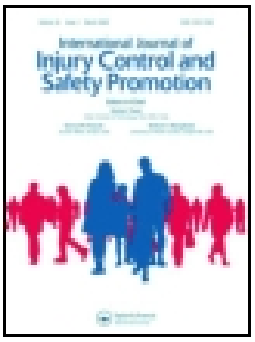


Taylor & Francis
Taylor & Francis Group

International Journal of Injury Control and Safety Promotion

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/nics20

Statistical considerations for research

Shrikant I. Bangdiwala

To cite this article: Shrikant I. Bangdiwala (2021): Statistical considerations for research, International Journal of Injury Control and Safety Promotion, DOI: 10.1080/17457300.2021.1957583

To link to this article: https://doi.org/10.1080/17457300.2021.1957583

Published online: 04 Aug 2021.

INTERNATIONAL JOURNAL OF INJURY CONTROL AND SAFETY PROMOTION
https://doi.org/10.1080/17457300.2021.1957583

ACCIDENTAL NOTE

Statistical considerations for research

Shrikant I. Bangdiwala[a,b,c]

[a]Population Health Research Institute and Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada; [b]Institute for Social and Health Sciences, University of South Africa, Johannesburg, South Africa; [c]Violence, Injury & Peace Research Unit, South Africa Medical Research Council, Tygerberg, South Africa

Introduction

The ultimate objective of research on injuries is to generate knowledge that could be used to reduce trauma and injuries worldwide. Knowledge is generated from interpreting and contextualizing information, while information, both qualitative and quantitative, comes from synthesizing data. The science of statistics provides researchers with the necessary tools to quantify the variability in the observations, as well as identify patterns and relationships in quantitative data.

Many folks think of statistical sciences as a branch of mathematical sciences. The science of statistics relies on mathematical sciences to handle quantitative data, but it is closer to philosophy in that it aims to understand nature and its processes, relationships, characteristics and patterns (see Figure 1). It postulates or hypothesizes these characteristics, abstractly, theoretically, but then desires to verify if these postulates are true. Statistical science accepts that the truth can never be known with certainty. Thus, statistical thinking differs conceptually from mathematical thinking by recognizing that there is inherent variability in nature, and that chance plays an important role in what is observed. Mathematics is a necessary tool for statistics, just like a hammer and a saw are necessary tools for a carpenter. However, the application of statistical thinking and methods helps researchers understand nature and its processes amidst uncertainty. Statistics focuses on quantifications or measurements. Thus, four statistical quantification aspects must be considered - quantifying uncertainty or variability, quantifying probability or chance, quantifying risk and exposure, and quantifying the strength of relationships – all of which are essential elements of injury prevention and control research (Bangdiwala & Banerjee Taylor, 2011).

Figure 1 summarizes the two main processes of the science of statistics, sampling and inference. The trigger for these processes begins with a research question, a desire to describe a characteristic, pattern or relationship in some population of interest. The first statistical process is sampling, whereby a subset from the population of interest (target population) is selected. The resulting sample is studied, and quantitative data are summarized (relevant 'statistics' are calculated). The next step is the process of inference, whereby 'educated guesses' about the population characteristics, patterns or relationships are made based on the observed statistics, including a quantification of the uncertainty around them. Quantifications of characteristics, patterns and relationships in the population are called parameters and labelled using Greek letters; the quantifications of characteristics, patterns and relationships in the sample are called statistics, or estimates of the parameters.

This manuscript presents statistical considerations for research that start with the specification of the research question, and end when the research question's answer is presented to the relevant stakeholders. In the next section, we specifically address 'translation' of the research question specified in non-statistical terms into a statistical question. This leads to statistical aspects of study design, followed by a discussion of the statistical implications from how a study is conducted, to the statistical methodological choices for summarizing the data and interpreting the information in it. Finally, we conclude with methods for 'back-translation' of the statistical results to answer the research question, and presenting the results to the non-statistician stakeholders.

The research question

The research question is the foundation of any research endeavour; it guides the entire research process. Ideally, it is an important question to be answered, one that has not yet been answered. It should be clear, precise, and focussed. In clinical epidemiology, research questions for experimental study designs usually specify the PICOT(S) criteria (Thabane et al., 2009), i.e. the Population being studied, the Intervention and Comparator arms, the Outcomes to be evaluated, and the Timelines and Setting of the study, so as to contextualize it.

The research question can be of different types, either exploratory (e.g. what factors are associated with a particular outcome Y?) or hypothesis-driven (e.g. is factor F associated with a particular outcome Y?). From a statistical perspective, exploratory research questions are ones that lead to descriptive analyses, while hypothesis driven research questions lead to inferential analyses. Descriptive analyses involve

CONTACT Shrikant I. Bangdiwala Shrikant.Bangdiwala@PHRI.CA

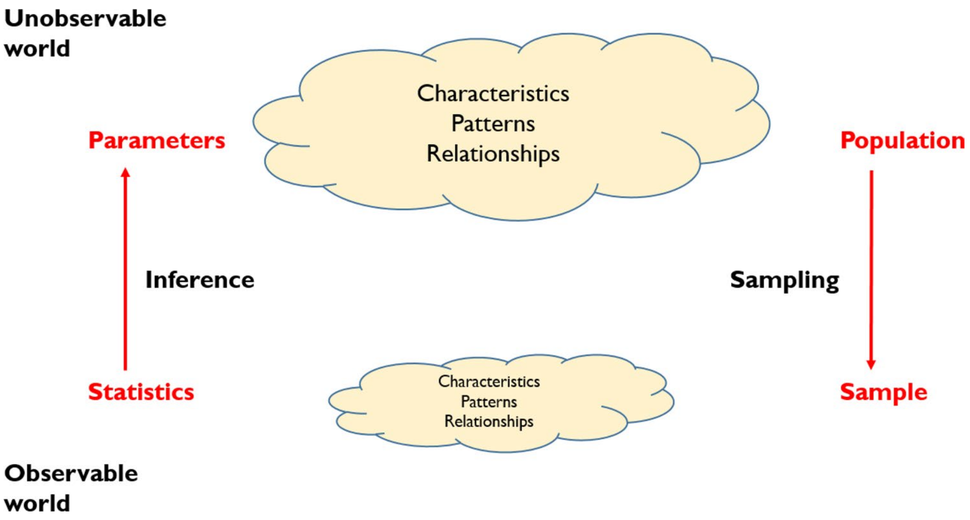


Figure 1. Schematic diagram of key statistical processes – sampling and inference.

estimating parameters in the sample, providing a measure of the variability of the estimate (e.g. standard error), and providing bounds on the uncertainty around those estimates (e.g. confidence intervals, credible intervals). On the other hand, inferential analyses involve taking the information from the sample statistics to make probability statements about the population parameters. This is done either with procedures for formal testing of hypotheses, or with construction of confidence intervals, the set of values of the parameter that one has confidence could be the true value in the population.

Whether exploratory or hypothesis-driven, the research question is 'translated' statistically to help design the study that will enable answering it. Exploratory research questions are 'translated' statistically usually into describing the distributional properties (behaviour) of a variable of interest, or more commonly, into looking at the association between variables, bivariately, or in multivariable regression models. Hypothesis-driven research questions lead to statistical tests of hypotheses.

