AAS AAS 17–614

RELEVANCE OF THE AMERICAN STATISTICAL SOCIETY'S WARNING ON p-VALUES FOR CONJUNCTION ASSESSMENT

J. Russell Carpenter[*] Salvatore Alfano[†] Doyle T. Hall[‡] Matthew D. Hejduk[§] John A. Gaebler[¶] Moriba K. Jah[‖] Syed O. Hasan[**] Rebecca L. Besser[††] Russell R. DeHart[††] Matthew G. Duncan[‡‡] Marissa S. Herron[§§] and William J. Guit[¶¶]

On March 7, 2016, the American Statistical Association issued an editorial paper on the "context, process, and purpose of p-values." According to the paper, "the statement articulates in non-technical terms a few select principles that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community." These principles would appear to have some relevance to the spacecraft conjunction assessment community.

INTRODUCTION

On March 7, 2016, the American Statistical Association (ASA) issued an online editorial paper[1] on the "context, process, and purpose of p-values." According to the paper, "the statement articulates in non-technical terms a few select principles that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community." This "consensus" statement was accompanied by 21 "commentaries" expressing a diversity of opinions among the panel of experts ASA convened. In the present work, we express our view that the ASA p-value warning has relevance to the space object Conjunction Assessment (CA) community.

The ASA Editorial gives an informal definition of a p-value as follows: "Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (for example,

[*] Deputy Project Manager/Technical, Space Science Mission Operations Project, NASA Goddard Space Flight Center, Code 444, Greenbelt, MD 20771.

[†] Senior Research Astrodynamicist, Center for Space Standards and Innovation (CSSI), 7150 Campus Drive, Suite 260, Colorado Springs, CO 80920-6522.

[‡] Senior CARA Analyst, Omitron Inc., 555 E. Pikes Peak Ave, #205, Colorado Springs, CO 80903.

[§] Astrorum Consulting LLC

[¶] Graduate Research Assistant, Aerospace Engineering Sciences, University of Colorado at Boulder, Boulder, CO 80302.

[‖] Associate Professor, Aerospace Engineering and Engineering Mechanics, The University of Texas at Austin, Austin, Texas 78712.

[**] Earth Observation System Lead Collision Avoidance Engineer, KBRwyle, NASA Goddard Space Flight Center, Code 428, Greenbelt, MD 20771.

[††] Systems Engineer, KBRwyle, NASA Goddard Space Flight Center, Code 444, Greenbelt, MD 20771.

[‡‡] SpaceNav LLC, Denver, Colorado 80211.

[§§] Deputy, NASA Robotic Conjunction Assessment, NASA Goddard Space Flight Center, Code 595.1, Greenbelt, MD 20771.

[¶¶] EOS Aqua Mission Director, Mission Validation and Operations Branch, NASA Goddard Space Flight Center, Code 584, Greenbelt, MD 20771.

the sample mean difference between two compared groups) would be equal to or more extreme than its observed value." A slight rephrasing of the definition will place it into the CA context: The collision probability, P_c, is the probability, under a specified set of modeling assumptions, that the estimated distance between two space objects would be equal to or less than the value we infer from our observations. More specifically, p-values are often used in statistical hypothesis testing to assess whether or not a decision to reject a hypothesis is justified. In CA, we are faced with a similar decision concerning whether or not to perform some kind of mitigation, such as a maneuver.

IS P_C A P-VALUE?

Based on the ASA's informal definition quoted above, it would seem that the answer to the question posed by this section is yes. However, the more familiar context for p-value usage is that of hypothesis testing, which begs the question as to whether P_c is used by the CA community, explicitly or implicitly, in a statistical hypothesis test. In such tests, one seeks to use uncertain or noisy data to inform a decision whether to reject a model hypothesized to give rise to the observed data. In one of the simplest hypothesis tests, one computes how unlikely the observations would be if the hypothesized model were true. The degree to which the observations are likely given the model is quantifiable as a probability, which is known as p. Therefore, when p is "small enough," it is common practice to formally reject the null hypothesis that the model is true, conditional on a particular set of observations and the strategy used to collect those observations.

