Statistical Errors in Aviation Psychology: Commonsense Statistics in Aviation Safety Research

Christopher D. Wickens

STATISTICAL ERRORS IN AVIATION PSYCHOLOGY: COMMONSENSE STATISTICS IN AVIATION SAFETY RESEARCH

Christopher D. Wickens
AlionScience & Colorado State University Fort Collins Colorado

I discuss problems with the use of null hypothesis significance testing, as it is particularly applied to safety research such as that in aviation psychology. Such problems are manifest in the inherent bias of traditional statistics to avoid type 1 statistical errors, and hence to discourage findings of safety improving effects as significant, when low powered experimental designs are required by necessary constraints. In contrast, I offer several approaches or remedies. Researchers should think about the decisions made by consumers of their research, based on the costs and values of those decisions; they should form alternative hypotheses, use smart planned comparisons where possible, and present data on the size of effects that do not meet conventional .05 levels of significance. Meta analyses are also encouraged.

In a hypothetical research project, investigators have examined an instructional program to train pilots to better understand the flight management system modes, and respond appropriately to unexpected surprises. A group of 20 line pilots from commuter airlines are selected to go through either conventional training or the augmented "understanding" instructional program. A transfer of training experiment is then done and after a 1 week delay pilots are confronted with an unexpected configuration of the FMS in a high fidelity simulator; the time until the initial correct diagnosis and response is recorded for each of the 10 pilots in the two groups. The authors report a mean RT of 9.5 seconds for the "understanding" group and of 14 seconds for the control group, a non-significant ($p>.05$) effect. A follow up study, with a slightly revised "understanding" curriculum is carried out later with 16 pilots (8/group), and it also provides a non-significant ($p>.05$) benefit, here of 3 seconds. It is concluded, based on the two studies showing no significant benefit, that the new curriculum is non effective. The developer of the curriculum points out to the investigators that if the samples of the two studies are pooled, with a resulting N=18/group, the mean benefit, now of approximately 4.5 seconds would have proven significant ($p<.05$). Furthermore, examining more closely the statistics that underlay the two experiments, the developer noted that the two p-values were, respectively, .07 and 0.11.

This hypothetical (but plausible) research scenario illustrates the potentially serious flaws in the manner that classical null-hypothesis significance testing (NHST) is applied in our safety-critical profession of aviation psychology. The likely conclusion by the research sponsor, and airline training groups that the technique was "ineffective", quite possibly results in a decision not to adopt it, and perhaps the resulting failure to take steps that could prevent serious FMS-related mishaps down the line.

In the following, I will outline some of the main concerns underlying the above sequence of events, and suggest remedies that might ameliorate some of these concerns. I will be drawing on some of my previous thinking about "common sense statistics" (Wickens, 1998), which itself was inspired by an earlier article on aviation safety by Don Harris (1991), as well as the more recent seminal article by Cumming (2014) on "the New Statistics". Cumming writes much more than room allows here to summarize that is relevant for our profession.

Five flaws in conventional statistical thinking.

Flaw #1. The p-value is a dichotomous , black-white cut off of significance, at p=0.05.

Fisher (1925), who developed the concept of the p value, never intended it to be used as a dichotomous criterion. The concept represents a continuum of the degree of evidence, in support of a hypothesis given the data. No different from degree of altitude on approach to a runway, or degree of temperature, there may be certain relatively important values along these continua, akin to a 25,000 foot "sterile cockpit" altitude on approach; or a 32 degree freezing point, but this certainly does not mean other changes in the variable are unimportant or to be disregarded; despite assertions (by many reviewers) that a $p>.05$ is "just non significant, and should not be talked about as if it were" (paraphrasing from several reviewers of my submitted

manuscripts). Indeed such dichotomous thinking can often lead to what I have referred to as "statistical illogic" of dichotomous thinking as in the following case: An ANOVA reveals a "significant" workload effect on performance, across three levels, low, medium and high; but then separate post-hoc comparisons reveal that the low-medium contrast is NS or "statistically equal", as is the medium-high contrast, while the low-high contrast is significantly ($p<.05$) different. So if L=M and M=H. Then in the logic of comparisons L must equal H. But the third comparison shows that this is not true: a contradiction.