For example, S. Paul and collaborators wanted to estimate the prevalence and describe the profile of unintentional injuries in children 1–5 years of age in rural India, and explore the potential predictors (Paul et al., 2019) They explored the roles of parent supervisory behaviours and child injury risk-taking behaviours along with other factors using multivariable logistic regression models. Afukaar reviewed the hypothesis that speed is the main cause of road traffic crashes by examining the effectiveness of various speed control measures in Ghana (Afukaar, 2003). He concluded that reducing vehicle speeds could be an effective intervention to reduce traffic crashes in low-income countries, if coupled with strict law enforcement of speed limits or the use of passive speed reduction measures.

There are various kinds of statistical hypotheses – difference (2-sided), superiority or inferiority (1-sided), equivalence, non-inferiority. Table 1 lists example statistical hypotheses for comparing two groups A and B on the distribution of an outcome variable Y. Similarly, research questions dealing with association between two variables X and Y are presented in Table 2.

The statistical formulation of hypothesis testing follows the logic of 'indirect proof', similar to the judicial system, where the hypothesis is called the claim (accusation = alternative hypothesis) and one initially assumes the *status quo* (innocence = null hypothesis). After a review of the evidence, the result is either a decision in favour of the claim (guilty = reject the null hypothesis) or in favour of the *status quo* (not guilty = data are consistent with the null hypothesis) (Bangdiwala, 1989).

An important aspect of formal hypothesis testing is the need to specify numeric thresholds on which to base the judgement of which hypothesis is more consistent with the observed data, the null or the alternative hypotheses. These thresholds are set by the investigator, not the statistician, but they depend on the type of hypothesis, and they impact the necessary sample size for the study. If testing for a difference or for superiority, $\delta > 0$ is the 'smallest clinically meaningful difference' in the outcome that is considered important to detect, in the judgement of the investigators. If testing for non-inferiority, $\Delta > 0$, is the 'largest tolerable difference' in the outcome, in the judgement of the investigator. If testing for equivalence, $\varepsilon > 0$ is the 'largest tolerable margin of error' in the outcome, in the judgement of the investigator. As will be seen below, the smaller the difference between the alternative and the null hypothesis, the larger the necessary sample size. Conceptually, the thresholds are typically such that $\delta > \Delta > \varepsilon > 0$, so that sample sizes for equivalence hypotheses are much larger than for non-inferiority hypotheses, which in turn are larger than for superiority (or difference) hypotheses.

Study design considerations

Once the research question is 'translated' statistically, the next step is to determine the type of study design that best

Table 1. Hypothesis driven research questions and corresponding statistical hypotheses for formal testing when comparing two groups (A and B) with respect to outcome Y (assume higher values of Y are 'better').

Research question in layman's terms	Research question in statistical terms	What type of test is it called?	Corresponding null and alternative hypotheses
Are A and B different in Y?	Is the distribution of Y in A different of the distribution of Y in B?	2-sided test of superiority	H_0: $Y_A - Y_B = 0$ H_1: $Y_A - Y_B = \delta_{2s} \neq 0$
Are A and B equal in Y?	Is the distribution of Y in A equivalent to the distribution of Y in B?	Test of equivalence	H_0: $\lvert Y_A - Y_B \rvert \geq \varepsilon$ H_1: $\lvert Y_A - Y_B \rvert < \varepsilon$
Is A better than B? [*the complementary RQ 'Is B better than A?' is treated similarly*]	Is the distribution of Y in A shifted positively from the distribution of Y in B?	1-sided test of superiority	H_0: $Y_A - Y_B = 0$ H_1: $Y_A - Y_B > \delta_{1s} > 0$
Is A not worse than B? [*the complementary RQ 'Is B not worse than A?' is treated similarly*]	Is the distribution of Y in A not shifted to far negatively from the distribution of Y in B?	Test of non-inferiority	H_0: $Y_A - Y_B \leq -\Delta$ H_1: $Y_A - Y_B > -\Delta$, where $\Delta > 0$

Table 2. Hypothesis driven research questions and corresponding statistical hypotheses for formal testing of the association between variable X (exposure or independent variable) and Y (outcome or dependent variable).

Research question in layman's terms	Research question in statistical terms	What type of test is it called?	Corresponding null and alternative hypotheses
Are X and Y associated?	Are X and Y independent?	Test of association	H_0: measure of association = 0 H_1: measure of association = $\rho \neq 0$
Does X predict Y?	Does knowing the value of X provide some information on the likely value of Y?	Test of association	H_0: measure of association = 0 H_1: measure of association = $\rho \neq 0$
Does X cause Y?	Do values of X determine the likely values of Y?	Test of association	H_0: measure of association = 0 H_1: measure of association = $\rho \neq 0$

addresses the research question. Study designs can be either observational or experimental. Typical ones in the injury field are mainly observational, such as case-studies (e.g. biomechanics crash analysis), case-series (e.g. black-spot analyses), cross-sectional surveys, retrospective case-control studies, case-crossover studies, or prospective longitudinal cohort studies. Experimental studies are ones where the investigator manipulates some exposure (intervention) factor, and prospectively observes the outcome.

Selecting the individuals for a study is a crucial aspect of study design. In observational studies, the target population is defined, and probability samples are drawn from it (see Figure 2). The sampling process aims to obtain a sample in an unbiased manner, and one that is 'representative' of the target population. Representative means that the actual sample obtained is similar to the population in all characteristics that are meaningful. Figure 2 illustrates the role of chance in the sampling process.

Samples should be obtained in an unbiased, objective manner, and be representative of the desired target population. For objectivity, we rely on chance probability for selection – if equal for all (simple random sample), then chance determines who is in our sample. We may want to modify the selection probabilities in order to favour certain segments of the population (probability random sample), but chance is still deciding who is selected in the specific segments.

Figure 2 illustrates the concept of representativeness in samples, as well as the role of chance. In Population A, all individuals are exactly of the same size and pattern; i.e. there is no variability. A sample of size n = 1 is sufficient to get a representative sample of the population. In Population B, the distribution of size is 4/6 big, 2/6 small, while the distribution of pattern is 2/6 solid, 3/6 spotted, 1/6 lined. If we take a random sample of size n = 3, there are $C_3^6 = \dfrac{6!}{3!{*}3!} = 20$ equally likely random samples. Samples A, B, C, and D are four out of the possible 20 equally likely random samples one could get. Sample A is representative of size and of solid pattern, while Sample B is representative of solid pattern only. Sample C is representative of size and of solid pattern as well, while Sample D is not representative neither of size nor of any pattern. Note that Sample B covers all possible patterns, so one could say it has 'good pattern coverage;' but it is not representative of the distribution of patterns in the population. Similarly, Sample D has included all spotted individuals, so it has perfect inclusion of spotted individuals; but it is not representative of patterns. Since we only draw one random sample, chance decides which of the 20 possible samples we could have gotten, and we have to understand the possibility of it not being representative of what we may want it to be; that is the uncertainty from sampling one must deal with.

