# Checklists and Examples for Registering Statistical Analyses

For well-designed *confirmatory* research, all analysis decisions that could affect the confirmatory results should be planned and registered prior to data collection. These decisions include: the specific statistical test for each confirmatory hypothesis, whether the test is one-sided or two-sided, the criterion for acceptable evidence, any transformations or adjustments to the data, any criteria for excluding or deleting data, and any corrections for multiple analyses. If this information cannot be pre-specified, the research is exploratory rather than confirmatory.

An experiment may include exploratory analyses and/or post hoc analyses as well as confirmatory analyses. Exploratory and post hoc analyses can be adapted as the data are being analyzed, but must be appropriately distinguished from confirmatory analyses. If a statistical test, data transformation, or data exclusion decision that was not pre-registered is used in a final confirmatory analysis, a deviation from the registered analysis occurs and will need to be carefully justified (KPU Registry, 2015).

Checklists and examples are provided below for different types of statistical analyses, including standard classical analyses, resampling (permutation, randomization, bootstrap) analyses, Bayesian analyses, and classification analyses. These checklists and examples are intended for confirmatory analyses, or for fully specified exploratory analyses as described in KPU Registry (2015). A concise summary of key points pertaining to independence of observations for statistical analysis is also presented because this has been a concern for some experiments.

## Standard Classical Hypothesis Tests

Standard classical hypothesis tests determine the probability of obtaining the experimental outcome if the null hypothesis is true. The null hypothesis is rejected if this *p*-value is below a pre-specified criterion. An alternative approach is to reject the null hypothesis if the confidence interval estimated for a parameter does not include the value for the null hypothesis.

### *Checklist*

___ describe the criteria for excluding any data from the analysis;

___ describe any data reduction, transformations, or adjustments to the raw data;

___ specify whether the analysis is one or two sided;

___ specify the *p*-value (e.g., $p \leq .05$) or confidence interval (e.g., 95%) that is considered acceptable evidence, and describe any corrections for multiple analyses;

___ specify the specific statistical method (e.g., ANOVA, t-test, etc.);

___ verify that the dependent or outcome variable and the independent or predictor variables are clearly indicated;

___ verify that the unit of analysis is clearly indicated (e.g., whether the outcome observations for the statistical analysis are the individual trials or are scores for a participant).

### *Examples*

Example 1. To analyze overall psi, a *z*-score binomial test with continuity correction will evaluate whether the overall rate of direct hits for all trials in the experiment is greater than 25%, with significance set at $p \leq .05$ one-tailed.

Example 2. The difference between the two conditions will be analyzed with a two-sample t-test with the mean hit rate for each participant as the unit of analysis and significance set at $p \leq .05$ two-tailed. Trials with invalid responses will be excluded from the analysis. All data will be excluded for any participant with more than 5 invalid responses.

## Resampling Methods (Permutation, Randomization, Bootstrap)

Resampling methods derive *p*-values and confidence intervals from simulations based on the observed data without an assumption that the data conform to a particular theoretical probability distribution. Resampling methods have few theoretical assumptions and therefore usually inspire greater confidence in the results. However, certain key assumptions must be met (see the section below on Dependence in Statistical Analyses).

Permutation and randomization methods were originally developed for randomized experiments that draw inferences about cause and effect, whereas bootstrap methods were originally developed for random-sample surveys that draw inferences about a population. In general, these methods are optimal for the original type of research. Permutation and randomization methods typically simulate the randomness in the experimental design and have few options. Bootstrap methods usually do not simulate the randomness in an experimental design, are more susceptible to bias, and have more options.

For a permutation or randomization test for comparing the difference between two groups, a distribution of outcomes is generated by randomly assigning each observed data point to one of the two groups. For comparison, a common bootstrap strategy for comparing two groups is to transform each data point by subtracting the mean for the group and then randomly sample with replacement from within each group to generate a distribution of outcomes. Other bootstrap strategies could be used.