Flaw #2. p = 0.05 represents a "decision rule".

It does not. The decision rule is defined by alpha, set by the experimenter. It can be at any p value chosen (convention often does select .05), and so the black-white thinking associated with the p value is more correctly associated with alpha. Whereas Fisher conceived of p as a continuous "evidence variable", it was Neyman and Pearson (1933) who developed the logic of the decision rule often associated with NHST; (Hubbard & Bayarri, 2003), to firmly either "accept" the alternative hypothesis, or "accept the null" (reject the alternative). If we can think of the traditional statistical analysis package as a form of automation, the distinction between the p value and alpha very closely parallels the distinction between stage 2 automation (information integration and inference): the p value (or its closely associated confidence interval) and stage 3 automation (decision making): the alpha level, with its associated decision to "accept" or "reject" an effect as meaningful. As we have pointed out elsewhere (Parasuraman Sheridan & Wickens, 2000: Onnasch, Wickens et al, 2013), errors of automation at stage 3 have more problematic consequences than those at stage 2. And as we see below, such automation can easily make errors.

Flaw #3 NHST is biased toward the status quo.

Table 1 presents the standard matrix underlying NHST. Across the top there is some "ground truth" effect that exists in the world (population) that we are trying to confirm. Here, this might be the truth that our experimental manipulations will make people better pilots and hence improve safety. We run the experiment, compute the statistics and derive a conclusion, based on whether our p value exceeds or is less than alpha (which is conventionally set to be.05). By accepting a criterion of .05, our decision rule is designed to "assure" that given our data, there **is** only one chance in 20 that we will conclude there is an effect, if the experiment is repeated multiple times (Cumming, 2104), when there is actually none to be found in the population; an errant conclusion resulting from the contributions of randomness to the data. This is the **type 1 error.**

Table 1.
The conventional table of statistical decisions with NHST

	State of the world: "the truth"	
Experimental results	Improve safety	No improvement
Disconfirm Ho (e.g., $p<.05$): An effect		Type 1 error (.05). Strongly discouraged
Confirm Ho ($p>.05$). "NS"	Type 2 error	

In contrast, conventional NHST is silent on the probability of concluding that there is **no** effect, when there actually **is** one, as shown in the bottom row, concluding a "non significant effect" such as that described in our story above: **a type 2 error**. This is a real number, which can be estimated from statistical power calculations (Cohen, 1988). But application of conventional statistics places far less emphasis on this, than it does on keeping the type 1 error below .05; and as a result, most experiments show a bias toward a considerably higher probability of the type 2 than the type 1 error. In essence, there is a direct analogy to our criminal justice system that cares much more to avoid an innocent person being found guilty than the converse; and hence requiring unanimous jury decisions to convict, and only one dissenting member to exonerate. The standard of evidence for guilt is set very high, just as in traditional NHST, the standard of avoiding a type 1 error is set high. This state of asymmetric concern for avoiding type 1 more than type 2 errors is an outgrowth of concern in basic sciences, that somehow type 1 errors are "worse" than type 2 errors, and a case can be made that the scientific community does not want a plethora of effects claimed, that turn out to be "untrue". But should this bias apply equally to safety research? As I argue below, it should only apply less severely, leading to the 4th "flaw".

Flaw #4 NHST does not address values in decision making.

Table 2 presents a classic decision table in expected value theory, populated by the specific characteristics of our automation training example above. It is similar in some respects to table 1, but also quite distinct. The

two possible states of the world regarding a "ground truth" are again shown across the two columns, and the two rows again represent decisions. However these are not decisions to reject or accept the null hypothesis by the researcher, but instead represent decisions, made the consumer of our safety research, to either adopt the concept suggested y the experimental results (e.g., implement the training plan) or ignore it. This is a very different form of decision than that made by the researcher to "decide" to say in print, whether the effect is "significant" or not.