Representativeness is how similar the sample is to the target population, and chance alone may not work in reaching this goal, especially if sample size is low. Size matters, since the laws of probability help in large samples and not in small samples – bigger is better to improve the chances for representativeness. Another way to do it is to stratify the selection so as to ensure representativeness in certain parameters (e.g. age group distribution, sex distribution, SES distribution.)

A sample needs to always be understood as a subset of the target population, and the fact that we have not observed the entire population gives rise to 'sampling error', a source of uncertainty in our results. We can minimize it by taking larger samples, but we can never eliminate it totally, unless we study the entire population. We must just be aware of it in our interpretation of results.

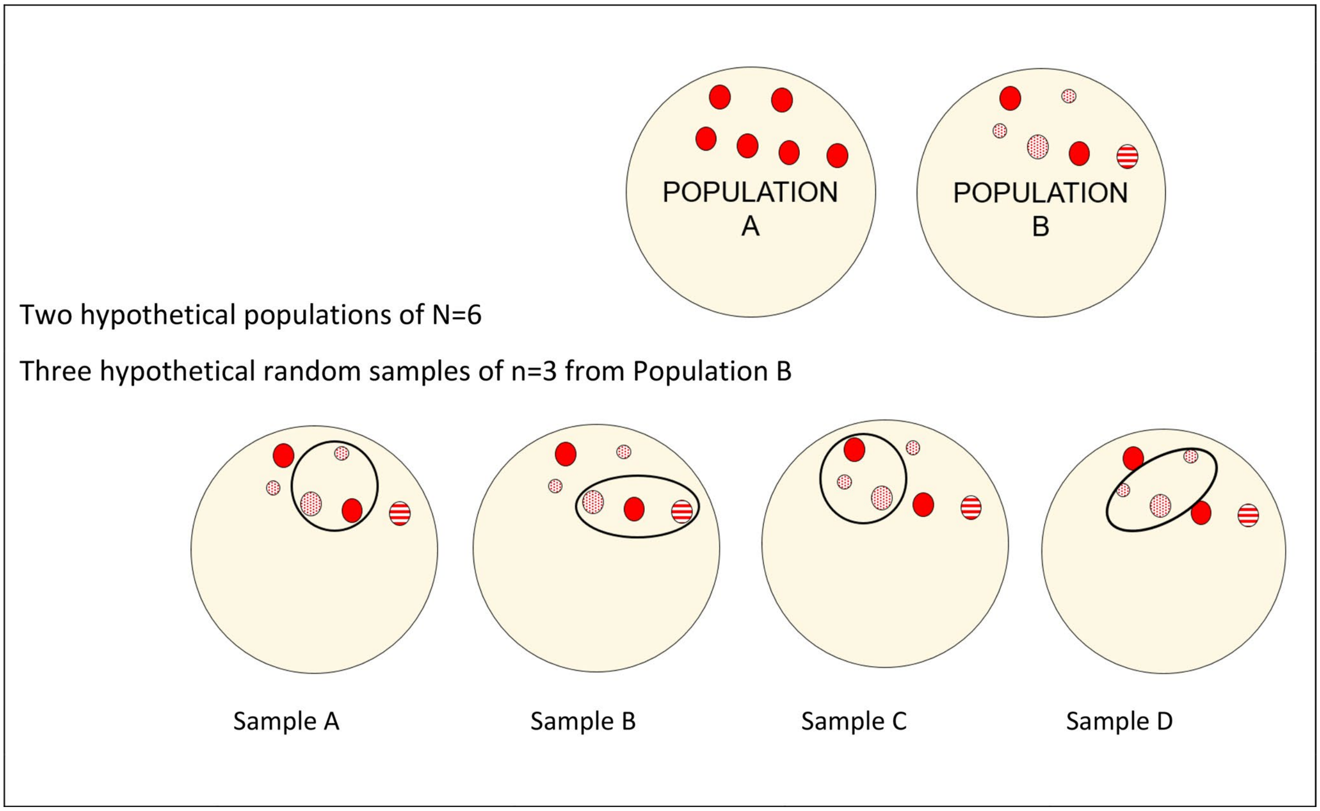

POPULATION A
POPULATION B
Two hypothetical populations of N=6
Three hypothetical random samples of n=3 from Population B
Sample A
Sample B
Sample C
Sample D

Figure 2. Diagram to illustrate the role of chance when drawing a random sample from a population, and the concept of 'representativeness'.

In experimental studies, the study participants are not selected through a random or chance-based probability sampling process. Strict eligibility inclusion and exclusion criteria define the target population, and eligible individuals that conveniently are approached and consent to participate are then recruited into the study. The process of allocating them to the various intervention arm is what now must be objective, unbiased and chance dependent. We use the same word 'random' for this process of randomization. Only with this process do we have the 'statistical endorsement' that the evaluation of the effectiveness of the interventions is valid. Similar to the sampling process in observational studies, the randomization process in experimental studies can be simple, stratified, cluster-based, multi-stage or complex.

Another important aspect of designing any study, is determining how many individuals to study. From a statistical standpoint, the more people studied the better, since larger sample sizes have a better chance of being representative of the population. However, from a practical standpoint, more people studied is most costly, so that some criteria are needed to decide the 'necessary' sample size. In exploratory research questions, the necessary sample size is the one that will give the investigator a desired precision for the estimated unknown population parameter, while in hypothesis driven research questions, the necessary sample size is the one that will give the investigator the desired power to reject the null hypothesis when the alternative hypothesis is the correct one.

There are several important considerations that enter into the calculations of necessary sample size when formally testing a hypothesis – (i) the highest probability of falsely rejecting the null hypothesis that the investigator is willing to tolerate (called the level of significance of a statistical test); (ii) the lowest probability of correctly rejecting the null hypothesis that the investigator wishes to have (called the power of a statistical test); (iii) the difference between the null and the alternative hypotheses values for the parameter of interest; and (iv) the variance of the statistic. Conceptually, these are related to the sample size n as follows:

$$n_{testing} = \frac{(desired\ power)(variability)}{(desired\ significance)(difference)^2}$$

If the research question is not hypothesis-driven, the interest is in estimating rather than testing. In this case, the considerations that enter into the calculations for the necessary sample size are slightly different – (i) the desired confidence level for the uncertainty bounds around the estimate; (ii) the width of the confidence interval (i.e. the precision desired for the estimate); and (iii) the variance of the statistic.

$$n_{estimating} = \frac{(desired\ confidence)(variability)}{(width\ of\ confidence\ interval)^2}$$

Regardless of whether planning an observational or an experimental study, understanding the sources of variability is crucial for controlling their impact on the study results. Recognizing the presence of variability is one of the key, fundamental concepts of the science of statistics. The term 'variable' for measurements or assessments of factors, recognizes the inherent variability in them. If there is no variability in a process or in a group, there is no role for the science of statistics. If there is high variability (i.e. considerable heterogeneity among the observations), it may not be possible to estimate a parameter with precision, or to quantify an association that is small, or to detect a meaningful difference when comparing two or more groups. One can think of variability as the 'noise' in the system, and the measure of association or intervention effect as the 'signal.' If the noise is high, it is difficult to detect signals; they must be quite large in order to be detected. Conversely, if the variability is low, it is easier to detect signals, even low ones.