What we have just described bears more than a passing resemblance to the CA process. The usual purpose of CA is to inform a decision, made under uncertainty, about whether to mitigate the risk posed by a conjunction or not. Given assumptions on the physics and sensors, we use observations to infer statistically probable trajectories. We propagate these orbits to an interval associated with the predicted conjunction, along with models of how uncertainty in the observations and the propagation models maps into uncertainty in the orbits, usually in the form of a covariance matrix. We use the propagated orbit and its covariance to compute a probability, P_c, and if P_c is "small enough," we generally take no action to mitigate the risk of a collision. All that we need to cast the CA process into the realm of a hypothesis test is to formally state the null hypothesis that we are rejecting when we find small values of P_c. Our rephrasing of the ASA definition points to how this null hypothesis, $\mathcal{H}_o$, for CA might be formulated:

> **$\mathcal{H}_o$:** The estimated distance between two space objects is less than or equal to the combined hard-body radius of the objects.

Thus, when P_c is appropriately small, and when the predicted miss distance is adequately large, this suggests that the observed tracking data, when fit to our orbit models, are inconsistent with a collision between the objects, and hence we are rejecting the null hypothesis above. When we decide to perform a CA risk mitigation maneuver, we have at least implicitly accepted (or technically, failed to reject) such a null hypothesis.[*]

Despite the apparent correspondence between P_c and p-values, this paper's appendix describes how they are different. The appendix provides two examples of hypothesis tests involving p-values

[*] Such a decision is subject to errors of two types. An error of Type I is rejecting the null hypothesis when it is true, which in the context of this paper's hypothesis is a *missed detection* or *missed alarm*. An error of Type II is is accepting the null hypothesis when it is false, which in the present context leads to a needless mitigation maneuver, or *false alarm*. Although any individual decision will be either right or wrong, our goal ought to be establishing a decision procedure whose error rates "in the long run" will limit decision errors to acceptable rates.

that compare and contrast a classic textbook example of a test on the mean of a normal distribution with a test that approximates a CA decision. The appendix shows that in fact P_c is not quite the same as a p-value, but more closely resembles the complement of a confidence value derived from the estimated miss vector and its associated covariance. The corresponding confidence interval has as its lower endpoint the combined hard-body radius. Since this confidence interval corresponds to a family of significance tests, there is an approximate correspondence between P_c and p-values, but they are conceptually different quantities.

THE SIX ASA PRINCIPLES

Having established that the CA process is at least implicitly using P_c in the manner of a p-value, we now examine the six principles ASA advocates to improve how analysts and decision-makers use p-values. In the following tables, we place direct quotations of the six ASA principles concerning p-values from Reference 1 into a side-by-side comparison context with our rephrasings of the principles to place them into the CA context.

Table 1: ASA's Principle 1

"P-values can indicate how incompatible the data are with a specified statistical model."	
ASA's Explanation	**Annotated for CA Context**
A p-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The most common context is a model, constructed under a set of assumptions, together with a so-called "null hypothesis." Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. The smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the p-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis *or the underlying assumptions.* [emphasis added]	A p-value provides one approach to summarizing the incompatibility between a particular set of [tracking] data and a proposed model for the data [which is that the space objects being tracked will collide]. The most common context is a model, constructed under a set of assumptions [e.g. "covariance realism," etc.], together with a so-called null hypothesis. [. . .] The smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis [that the space objects will collide], if the underlying assumptions used to calculate the p-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis [that the space objects will collide] or the underlying assumptions [e.g. the combined covariance is realistic, etc.].

Table 2: ASA's Principle 2

"P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone."	
ASA's Explanation	**Annotated for CA Context**
Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.	[Owner/Operators] often wish to turn a p-value into a statement about the truth of a null hypothesis [that the space objects will collide], or about the probability that random chance produced the observed data. The p-value is neither. It is a statement about [whether tracking data are consistent with the hypothesis that the space objects being tracked will collide], and is not a statement about [whether or not the space objects will actually collide].