Table 2.
The classic expected value decision matrix.

	State of the world	
Decision	Improves safety (P)	Does not improve safety (1-P)
It works: adopt the procedure	Mishaps saved cost of adoption	Cost of adoption
It does not work. Discard the procedure	Unnecessary mishaps created	No cost

Most importantly, we can now depict specific costs and benefits of different outcomes, particularly for the two types of decision "errors" that corresponded to the type 1 and type 2 errors in Table 1. Rather than, as in table 1, simply saying "type 1 errors are worse than type 2 errors", one can begin to put some approximate numbers on these to make a more objective judgment, as indicated by the cells of table 2.

Flaw #5 NHST does not address probabilities in decision making.

A second feature of the decision matrix in table 2, is the explicit presentation of some a-priori estimation of the prior probability (p) that the state of the world is true, as shown across the top row. These are quite different from the probability value depicted in table 1, which only represents the probabilities that the alternative hypothesis (H1) is falsely accepted, **given the data** observed in the experiment. The two probabilities have totally different meaning. Of course there is often no basis for making such prior probability estimation in Table 2, but by setting it a 50/50 for example, you might be essentially saying that "independent of, or prior to observing my experimental results, I think it is just as likely that my intervention will improve safety as not". This starts both hypotheses with even odds of confirmation

There is certainly plenty of debate in the statistical and experimental world about the role of prior probabilities in hypotheses testing, an issue defined as "Bayesian statistics" (e.g., Berger, 2000; Cumming, 2014); but a little common sense can be applied. If several prior experiments have suggested evidence that the instructional intervention "works" (i.e., in the above example, the results of the first experiment should have made it clear that it was more likely to work than not), then one can approach an additional experiment with some bias to assume that it works, or at least an equal footing. Of course, this guidance is only feasible to the extent that we can decide how big an effect is defined as "it works", and this requires some estimation of an effect size defined as "working". This procedure is the hallmark of statistical power analysis, and once an effect size is chosen (e.g., 50% of the variance accounted for), then it is possible to specifically compute the probability of a type 2 error in one's experiment, And it may then be possible to set alpha at that same level of probability, hence providing more equal odds for the two types of errors.

In summary, there are two general points to be made here: (1). The consumer of the research, who is the ultimate implementer of safety-critical procedures should be provided by the researcher with more data upon which to base her decision, than simply an "accept/reject" output of a statistical decision rule which implicitly or explicitly sends the message, in one case that "there is nothing there" (2) application of this decision rule, without adequate statistical power, provides an inherent bias against adopting safety improvement procedures or equipment.

What is to be done?

Below, I have outlined two general categories of remedies for this state of affairs; changes to how the researcher should approach experimental design and analysis, and changes to the way data are presented in written reports and articles.

Design and Analysis.

Increasing statistical power. It should be apparent that increasing statistical power, typically by running more subjects, which will reduce the estimated variance in effect size, will increase the confidence that a given effect is "true", and hence decrease the probability of a type 2 error. If this probability can be decreased to 0.05, then there is the desired symmetry between the two types of errors, and not the inherent bias against concluding safety-improving effects. Unfortunately however there are two issues that mitigate against such an increase in N in aviation human factors research (and research in other safety critical professions). First, it may be extremely difficult to obtain the participation of highly skilled professional workers in high fidelity simulation experiments and, with a restricted budget the researcher may well feel lucky to get even the 20 qualified line-pilots of our first example to find the time. Second, for reasons I have articulated at this symposium in 2009 (Wickens 2009), the very responses that may be most safety critical (and safety compromising) are those in which the pilot or controller might be most surprised (not expecting); the so called "black swan" event. This would be true of the unexpected first failure event used to estimate the success of training in our example above. Responses to a single, first failure event, of which there is by definition only one per pilot, do not afford the luxury of averaging across trials to reduce variance (and increase statistical power), and in contrast to so many other variables. Yet such first failure responses are unique in their ability to yield worst case response times and detection rates (Wickens Hooey et al., 2009). It is often these unexpected events that compromise safety.