At the design stage, study investigators and statisticians jointly develop systems to minimize the variability in the study (see Table 3a). Criteria are set to homogenize as much as possible the characteristics of study participants, a common protocol is set so as to implement uniform procedures for all study and data related processes, training of staff and calibration of instrumentation is performed. All these processes are attempts to reduce the variability, increase precision of estimates and ensure high quality, accurate (unbiased) data.

Another important aspect of a study's design is defining the variables to be measured/observed. The Haddon matrix (Haddon, 1970; Runyan, 1998) is a useful tool for identifying potential risk factors to consider in a study of injuries, either as exposure factors, as potential confounders, or effect modifiers of the effects of interest. From a statistical standpoint, the scale of measurement – whether on a continuous numeric scale, a numeric discrete count variable, an ordinal discrete scale, a nominal categorical scale, or a binary scale, will determine the behavior of the variable, and the possible distributional properties that may be relevant when analyzing the data.

In all studies, there are many other measurable and unmeasurable factors that are not under the control of the investigator, and may introduce uncertainty in the study results. If measured, they can be controlled at the analysis stage using multivariable techniques. In experimental studies, the investigator has the possibility of controlling who receives what exposure, through the process of random allocation of the intervention, so that the key primary hypothesis-driven research question can be validly answered. If a large number of units/individuals are randomized, one can more-or-less safely assume that the other potential confounding variables are distributed similarly among the treatment arms, and so they need not be adjusted for in the analyses.

La and collaborators estimated the prevalence and studied which factors may be associated with road traffic crashes among bus drivers in Hanoi, Vietnam (La et al., 2013). They used multivariable logistic regression models to determine which factors assessed on bus drivers were related to the crash prevalence in their observational study. Wijlhuizen and collaborators conducted a multifactor community intervention in The Netherlands to reduce falls in older persons (Wijlhuizen et al., 2007). Because of differences in characteristics of participants in the intervention and control communities at baseline, they needed to use logistic regressions that included those variables in the models.

Study conduct

Once the sample is obtained, measurement of the variables (independent factors, outcomes) takes place. This is a feature of research that is usually not the purview of statisticians, but all statistical analyses will be impacted by the quality of this process. Biases can creep in in multiple ways. Self-selection bias as some individuals in the sample may be non-participants, or responder-bias from some individuals choosing to not respond to particular questions. Measurement error from imprecision of the measurement tools, whether devices (e.g. uncalibrated laboratories; non-standardized weight scales) or psychological scales (e.g. inventories to measure depression) can be minimized by standardization of methods, calibration. and training of personnel. Interviewer bias or responder bias are other sources of uncertainty in the data.

As far as the investigators are concerned, the bulk of the work in a research study is the conduct of the study, i.e. the actual selection of study participants, following them, collecting the data, managing the data. These are essential and necessary steps, but ones in which the statistician takes a back seat. After being closely involved in the planning of the study, the actual conduct or execution is left to study investigators and study staff. The progress of all studies is monitored internally by the study investigators, and experimental studies are often further monitored by external 'data and safety monitoring committees.' The statistician may be involved in preparing reports and presenting them to the monitors, but is not directly involved in the aspects of conducting the study. When the study is being conducted, investigators and statisticians are jointly involved in monitoring the study processes, to ensure that the study integrity is maintained as designed (see Table 3b).

The conduct of the study has major implications for the statistical analysis. The statistician depends on getting high quality data. If the data are collected with imprecision or inaccuracy, the resulting analyses will be imprecise and inaccurate. Imprecision implies higher variability ('noise'), which makes it harder to detect important signals. Inaccuracy implies the data are biased, and methods to deal with it are quite cumbersome, including trying to measure the bias in order to correct/adjust for it in the analyses. A major challenge is incomplete and missing data, which not only reduce the number of observations, but can lead to potential biased results if missing in a non-random way (MNAR). The missingness mechanism is important, and the hope is that, if missing, it be missing completely at random (MCAR), or at least, missing at random (MAR).

Table 3a. Statistical considerations of the design of observational and experimental studies.

	Observational study	Experimental study
Type of study	• Retrospective, Cross-sectional; Prospective	• Prospective
How to study	• Broadly or narrowly focussed • Unadjusted or adjusted models	• Parallel arms, factorial designs, cross-over designs • Who will be aware of the treatment assignment (i.e. who are to be blinded/masked?)
Who to study	• Eligibility criteria	• Eligibility criteria
Number of individuals	• Necessary sample size for adequate precision if research question is exploratory • Necessary sample size for adequate power if research question is hypothesis driven	• Necessary sample size for adequate power since research question is hypothesis driven
Selection of study participants	• Sampling process (probability samples)	• Recruitment process • Process for allocation to intervention (randomization)
What to study	• Estimation of distributional or association parameters	• Estimation of effect of different intervention and control arms
Selection of variables to study	• Relevant outcomes and exposure variables, and their definitions • Possible confounders, mediators and moderators • How to measure/assess variables • How often to measure/assess variables	• Outcome variables – primary, secondary, tertiary – and their definitions • Possible confounders and effect modifiers • How to measure/assess variables • How often to measure/assess variables
How to collect data	• Primary or secondary data collection • Data capture systems	• Primary or secondary data collection • Data capture systems
How to manage data	• Develop data management plan (DMP) – processing, editing, storage, back-up	• Develop data management plan (DMP) – processing, editing, storage, back-up
How to ensure study integrity	• Finalize the study protocol, manuals of operations • Develop quality assurance plan (QAP) – staff training & certification, instrumentation calibration, internal data monitoring plan • Consider external advisory/monitoring boards • Receive approval from ethics boards	• Finalize and register the study protocol and draft statistical analysis plan (SAP) • Finalize manual of operations • Develop quality assurance plan (QAP) – staff training & certification, instrumentation calibration, internal data monitoring plan • Establish an external Data and Safety Monitoring Board (DSMB) • Develop the DSMB Charter, which states membership, their roles and responsibilities, frequency and content of meetings, and statistical analysis plans for interim analyses of safety and efficacy • Receive approval from ethics boards

Methods to complete the data, such as multiple imputation, may be necessary.

Study analysis

The role of statistics in the analysis stage of studies is well known. What is not well recognized is that there is not one and only one unique way to analyze data to answer the research question; there are many ways. Non-statisticians like to have ‘cookbooks’ of what analytical method to use for what situation. The methods provided in such cookbooks are usually the most common way, often due to historical or convenience reasons. What is important is to acknowledge the possibility of alternative approaches to analysis, being transparent, and providing some justification for the method chosen.

All methods are based on several conditions being met, the ‘model assumptions’. ‘Model’ is a broad term, encompassing probability models (i.e. the shape and properties of the distribution of possible values of a variable and their likelihood in a population) and relationship models (e.g. regression models, classification models). No real data fits perfectly any model; it may approximately fit it. There are ‘goodness of fit’ tests to assess whether a specific variable fits a certain type of probability model like the bell-shaped Gaussian distribution or the skewed exponential distribution, and there are methods to assess whether a particular relationship model fits the observed data well or not (e.g. analyses of residuals).