An exact evaluation of all of possible outcomes is not feasible for most resampling analyses; therefore, resampling methods typically provide *estimates* for p-values and for confidence intervals. The accuracy of an estimate depends on the number of simulated outcomes or replications, and confidence intervals can be obtained for the estimates. Recommendations for the number of simulated outcomes vary widely. Often 5000 are generated, which gives exact binomial 95% confidence intervals for *p*-values of: .044-.056 for $p = .05$, .0074-.0132 for $p = .01$, and.0003-.0023 for $p = .001$.  10,000 or more simulated outcomes provide greater assurance of accuracy to two decimal places (.046-.054 for $p = .05$ and .0081-.0121 for $p = .01$). For permutation tests, the randomization of the data is without replacement. For bootstrap methods, random samples are drawn with replacement. Different writers have different meanings for the term “randomization test,” but the term usually indicates methods for randomized experiments.

### *Checklist*

___ describe the criteria for excluding any data from the analysis;

___ describe any data reduction, transformations, or adjustments to the raw data;

___ specify whether the analysis is one or two sided;

___ specify the *p*-value (e.g., $p \leq .05$) or confidence interval (e.g., 95%) that is considered acceptable evidence, and describe any corrections for multiple analyses;

___ verify that the unit of analysis is clearly indicated (e.g., whether the outcome observations for the statistical analysis are the individual trials or are scores for a participant);

___ specify which variable is randomly sampled or randomly assigned and whether the randomization is without replacement (permutation) or with replacement (bootstrap);

___ specify the number of simulated outcomes or replications that will be done, or the algorithm and criteria for determining the number;

___ specify any constraints on the random sampling of the data, such as sampling within groups;

___ specify the summary measure or statistic that will be used as the overall outcome for each replication or simulated outcome (e.g., sum of ratings for all trials, mean value for the participants, mean standardized score, Pearson correlation coefficient, t-test t-value, etc.)

___ for bootstrap methods, specify whether a simple nonparametric bootstrap will be used, or which modification(s) to reduce bias will be used (e.g., bias-corrected, bias-corrected-and-accelerated, iterated, parametric, studentized, block, etc.).

### *Examples*

Example 1. Statistical significance for a free response experiment will be determined by random permutation of the observed target sequence while keeping the judges' ratings on each trial fixed. 10,000 simulated experiments will be generated. The sum of the ratings for the simulated targets will be the outcome for each simulated experiment. Significance is set as $p \leq .05$ one-tailed.

Example 2. For this reaction time test, a participant's reaction time on a trial is expected to be increased or decreased due to psi. The type of trial (increase or decrease) is randomly selected without feedback to the participant and is not balanced. Trials with reaction times of greater than 2 seconds will be excluded. For each participant, the average reaction time for the "decrease" trials will be subtracted from the average reaction time for the "increase" trials. This difference value for one participant will be the unit of analysis. Statistical significance will be determined by a randomization test that simulates the experiment by randomly assigning "increase" or "decrease" to the observed reaction time for each trial (with replacement and without attempting to match the exact distribution of trial types in the original experiment). Significance is set as $p \leq .05$ two-tailed. The specific *p*-value for the analysis will also be reported. 100,000 simulated experimental outcomes will be generated in order to have reliable *p*-values in the range of .001 two-tailed. The summary measure for each simulated experiment will be the mean of the difference values for all participants. 95% confidence intervals will be derived from a bias-corrected-and-accelerated bootstrap using the observed difference values with 5000 replications.

For example 2, a bootstrap could be used to test hypotheses as well as to generate confidence intervals. However, a randomization test simulating the random assignment in the experiment is less prone to bias and is conceptually more straightforward (but does not easily provide confidence intervals).

Simulations without replacement (permutation) as in Example 1 are usually more conservative and have slightly lower power than simulations with replacement as in Example 2. Arguments can be made for either strategy when the randomization in the original experiment is with replacement (not balanced).