Formulate an alternative hypothesis. An alternative hypothesis can be explicitly formulated to provide equal footing to the null hypothesis of "no effect". Statistical power calculations require this to be done. However these are typically based on deciding an effect size that is "meaningful" (e.g., 75% of the variance accounted for). But it is often more compelling to express this in meaningful performance units. In our example above, the researcher may state that under cases of possible spatial disorientation, in which timely diagnosis of an automation-induced upset is critical, any time-savings of greater than 3 seconds is *the* desired effect. Hence H1 = 3, while Ho = 0, and a savings of 2 sec would stand as more confirming of H1 than Ho. And it is always possible to "reserve judgment" for such intermediate effects. Similar "point estimates" of an alternative hypothesis might be made for the % improvement in performance that results from a particular display, or training innovation. In our field of aviation, where, because of its physical/spatial constraints, safety margins can so often be defined explicitly in terms of time and distance (separation), such alternative hypothesis point estimations are quite feasible.

Use "smart statistics" and planned contrasts. [and tolerate them if you are a reviewer].The data in figure 1 provide a case study of a typical experimental result. Two displays to aid mid-air collision avoidance are contrasted and both tested under low vs high workload conditions. Using conventional statistics, an ANOVA is performed and reveals a significant "p<.05" effect of workload, but a "NS p>.05" effect of both display type and its interaction with workload. Conventional statistics says "end of story". However what I have labeled "smart statistics" would make two points. First, the condition and effect you really care about is collision avoidance under **high workload** (potentially a worst-case, safety critical scenario). Hence what you should really do is to focus a **planned comparison** on the high workload condition. Indeed Cumming (2014) has effectively argued that going into an experiment, the investigator should **know** what s/he is looking for in terms of specific effects, and the omnibus F test ("is anything happening?") is a rather indirect way, of asking whether "something specific that you care about" is happening.

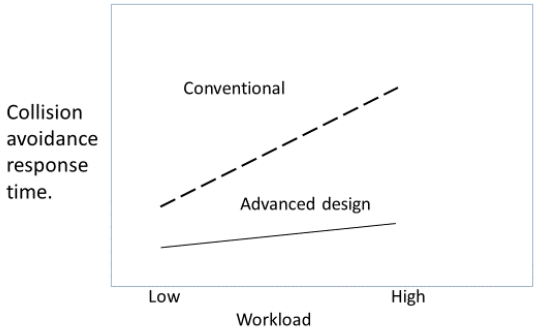

Figure 1. Hypothetical example of Results in a 2X2 experimental design

A second role of "smartness" in statistical testing, and particularly for most safety critical research, is that the comparisons should be 1 tailed, which provides more statistical power, rather than 2-tailed. What this essentially means is that you care if the effect is one that is safety-improving (i.e., typically shortens RT, improves accuracy or lowers workload). But you DON'T care whether there is no effect at all, or the effect works in the other direction. In either of the latter two cases, your proposed safety improvement is **not working**.

Presentation of experimental results.

Present the "raw data". It is worth highlighting again, the importance of presenting more, rather than less "raw data" to the readers of your article. By "raw data", we do not of course mean the individual subject measures, but we do mean graphs, 95% confidence intervals, effect size measures, and all statistical test measures (not just those of the magical "significant $p<.05$" type). The added relevance of this last guidance to meta-analyses will be described below. Here, if we think about statistics and stats packages in terms of the stages and levels of automation framework (Parasuraman et al., 2000), later stage automation is good if it is correct, but more problematic if it is in error of either the type 1 or type 2 variety. As we know, a stats package that simply tells you to accept or reject the null hypothesis is an example of late stage decision aiding automation, and, of course, can have a (typically) 5% chance of being in error. The best mitigation of this, in human-automation interaction research, is to let automation provide and convey to the reader more assistance in the earlier stage of integration and inference, and here, that specifically means providing graphed data and confidence intervals, along with the full array of inferential statistics.