All models are approximations (simplifications) of reality. A quote attributed to British statistician George E. P. Box is, ‘*all models are wrong; but some are useful*’ (Box, 1976). This highlights the fact that all models are approximations to reality. He is also attributed to saying that ‘*there never was, or ever will be, an exactly normal distribution or an exact linear relationship*’ (Box & Luceño, 1997).

The construction of the models needs to be transparent. There are many ways to build a model. If you ask two different statisticians to answer the same research question with the same data, you are likely to get at least five different models! Using Box’s criterion, all of them are wrong; hopefully one of those is more useful.

Models can be characterized as either explanatory (association models) or predictive models. Explanatory models are ones interested in identifying variables that have a scientifically meaningful and statistically significant relationship with an outcome, and their focus is that the model ‘makes sense’. The goal in predictive models is to use the associations between predictors and the outcome variable to generate good predictions for future outcomes; i.e. predictive accuracy; their focus is on ‘good fit’ and less care is placed on the predictors, which may not have any theoretical value, nor statistical significance or scientific meaning.

All models are built with different criteria: parsimony, inclusivity, statistical ‘goodness of fit’ (focus on variance explained and significance testing criteria), or epidemiological ‘understanding of the effect of the exposure’ (focus on modifiers or confounders of the relationship between

Table 3b. Statistical considerations for the conduct of observational and experimental studies.

	Observational study	Experimental study
Internal monitoring for study integrity	• Oversee interviewers, data collectors • Consider re-training of staff, re-calibration of instrumentation • Verify data completeness, timeliness, accuracy and precision • Verify compliance of staff with study protocol • Distributed and centralized statistical monitoring of data if a multicenter study • Frequent and periodic reports to study sites and to study Steering Committee	• Oversee interviewers, data collectors • Consider re-training of staff, re-calibration of instrumentation • Verify data completeness, timeliness, accuracy and precision • Verify compliance of staff with study protocol, focused on intervention dispensing • Verify compliance with randomization procedures • Verify adherence of participants with intervention regimens • Verify whether blinding/masking is maintained as per protocol • Monitor overall adverse event rates • Distributed and centralized statistical monitoring of data if a multicenter study • Frequent and periodic reports to study sites and to study Steering Committee
External monitoring for study integrity	• Optional	• DSMB meetings

exposure and outcome). The purpose may be to confirm contextual theories and thus focus on making sense of all variables, or in predictive accuracy and thus focus on 'discovering' relationships.

Approaches to building models are guided by the criteria, and include ones that let the software think for you (e.g. '*stepwise*' – *forward, backward, 'all possible models')* or ones that involve the careful implementation of criteria by the investigator. Often, unfortunately, the model building approach is not pre-specified, and there is the danger of 'torturing the data so it confesses to what you want', which recently is called 'p-hacking' (Streiner, 2018).

The clinical epidemiology approach to model building is focussed on the relationship between an exposure and the outcome, and considers including other variables that may modify or confound that relationship. Other approaches look at whether variables may moderate or mediate a hypothesized causal association. It is considered preferable that association models be constructed with the model builder 'managing' the process as opposed to letting the computer manage the process. The opposite is preferred for predictive models. In either case, one should explain the method of model building, i.e. be transparent of what was done. After any modelling exercise, one should verify the assumptions of your specific type of regression model, assess the goodness of the fit, and interpret the results.

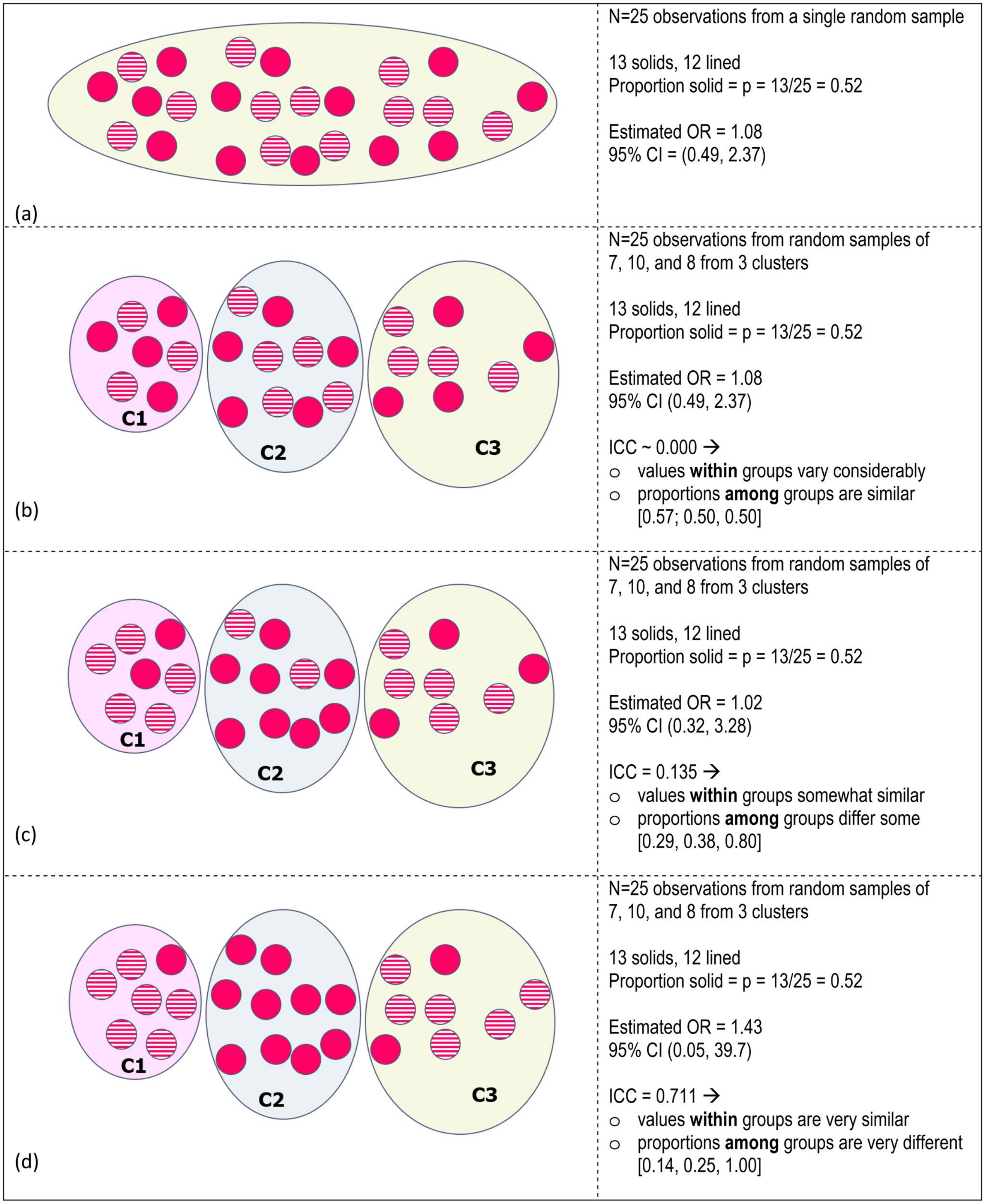
One could also think of any statistical analysis as a painting; the study design is the canvas, the colours are the data, the brushes are the methods, the research question is what the artist was commissioned to do. The statistician has the 'freedom' of an artist, to answer the research question applying the different methods of choice, with the data available. Modelling is an art, and all statisticians are artists.

