0.001 should be reported as P<0.001. Notable

exceptions to this policy include P values arising from the application of stopping rules to the analysis of clinical trials and from genetic-screening studies.

- For tables comparing treatment groups at baseline in a randomized trial, significant differences between or among groups (i.e., $P<0.05$) should be identified in a table footnote and the P value should not be provided in the table.
- In manuscripts that report on randomized clinical trials, authors should provide a flow diagram about the subjects in CONSORT format and all of the relevant information required by the CONSORT checklist. If necessary, all this information should be provided in a supplementary appendix separately from the manuscript. The CONSORT statement, checklist, and flow diagram are available on the CONSORT website.

Science: Statistical Description Checklist

- Authors should describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the results.
- Data pre-processing steps such as transformations, re-coding, re-scaling, normalization, truncation, and handling of below detectable level readings and outliers should be fully described; any removal or modification of data values must be fully acknowledged and justified.
- Descriptive statistics should be presented for variables that are integral to subsequent analyses and interpretation of the study findings. The number of sampled units (N), mean and median, upon which each reported statistic is based must be stated.
- For continuous variables, distributions should be described using graphical displays such as scatterplots, boxplots, or histograms or by reporting measures of central tendency (e.g., mean or median) and dispersion (e.g., SD, interquartile range).
- For continuous variables that are approximately normally distributed, mean and SD are suitable measures for center and dispersion, respectively. For continuous variables with asymmetrical distributions, median and range (or interquartile range) are preferred to mean and SD. All measures of central tendency or dispersion that are used should be identified.
- For very small samples sizes (e.g., $N < 20$), presentation of all data values in tabular format is desirable unless presentation would violate restrictions for privacy or confidentiality for human subjects. Units should be supplied for all measurements.
- Methods used for conducting statistical tests (e.g., t-test, Wilcoxon signed rank test, Wald test of regression coefficient) and for constructing confidence intervals (e.g., normal-based 95% CI: mean ± 2SD, likelihood ratio-based interval) should be clearly

stated. Mention methods used in the Materials and Methods and then provide the individual test name in the figure legend for each experiment.

- The testing level (alpha) and whether one-sided or two-sided testing was used should be reported for each statistical test; typically two-sided testing is appropriate, but if one-sided testing is used its use should be justified.
- Adjustments made to alpha levels (e.g., Bonferroni correction) or other procedure used to account for multiple testing (e.g., false discovery rate control) should be reported.
- When Bayesian analyses are conducted, any assumptions made for prior distributions must be fully described.
- Sufficient information should be supplied to allow readers to judge whether any assumptions necessary for the validity of statistical approaches (e.g., data are normally distributed, survival data are consistent with proportional hazards) have been verified.
- An accounting of missing data values should be provided; if imputed data values are used in statistical analyses, the methods used for imputation should be fully described.
- Any novel or advanced method used in statistical analysis or computational algorithms should be fully described with references and examples so that readers can reproduce the results. At the time of publication, we may request the code or data used for computation as supplementary information.
- Authors should present results in complete and transparent fashion so that stated conclusions are backed by appropriate statistical evaluation and limitations of the study are frankly discussed.
- Point estimates of population parameters (e.g., mean, correlation coefficient, slope) or comparative measures (e.g., mean difference, odds ratio, hazard ratio) should be accompanied by a measure of uncertainty such as a standard error or a confidence interval.
- Results of each statistical test should be reported in full with the value of the test statistic and p-value, and not simply reported as significant or non-significant; more than two significant digits on p-values are usually not needed except in situations of extreme multiple testing such as in genetic association studies where stringent corrections for multiple testing might be used.
- Any results that are reported to constitute a blinded, independent validation of a statistical model must be accompanied by a detailed explanation that includes:
    1) specification of the exact “locked down” form of the model, including all data processing steps, algorithm for calculating the model output, and any cutpoints that might be applied to the model output for final classification;

2) date on which the model or predictor was fully locked down in exactly the form described;
3) name of the individual(s) who maintained the blinded data and oversaw the evaluation (e.g., honest broker); and
4) statement of assurance that no modifications, additions, or exclusion were made to the validation data set from the point at which the model was locked down and that neither the validation data nor any subset of it had ever been used to assess or refine the model being tested.