Choose language carefully. Be very careful that the language you use in the text, does **not** convey the impression that effects which might be important for safety improvement but fail to reach the magic .05 levels are to be disregarded. The offenses here, ranked from worst to better for such a phrase to describe, say a .07 effect would be to say: "there was no effect", "not different", "not significantly different". Even if you **do** report the p values of such $p>.05$ effects in the results, time-limited readers of only a Discussion, or Abstract may not have taken the time to find those statistics. More plausible approaches (although here I have had to argue with editors) would be to label such effects as "marginally significant" or "approaching conventional levels of statistical significance" or even a "non-significant trend". It is equally important, when such effects are in evidence to describe in the text (and not just tables and graphs), their actual magnitudes, in terms such as the 4 second savings in response time, or the 30% gain in accuracy.

Accumulating evidence over experiments.

Earlier, we referred to "prior probabilities" for assuming that an effect might actually exist in the world, before we have seen the data from our current experiment. Of course the best source of such prior odds comes from other research on the topic, that may have used the same or similar variables to reveal the effect in question. Literature reviews can qualitatively summarize that research. But the ideal tool for this is the meta-analysis Rosenthal, 1991; Cumming, 2014; Onnasch et al., 2014). Various meta-analytic approaches can actually yield a quantitative estimate of the "collective wisdom" of that prior research, which may enable our researcher to not only express that an effect is likely to be there (or not), but also can give a point estimate of how large it is; that is, an explicit alternative hypothesis. The importance of meta-analyses has two implications: first, in a review of the literature, even an informal meta-analysis can establish the prior likelihood of finding the effect. Second, in helping others do their own meta-analyses, reporting the null effects that you do observe in your own data (or effect sizes of "NS $p>.05$" effects), you can provide data for the analyses of others that are not inherently biased toward more positive results.

Conclusions

In conclusion, we note that many of the concerns that have brought about the emergence of the .05 level of significance to avoid type 1 errors are well formulated, and we do not argue for a total abandonment of the tenets of NHST. However I argue that people should clearly understand the biasing implications of the black-white approach fostered by alpha levels, particularly when "the effect" that is examined (and is often rejected because of NHST) is one that has safety-improving implications. I hope that some of the remedies suggested above, can be adopted by the aviation psychology research community when they consider the decisions that the consumers of their work may make.

Dedicated to Tom Wickens: 1943-2012

References

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd Ed.). New York: Springer-Verlag

Cohen, J. (1988) *Statistical Power analysis for the behavioral sciences*. 2nd Edition. Hillsdale, N.J,: Erlbaum.

Cumming, G. (2014) The New Statistics: Why and How *Psychological Science*., *12,* 7-29.

Fisher, R.A. (1925) *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Harris, D (1991). The importance of type 2 error in aviation safety research.In E. Farmer (Ed.) *Stress and Error in Aviation* (pp 151-157). Brookfield Vt.: Avebury.

Hubbard, R. & Bayarri, M. (2003). Confusion over measures of evidence (p's) versus errors (alpha's) in classical statistical testing. *The American Statistician*. *57*, 171-188.

Neyman, J. & Pearson, E. (1933) On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. A.231 289-337.

Onnasch, L., Wickens, C., Li, H. & Manzey, D. (2014) Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. Human Factors. 56(3), 476–488.

Parasuraman, R., Sheridan, T.B., & Wickens, C.D. (2000). A model of types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 30, 286-297.

Rosenthal, R. (1991). Meta-analytic procedures for social research (Rev ed.). Beverly Hills, CA: Sage.
Wickens, C.D. (1998) Commonsense Statistics. Ergonomics in Design. Oct. pp 18-22.

Wickens. C.D. (2009). The psychology of aviation surprise: an 8 year update regarding the noticing of black swans. In J, Flach & P. Tsang (eds). Proceeedngs 2009 Symposium on Aviation Psychology: Dayton Ohio: Wright State University

Wickens, C.D. Hooey, B. Gore, B.F., Sebok, A. & Koenicke, C. (2009) Identifying Black Swans in NextGen: Predicting Human Performance in Off-Nominal Conditions. Human Factors. 51, 638-651.