The 'freedom' to analyze should not be abused. There are fundamental assumptions in most methods, that if substantially violated, will lead to misrepresentation of the data, incorrect results, and incorrect interpretations. If data are biased or inaccurate, the results of analyses will be biased or inaccurate. If data are of poor quality or imprecise, the results of the analyses may be inconclusive. If data are missing or incomplete, the results may be misleading. There are methods to try to overcome such problems with the data, but they are based on meeting other assumptions. Again, being transparent and acknowledging the statistical limitations in the data, but also in the methods chosen for analysis, is important.

Just like the research question determines the study design, the study design determines the analysis method. Most standard methods in statistical analysis depend on sampling processes or randomization processes that ensure the observations can be treated as 'statistically independent,' i.e. uncorrelated. However, some study designs – e.g. multi-stage sample surveys, clustered randomized trials, repeated measures longitudinal studies – produce hierarchical data, and such data are correlated. Figure 3 illustrates the impact of correlated hierarchical data on estimates and the uncertainty around them. In Figure 3a, we have 25 observations from a random sample, and the odds of being solid versus lined is 1.08 (95% CI = 0.49, 2.37). In Figure 3b–d, the total variability remains the same, but we assume that the 25 observations have come from random samples from 3 clusters. The odds of being solid versus lined is exactly the same in Figure 3b since the intracluster correlation coefficient (ICC) is essentially zero; the differences among clusters is small; all the variability is within clusters. In Figure 3c, there are differences among clusters, and the ICC = 0.135. We note that the odds have changed to 1.02 and the 95% CI (0.32, 3.28) is slightly wider. In Figure 3d, the differences among clusters are very large, the variability within clusters in diminished, and the ICC = 0.711 is quite large. The odds now are dramatically different (1.43) and the 95% CI is quite wide (0.05, 39.7). Thus, when the sampling is hierarchical, one must not ignore the correlation structure among the observations (Fisher, 1919).

Interpretation and dissemination of study results

The role of the statistician does not end with the completion of the statistical analyses; it is critical in the interpretation of research results. It is quite common to leave the interpretation of results to the non-statistician investigators. However, they may not be as familiar with the many



N=25 observations from a single random sample

13 solids, 12 lined
Proportion solid = p = 13/25 = 0.52

Estimated OR = 1.08
95% CI = (0.49, 2.37)

N=25 observations from random samples of 7, 10, and 8 from 3 clusters

13 solids, 12 lined
Proportion solid = p = 13/25 = 0.52

Estimated OR = 1.08
95% CI (0.49, 2.37)

ICC ~ 0.000 →

- values **within** groups vary considerably
- proportions **among** groups are similar [0.57; 0.50, 0.50]

N=25 observations from random samples of 7, 10, and 8 from 3 clusters

13 solids, 12 lined
Proportion solid = p = 13/25 = 0.52

Estimated OR = 1.02
95% CI (0.32, 3.28)

ICC = 0.135 →

- values **within** groups somewhat similar
- proportions **among** groups differ some [0.29, 0.38, 0.80]

N=25 observations from random samples of 7, 10, and 8 from 3 clusters

13 solids, 12 lined
Proportion solid = p = 13/25 = 0.52

Estimated OR = 1.43
95% CI (0.05, 39.7)

ICC = 0.711 →

- values **within** groups are very similar
- proportions **among** groups are very different [0.14, 0.25, 1.00]

Figure 3. Illustration of the impact of ignoring the correlation among observations within clusters: (a) Random sample with OR = 1.08 (95% CI 0.49–2.37); (b) Clustered random sample with ICC~0.000, OR = 1.08 (95% CI 0.49–2.37); (c) Clustered random sample with ICC = 0.135, OR = 1.02 (95% CI 0.32–3.28); (d) Clustered random sample with ICC = 0.711, OR = 1.43 (95% CI 0.05–39.7)..

nuances of the methodologies, or of the impact of departures from model assumptions. For example, issues like the effect of influential observations, the potential for bias from incomplete or missing data, collinearity among independent variables in linear regression models, small frequency counts in categorical data may lead to imprecise estimates, and lack of model fit to parametric probability models. Some slight departures to model assumptions may be acceptable if the method is 'robust' or if the sample size is 'sufficiently large.' In the injury field, we commonly study the extreme situations, and this can lead to estimates affected by 'regression to the mean.' The statistician's input is essential for proper interpretation of such 'statistical limitations' and their impact on study results.

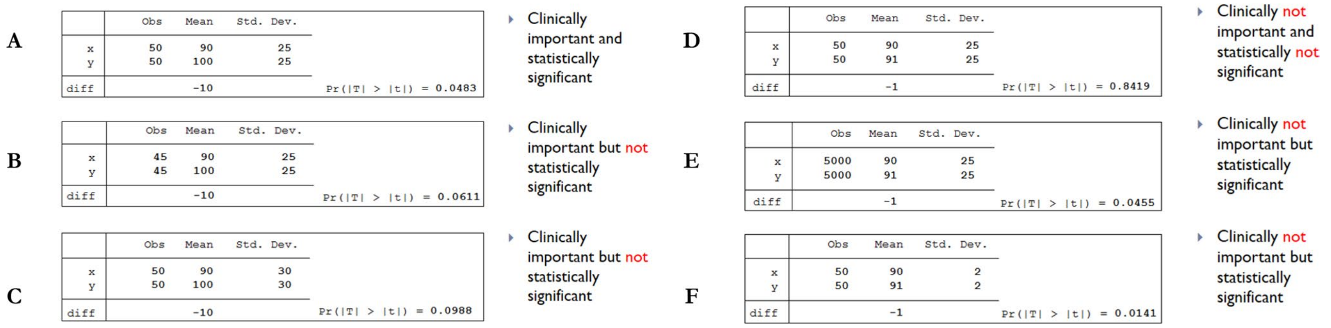
The classic mistake of many investigators is their wrong interpretation of statistical significance as if it were the same as clinical importance or clinical meaningfulness. Significance and importance cannot be further from each other. As Figure 4 shows, statistically significance has to do only with whether the effect observed is likely or not to be due to chance.

Situations A-B-C in Figure 4 are ones with a meaningful or important difference of 10 units between group x and group y, but the only one statistically significant (at the 0.05 level) is A. Situation B has just slightly smaller sample size while Situation C has slightly larger variability, leading to a probability that the observed difference is due to chance is >0.05, not statistically significant. Situations D-E-F are ones with a meaningless difference of 1 unit between group x and group y, but Situation E has a very large sample size and Situation F has very small variability, leading to a probability that the observed difference is due to chance to be <0.05, statistically significant. Note especially that lack of significance does not mean 'no effect' or 'no difference;' it simply means that the observed effect or difference is likely to be due to chance.