The EMBO Journal: Statistical Description Checklist

- The description of all reported data that includes statistical testing must state the name of the statistical test, the number (n) of independent experiments underlying each data point, and P values generated from each test. The P values should be reported as actual values, rather than simply reported as statistically significant or not. Discussion of statistical methodology can be reported in the materials and methods section, but figure legends should contain a basic description of n, P and the test applied.
- Descriptive statistics should include a clearly labelled measure of centre (such as the mean or the median), and a clearly labelled measure of variability (such as standard deviation or range). Ranges are more appropriate than standard deviations or standard errors for small data sets. Standard error or confidence interval is appropriate to compare data to a control.
- Graphs must include clearly labelled error bars. Authors must state whether a number that follows the ± sign is a standard error (s.e.m.) or a standard deviation (s.d.)
- Authors must justify the use of a particular test and explain whether their data conform to the assumptions of the tests.
- When using statistical methods based on the normal distribution authors should explain how they tested their data for normality. If the data do not meet the assumptions of the test, then a non-parametric alternative should be used.
- When making multiple statistical comparisons on a single data set, authors should explain what measures they implemented to avoid an inflated Type I error rate.
- For each experiment, the number of both technical and biological replicates should be clearly stated. Biological replicates provide external validity which indicates that the same results can be generated from multiple samples, while technical replicates provide internal validity which indicates that the same results can be generated from the

same samples. In general, observation values generated in technical replicates are averaged before any statistical inference tests are performed.

- In cases where n is small, appropriate statistical tests should be employed and justified in the text.
- For complex biological experiments the number of independent repeats of a measurement often has to be limited for practical reasons. In such a case, statistical measures applied to too small a sample size are not significant and they can suggest a false level of significance. We recommend that the actual individual data from each experiment should be plotted if $n < 5$. In cases where n is small, a justification for the use of the statistical test employed has to be provided. If n is not based on independent experiments (that is, n merely represents replicates of a measurement), statistics may still be useful, but a detailed description of the repeated measurement is required.

JAMA: Statistical Description Checklist

- In the Methods section, describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to reproduce the reported results. Such description should include appropriate references to the original literature, particularly for uncommon statistical methods. For more advanced or novel methods, provide a brief explanation of the methods and appropriate use in the text and consider providing a detailed description in an online supplement.
- In the reporting of results, when possible, quantify findings and present them with appropriate indicators of uncertainty, such as confidence intervals.
- Avoid relying solely on statistical hypothesis testing, such as the use of P values, which fails to convey important quantitative information.
- For observational studies, provide the numbers of observations. For randomized trials, provide the numbers randomized. Report losses to observation or follow up.
- For multivariable models, report all variables included in regression models, and report model diagnostics and overall fit of the model when available.
- Define statistical terms, abbreviations, and symbols, if included. Avoid nontechnical uses of technical terms in statistics, such as correlation, normal, predictor, random, sample, significant, trend. Do not use inappropriate hedge terms such as marginal significance or trend toward significance for results that are not statistically significant (i.e., $P \geqq 0.05$). For observation studies, methods and results should be described in terms of association or correlation and should avoid cause-and-effect wording. Terms

suggesting a causal relationship such as effect can be used for randomized clinical trials.

- For randomized trials, a statement of the power or sample size calculation is required (see the EQUATOR Network CONSORT Guidelines). For observational studies, a power calculation is not generally required when the sample size is fixed. However, if the sample size was determined by the researchers, then there should be some justification for the number sampled. In any case, describe power and sample size calculations at the beginning of the Statistical Methods section.
- It is generally not necessary to provide a detailed description of the methods used to generate summary statistics, but the tests should be briefly noted in the Methods section.
- Identify regression models with more than 1 independent variable as multivariable and regression models with more than 1 dependent variable as multivariate, rather than describing all regression models as multivariate regression. Report all variables included in models, as well as any mathematical transformations (eg, logarithmic transformations) of those variables. Provide the scientific rationale (clinical, statistical, or otherwise) for including variables in regression models.
- For regression models fit to dependent data (eg, clustered or longitudinal data), the models should account for the correlations that arise from clustering. Failure to account for such correlation will result in incorrect estimates of effects or confidence intervals. Describe how the model accounted for correlation of clustered data. For example, for an analysis based on generalized estimating equations (GEE), identify the assumed correlation structure and whether robust (or, sandwich) variance estimators were used. Or, for an analysis based on mixed-effects models, identify the assumed structure for the random effects, such as the level of random intercepts and whether any random slopes were included. Fixed-effects estimation should be described as conditional likelihood. Avoid the term fixed effects for describing covariates.
- Report losses to observation, such as dropouts from a clinical trial or those lost to follow-up or unavailable in an observational study. If some participants are excluded from analyses because of missing or incomplete data, provide a supplementary table that compares the observed characteristics between participants with complete and incomplete data. Consider multiple imputation methods to impute missing data and include an assessment of whether data were missing at random (MAR). Approaches based on “last observation carried forward (LOCF)” should not be used.
- Both randomized and observational studies should identify the primary outcome(s) before the study began, as well as any prespecified secondary, subgroup, and/or