## Bayesian Hypothesis Tests

Bayesian analyses are based on models of the uncertainty in the beliefs in a human mind rather than uncertainty in the physical world. Prior probability distributions represent the beliefs prior to the experiment and have a fundamental role in Bayesian analyses. Bayesian hypothesis tests are usually based on the Bayes factor, but other Bayesian methods can also be used.

### *Checklist*

___ describe the criteria for excluding any data from the analysis;

___ describe any data reduction, transformations, or adjustments to the raw data;

___ specify whether the analysis is one or two sided;

___ specify the software and specific test that will be used for the analysis;

___ specify the prior probability distributions and null model that will be used, including the values for the specific parameters used with the software indicated above;

___ specify the magnitude of the Bayes factor or other outcome measure that will be considered acceptable evidence;

___ verify that the unit of analysis is clearly indicated (e.g., whether the outcome observations for the statistical analysis are the individual trials or are scores for a participant).

### *Examples*

Example 1. A Bayesian hypothesis test will be applied to the overall rate of direct hits for all trials in the experiment. The binomial Bayes factor calculator provided by Rouder (2012) will be used with a uniform prior probability distribution (beta(1,1)), two-sided test, and $P = .25$ for the null hypothesis. A Bayes factor of 3 or greater will be considered acceptable evidence.

Example 2. The difference between the two conditions will be analyzed with a Bayesian two-sample test with the mean hit rate for each participant as the unit of analysis. The ttestBF function in the BayesFactor package developed Morey, Rouder, and Jamil (2014) will be used. This provides a two-sided test with a Cauchy prior distribution. The rscale parameter for the prior will be set to the default of $\sqrt{2}/2$. A Bayes factor of 3 or greater will be considered acceptable evidence. Trials with invalid responses will be excluded from the analysis. All data will be excluded for any participant with more than 5 invalid responses.

## Classification Analysis or Discriminant Analysis

Classification analyses typically use multivariate data to predict or classify the distinct category for an outcome variable. Statistical analysis is based on the accuracy of the predictions or classifications. Linear discriminant analysis is one of the most well-know methods for classification analyses, but other methods are often used.

Classification methods can be used, for example, to investigate whether physiological measures indicate that a participant is precognitively anticipating a random stimulus. The physiological measures preceding a stimulus are used to predict which type of stimulus occurs on a trial.

The criteria for making the predictions are developed with initial *training* or *learning* data. The algorithms for developing the prediction criteria are basically highly optimized post hoc analyses. For confirmatory research, a hypothesis test is based on applying the criteria to different data that were not used in developing the criteria. Given the optimized post hoc nature of the algorithms, it is usually very difficult to evaluate statistical significance when the criteria are applied to the data used to develop the criteria. Clear descriptions of the training process and the application of the criteria to new data are needed when registering confirmatory experiments that involve classification methods.

### *Checklist*

___ describe the variables that will be entered into the classification process, including the timing of the data points relative to the stimuli;

___ describe the criteria for excluding any data from the analysis;

___ describe any data reduction, transformations, or adjustments to the raw data prior to processing by the classification algorithm;

___ provide a brief conceptual description and references for the classification algorithm;

___ specify the amount of training data that will be used to develop the prediction criteria;

___ specify the amount of data that will be used for the confirmatory hypothesis test;

___ clearly state whether the data for the confirmatory hypothesis test contains any of the data used in developing the prediction criteria, and whether the participants are the same for training and for the hypothesis test;

___ specify the specific statistical test that will be used to determine the accuracy of the classifications (e.g., binomial test);

___ verify that the unit of analysis is clearly indicated (e.g., whether the test is based on analysis of the individual trials or on scores for a participant)