The concept of statistical significance and the calculation of a 'P-value' were introduced by Sir Ronald A. Fisher in the early years of the 20th century. Fisher viewed the P-value as an informal index of the discrepancy of the data with the assumed model (null hypothesis), and suggested the following interpretation for the P-value: '*If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts*' (Fisher, 1925). As shown in Figure 5, Fisher interpreted the range of P-values between 0.02 and 0.10 as inconclusive, requiring additional data from observation or experimentation. He was asked to be more specific within this 'grey zone,' so in 1926 he stated '*The value for which P = 0.05, or 1 in 20, is 1.96 or nearly 2; it is **convenient** to take this point as a limit in judging whether a deviation ought to be considered significant or not*' (Fisher, 1926). There were two important reasons it was convenient. First, in his work, he normally encountered the bell-shaped probability distribution of Gauss, for which 0.05 was close to covering points within ± 2 standard deviations from the mean. Second, in the early years of the 20th century, the tabulations of probabilities of the most commonly used distributions (the standard Gauss Z distribution, Student's t distribution, Fisher's F distribution, and Pearson's chi-squared distribution) had to be done by laborious hand calculations, so they only tabulated the extreme portions of the distributions, for probabilities 0.01, 0.02, 0.05, and 0.10. Thus, 0.05 was a convenient choice in the 'grey zone' (see Figure 5) as the cutpoint for whether the data are likely to be due to chance (not significant) or not likely to be due to chance (significant).

If the P-value for a particular test is large, then there is only one of three possibilities: there is too much uncertainty relative to the effect (large variability, small sample size, poor power); the study has some bias or confounding that is producing the result; or the data are consistent with the null hypothesis ['*what was observed possibly could be due to chance*']. On the other hand, if the P-value is small, then there is only one of three possibilities: a rare event has occurred (by chance); the study has some bias or confounding that is producing the result; or there is strong evidence that the null hypothesis is probably not true and should reject it ['*what was observed is possibly not due to chance*'].

The implications and recommendations should be logical, based on the findings, and explained thoroughly, with appropriate caveats. There are important steps to follow in understanding the observed effects/results. First, do they make sense; i.e. do they agree with expected results, in direction and magnitude? If not, what may be wrong, the expectations, the study, the data, or the analyses? If they make sense, then ask if they are meaningful, important or


A
Clinically important and statistically significant
B
Clinically important but not statistically significant
C
Clinically important but not statistically significant
D
Clinically not important and statistically not significant
E
Clinically not important but statistically significant
F
Clinically not important but statistically significant

```
         Obs    Mean    Std. Dev.
   x      50      90          25
   y      50     100          25
diff             -10               Pr(|T| > |t|) = 0.0483
```

```
         Obs    Mean    Std. Dev.
   x      45      90          25
   y      45     100          25
diff             -10               Pr(|T| > |t|) = 0.0611
```

```
         Obs    Mean    Std. Dev.
   x      50      90          30
   y      50     100          30
diff             -10               Pr(|T| > |t|) = 0.0988
```

```
         Obs    Mean    Std. Dev.
   x      50      90          25
   y      50      91          25
diff              -1               Pr(|T| > |t|) = 0.8419
```

```
         Obs    Mean    Std. Dev.
   x    5000      90          25
   y    5000      91          25
diff              -1               Pr(|T| > |t|) = 0.0455
```

```
         Obs    Mean    Std. Dev.
   x      50      90           2
   y      50      91           2
diff              -1               Pr(|T| > |t|) = 0.0141
```

Figure 4. Hypothetical data example to illustrate the difference between interpretation of a difference as statistically significant as opposed to clinically important or meaningful.

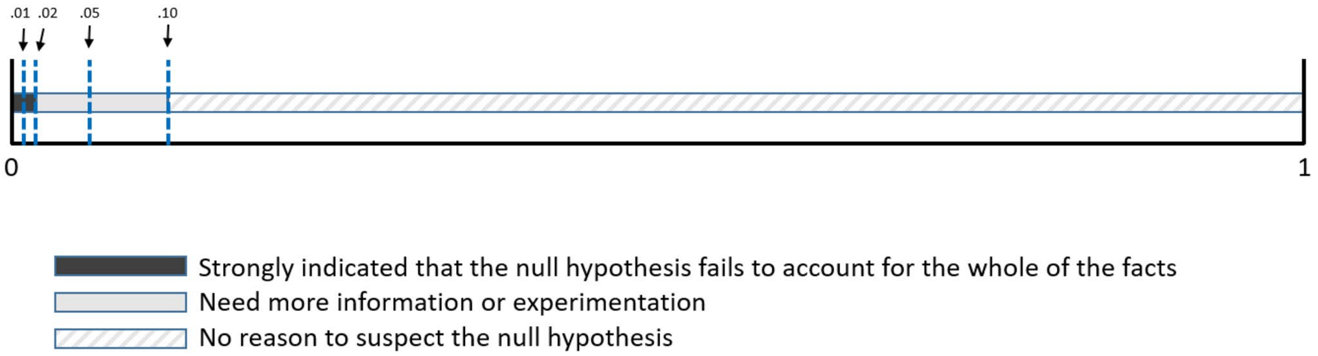

.01
.02
.05
.10
0
1
Strongly indicated that the null hypothesis fails to account for the whole of the facts
Need more information or experimentation
No reason to suspect the null hypothesis

Figure 5. The P-value, an index measuring whether the data are compatible with the null hypothesis, as interpreted by Sir. Ronald A. Fisher.

relevant. If unimportant, no need to proceed. If important, one must then verify if the observed results could be due to some biases (e.g. selection, measurement, recall, analytic) or could it be explained by confounding. If such issues can be ruled out, only then one assesses if one can ignore the role of chance; i.e. test if 'statistically significant.'

The use of P-values should be restricted to hypothesis-driven research questions that have been pre-specified prior to 'playing around with the data.' If the research question is exploratory, the statistically proper method for statistical inference is to provide the value of the estimated parameter or effect of interest in the sample, along with bounds of uncertainty around the estimate; a.k.a. a confidence interval.

Figure 6 illustrates the interpretation of confidence interval for 2-sided tests of superiority, tests for equivalence and 1-sided tests of non-inferiority when a positive difference is considered better. Confidence intervals A-B-C-D illustrate the interpretation of 2-sided superiority hypotheses when a positive difference is considered better. A-B-C are all consistent with a 2-sided statistical test being significant since they do not contain 0, but only A can be interpreted has showing 'clinical superiority' since it does not contain the 1-sided clinical superiority threshold, and B shows 'clinical difference' since it does not contain the 2-sided clinical difference threshold. Confidence intervals E-F illustrate the interpretation of testing for equivalence. Despite having a point estimate above 0, E is consistent with the null hypothesis of 'not equivalent' since part of the confidence interval is outside of the equivalence tolerance threshold. Hard to comprehend, but F would be consistent with significantly showing 'equivalence' despite having a point estimate that is negative, simply because its limits are within the thresholds. Finally, confidence intervals G-H illustrate the interpretation of the 1-sided test of non-inferiority, where a positive difference is considered better. G is consistent with the hypothesis of not-inferior since despite the point estimate being negative, the lower limit of the G confidence interval is still above the non-inferiority limit -Δ. H is consistent with the hypothesis of inferiority since the lower limit of the H confidence interval is less than the non-inferiority limit -Δ.