sensitivity analyses. Comparisons arrived at during the course of the analysis or after the study was completed should be identified as post hoc.

- For analyses of more than 1 primary outcome, corrections of P values for multiple testing should generally be used. For secondary outcomes, address corrections for multiple testing or consider such analyses as exploratory and interpret them as hypothesis-generating.
- For randomized clinical trials, protocols with complete statistical analysis plans should be cited in the Methods section and submitted as online supplementary content. Randomized clinical trials should be primarily analyzed according to the intention-to-treat approach. Deviations from strict intention-to-treat analysis should be described as “modified intention-to-treat,” with the modifications clearly described.
- At the end of the Methods section, briefly describe the statistical tests used for the analysis. Also include the statistical software used to perform the analysis, including the version and manufacturer, along with any extension packages. Do not describe software commands (eg, SAS proc mixed was used to fit a linear mixed-effects model). If analysis code is included, it should be placed in the online supplementary content.
- Analyses should follow EQUATOR Reporting Guidelines and be consistent with the protocol and statistical analysis plan, or described as post hoc.
- When possible, present numerical results (eg, frequency and/or rates) with appropriate indicators of uncertainty, such as confidence intervals (eg, measurement errors).
- Use means and standard deviations (SDs) for normally distributed data and medians and ranges or interquartile ranges (IQRs) for data that are not normally distributed.
- Avoid solely reporting the results of statistical hypothesis testing, such as P values, which fail to convey important quantitative information. For most studies, P values should follow the reporting of comparisons of frequencies and differences (eg, 0.8%, 95% CI −0.2% to 1.8%; P = .13). P values should never be presented alone without the data that are being compared.
- Report P values in the following manner: for P values less than .001, report as “P<.001”; for P values between .001 and .01, report the value to the nearest thousandth; for P values greater than or equal to .10, report the value to the nearest hundredth; and for P values greater than .99, report as “P>.99.” For studies with exponentially small P values (eg, genetic association studies), P values may be reported with exponents (eg, $P = 1\times10^{-5}$). In general, there is no need to present the values of test statistics (eg, F statistics or χ^2 results) and degrees of freedom when reporting results.

- For secondary and subgroup analyses, there should be a description of how the potential for type I error (false positive) due to multiple comparisons was handled, for example, by adjustment of the significance threshold. In the absence of some approach, these analyses should generally be described and interpreted as exploratory, as should all post hoc analyses.
- For randomized trials using parallel-group design, there is no validity in conducting hypothesis tests regarding the distribution of baseline covariates between groups. Therefore, this approach should be avoided. Tables of baseline participant characteristics should not include P values. Instead, report clinically meaningful imbalances between groups, along with potential adjustments for those imbalances in multivariable models.

This module was prepared by "Ethics Education Program on the Research Reliability Standards of International Medical Journals" (or "AMED International Journals Project") with funding support from "Research and Development of Material/Program for Research Integrity Education" of the Japan Agency for Medical Research and Development. Please visit the APRIN website for the names of the experts who prepared and/or reviewed this module.

Websites for Reference

[1] Nature
(http://image.sciencenet.cn/olddata/kexue.com.cn/upload/blog/file/2010/12/2010128212513557501.pdf; visited on 2019.01.23)

[2] New England Journal of Medicine
(http://www.nejm.org/page/author-center/manuscript-submission#electronic; visited on 2019.01.23)

[3] Science
(http://www.sciencemag.org/authors/science-editorial-policies; visited on 2019.01.23)

[4] The EMBO Journal
(http://emboj.embopress.org/authorguide#statisticalanalysis ; visited on 2019.01.23)

[5] JAMA
(http://jamanetwork.com/journals/jama/pages/instructions-for-authors; visited on 2019.01.23)