___ specify the *p*-value (e.g., $p \leq .05$) or other criterion that is considered acceptable evidence;

___ specify whether the analysis is one or two sided.

### *Examples*

Example 1. The measures for average skin conductance, peak EMG activity, and average heart rate during the 10 seconds preceding the stimulus will be used to predict whether the upcoming stimulus does or does not require the participant to press the button. Linear discriminant analysis will be used with the discriminant function developed from the data for 2 participants with 40 trials each. The discriminant function will be applied to the data for 20 other participants with 40 trials each. Statistical significance will be evaluated with an exact binomial analysis of the overall proportion of correct classifications for all trials, with $P$ =.5 under the null hypothesis. The significance level for the test will be $p \leq .01$ two-tailed. The peak EMG values will have a natural logarithm transformation before entering the discriminant analysis. Trials with the EMG reading going off scale will be considered as movement artifacts and will be excluded from the analysis. [This simplified hypothetical example does not contain information about baseline adjustments and other technical information that would be expected for registering an actual confirmatory experiment.]

## Dependence in Statistical Analyses

It may be useful to clarify issues relating to independence of data for statistical analyses.

Most standard classical statistical methods, including bootstrap, are based on the assumption that each observation or data point for the dependent (outcome) variable is independent from the other observations. Permutation and Bayesian methods are based on the slightly less stringent assumption that the observations are *exchangeable*, which is very similar to independence but allows balanced observations (closed deck or without replacement). Observations that are completely independent are not balanced (are open deck or with replacement), and are exchangeable. Statistical results cannot be assumed to be valid when the assumptions for independence or exchangeability are not met.

Situations with sequential dependencies among observations or with other confounding factors do not comply with these assumptions. Properly designed randomized experiments neutralize most confounding factors; however, sequential dependencies among observations are more difficult to address.

A sequence of responses by one person is prone to sequential dependencies. Humans tend to have habits and expectations that make repeated responses non-random, non-independent, and non-exchangeable. This applies to physiological measures as well as to overt responses. These sequential dependencies are a problem for data analysis when a human response on one trial is the dependent variable and a participant does multiple trials. Dependence problems have long been recognized and addressed with paired t-tests and repeated measures ANOVA. Dependence problems can occur when the participants do not receive immediate feedback on each trial, and can be greatly compounded if feedback is given. Standard methods such as repeated measures ANOVA may not adequately handle cases with immediate feedback.

Dependence problems have been an issue with presentiment and related studies of unconscious precognitive anticipation of random events. Most such studies have had human responses on individual trials as the dependent variable and have multiple trials per participant (Kennedy, 2014). The potential for dependence problems is high because the participants receive feedback on each trial.

The traditional strategy for analyzing parapsychological experiments avoids dependence problems by using as the dependent variable the random targets or random stimuli rather than the human responses (Kennedy, 2014). If properly generated, the random events are independent and/or exchangeable, and eliminate dependence problems when used as the dependent or outcome variable. This strategy can be used with immediate feedback if the targets or stimuli are generated with replacement and the response or prediction for a trial is inalterably fixed before the participant receives feedback for the trial or is derived using only data prior to feedback (Kennedy, 2014). This strategy can be and has been implemented with presentiment type studies, although analyses that have potential dependence problems have been more common.

Another strategy for handling sequential dependence is to use a score or average for a participant as the unit of analysis rather than the outcome of an individual trial. A paired t-test is an example of this strategy. The dependencies are absorbed into the score or average for a participant and should be reflected in the variance for the scores or averages (assuming each participant has a different set of random targets). However, this strategy may not adequately handle cases with immediate feedback.

## References

Kennedy, J.E. (2014). [Letter on methodology for presentiment studies], *Journal of Parapsychology*, 78, 273-274. Available at http://jeksite.org/psi/jp14let.pdf

KPU Registry (2015). Exploratory and confirmatory analyses. Available at http://www.koestler-parapsychology.psy.ed.ac.uk/Documents/explore_confirm.pdf

Return to KPU Registry Home (https://koestlerunit.wordpress.com/study-registry/)

Koestler Parapsychology Unit Registry for Parapsychological Experiments

Version of August 23, 2015