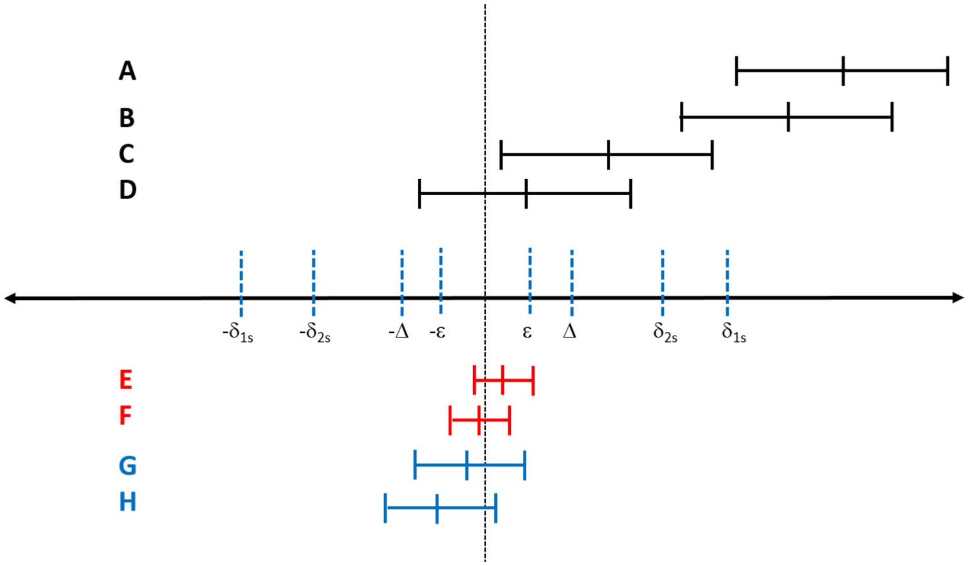


Figure 6. Interpreting confidence intervals beyond simply assessing whether it is statistically significant.

Concluding remarks

Statistical sciences help us in our desire to understand the various complex processes – biological, behavioural, psychological, sociological, cognitive - in nature. What we observe is subject to multiple sources of variability ('noise'), and statistical methods provide us the tools in our research to control or reduce the noise so that we can detect the 'signal amidst the noise'. We must accept we will never know the truth, but we can minimize the uncertainty by design, and quantify the remaining uncertainty in our analyses. Every analytic method involves a set of assumptions and simplifications that enable practicality for addressing the research question.

Variability in the processes under study is necessary for requiring the use of statistics. If there is no variability, the methods of statistics are not needed. If the signal is so much stronger than the noise, statistical arguments may not be needed except for the most sceptical to be convinced. The proper use of statistical methodology enables addressing the research question in a sound, defensible manner, one that accounts for the sources of variability in the observed data and provides an answer that accounts for its uncertainty.

In injury epidemiology we deal with understanding 'relationships' – we want them to be causal, but making the argument can be by design (e.g. randomization controlled trials) or by analysis (regression models). We start by getting a decent sample [representative, unbiased, of sufficient size, with coverage], get decent data [no measurement error, unbiased, no informative missing], do decent analyses [appropriate methods, assumptions satisfied], do proper interpretation [meaningfulness, direction and magnitude of effects], to then finally rule out the role of chance. Statistics is seen as helping in the final step, but the discipline of statistics is the science of dealing with uncertainty, and it comes in all stages, from design, to conduct, to analysis and interpretation.

Acknowledgement

Special thanks to Ananta Bangdiwala for reviewing a draft of this manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Afukaar, F. K. (2003). Speed control in developing countries: Issues, challenges and opportunities in reducing road traffic injuries. *Injury Control and Safety Promotion*, *10*(1–2), 77–81. https://doi.org/10.1076/icsp.10.1.77.14113

Bangdiwala, S. I. (1989). The teaching of the concepts of statistical tests of hypotheses to non-statisticians. *Journal of Applied Statistics*, *16*(3), 355–361. https://doi.org/10.1080/02664768900000043

Bangdiwala, S. I., & Banerjee Taylor, B. (2011). Statistical considerations. In G. Li & S. P. Baker (Eds.), *Injury research*, 383–396. Springer.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Box, G. E. P., & Luceño, A. (1997). *Statistical control: By monitoring and feedback adjustment*. John Wiley & Sons.

Fisher, R. A. (1919). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52*(2), 399–433. https://doi.org/10.1017/S0080456800012163

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.

Fisher, R. A. (1926). The arrangement of field experiments. Journal of the Ministry of Agriculture of Great Britain. *33*, 503–513. https://doi.org/10.23637/rothamsted.8v61q

Haddon, W. (1970). On the escape of tigers: An ecologic note. *American Journal of Public Health and the Nation's Health*, *60*(12), 2229–2234. https://doi.org/10.2105/ajph.60.12.2229-b

La, Q. N., Lee, A. H., Meuleners, L. B., & Duong, D. V. (2013). Prevalence and factors associated with road traffic crash among bus drivers in Hanoi, Vietnam. *International Journal of Injury Control and Safety Promotion*, *20*(4), 368–373. https://doi.org/10.1080/17457300.2012.748810

Paul, S., Mehra, S., Prajapati, P., Malhotra, V., Verma, K. C., & Sidhu, T. K. (2019). Unintentional injury and role of different predictors among 1-5 years children: A community based cross sectional study in a rural population of a developing country. *International Journal of Injury Control and Safety Promotion*, *26*(4), 336–342. https://doi.org/10.1080/17457300.2019.1595666

Runyan, C. W. (1998). Using the Haddon matrix: introducing the third dimension. *Injury Prevention: Journal of the International Society for Child and Adolescent Injury Prevention*, *4*(4), 302–307. https://doi.org/10.1136/ip.4.4.302

Streiner, D. L. (2018). Statistics commentary series: Commentary no. 27: P-hacking. *Journal of Clinical Psychopharmacology*, *38*(4), 286–288. https://doi.org/10.1097/JCP.0000000000000901

Thabane, L., Thomas, T., Ye, C., & Paul, J. (2009). Posing the research question: Not so simple. *Canadian Journal of Anaesthesia = Journal Canadien D'anesthesie*, *56*(1), 71–79. https://doi.org/10.1007/s12630-008-9007-4

Wijlhuizen, G. J., Du Bois, P., van Dommelen, P., & Hopman-Rock, M. (2007). Effect evaluation of a multifactor community intervention to reduce falls among older persons. *International Journal of Injury Control and Safety Promotion*, *14*(1), 25–33. https://doi.org/10.1080/17457300600935189