Table 3: ASA's Principle 3

"Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold."	
ASA's Explanation	**Annotated for CA Context**
Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision-making. A conclusion does not immediately become "true" on one side of the divide and "false" on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, "yes-no" decisions, but this does not mean that p-values alone can ensure that a decision is correct or incorrect. The widespread use of "statistical significance" (generally interpreted as "$p \leq 0.05$") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.	Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "[$P_c < 1 \times 10^{-4}$]") for justifying [that the conjunction is safe] can lead to erroneous beliefs and poor decision-making. A conclusion does not immediately become "true" on one side of the divide and "false" on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, "yes-no" decisions, but this does not mean that p-values alone can ensure that a decision is correct or incorrect. The [possibly] widespread use of [P_c thresholds] as a license for making a claim [about the risk of a conjunction] leads to considerable distortion of the [conjunction assessment] process.

Table 4: ASA’s Principle 4

“Proper inference requires full reporting and transparency.”	
ASA’s Explanation	**Annotated for CA Context**
P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain p-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference and “p-hacking,” leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided. One need not formally carry out multiple statistical tests for this problem to arise: Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted and all p-values computed. Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting.	[P_c] values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain [P_c] values (typically those passing a significance threshold) renders the reported [P_c] values essentially uninterpretable. [...] One need not formally carry out multiple statistical tests for this problem to arise: Whenever a [CA analyst] chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the [Owner/Operator] is not informed of the choice and its basis. [CA analysts] should disclose [...] all data collection decisions, all statistical analyses conducted and all [P_c] values computed. Valid [risk mitigation] conclusions based on [P_c] values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including [P_c] values) were selected for reporting.

Table 5: ASA's Principle 5

"A p-value, or statistical significance, does not measure the size of an effect or the importance of a result."	
ASA's Explanation	**Annotated for CA Context**
Statistical significance is not equivalent to scientific, human, or economic significance. Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different p-values if the precision of the estimates differs.	Statistical significance is not equivalent to scientific, human, or economic significance [and the significance of a risk derives from both its likelihood and its consequences]. Smaller [P_c] values do not necessarily imply the presence of [less risky conjunctions], and larger [P_c] values do not imply [an elevated risk]. Any [conjunction], no matter how [risky], can produce a small [P_c] value if the [covariance is small relative to the miss distance], and large miss distances may produce [large P_c values] if the [covariance is commensurately large]. Similarly, identical [miss distances] will have different [P_c] values if the precision of the estimates differs.

Table 6: ASA's Principle 6

"By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis."	
ASA's Explanation	**Annotated for CA Context**
Researchers should recognize that a p-value without context or other evidence provides limited information. For example, a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large p-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a p-value when other approaches are appropriate and feasible.	[The CA community] should recognize that a [P_c] value without context or other evidence provides limited information. For example, a [P_c] value near [1×10^{-4}] taken by itself offers only weak evidence against the null hypothesis [that the space objects will collide]. Likewise, a relatively large [P_c] value does not imply evidence in favor of the null hypothesis [the space objects will collide]; many other hypotheses may be equally or more consistent with the observed data [e.g. the tracking data are flawed, etc.]. For these reasons, [conjunction assessment] should not end with the calculation of a [P_c] value when other approaches are appropriate and feasible.

GREENLAND ET AL.'S TWENTY-FIVE MISCONCEPTIONS

It is also helpful to consider a more precise definition of p-value than ASA's informal definition. One of the twenty-one commentaries accompanying the publication of Reference 1 is Greenland, et al.,[2] which gives what we find an especially cogent and precise definition:

> Specifically, the distance between the data and the model prediction is measured using a test statistic (such as a t-statistic or a χ^2-statistic) [or a non-central χ^2 statistic based on Mahalanobis distance for some simplifications of the CA problem]. The p-value is then the probability that the chosen test statistic would have been at least as large as its observed value if every model assumption were correct, including the test hypothesis. This definition embodies a crucial point lost in traditional definitions: In logical terms, the p-value tests all the assumptions about how the data were generated (the entire model), not just the targeted hypothesis it is supposed to test (such as a null hypothesis). Furthermore, these assumptions include far more than what are traditionally presented as modeling or probability assumptions – they include assumptions about the conduct of the analysis, for example that intermediate analysis results were not used to determine which analyses would be presented.
>
> Now it is true that the smaller the p-value, the more unusual the data would be if every single assumption were correct; but a very small p-value does not tell us which assumption is incorrect. For example, the p-value may be very small because the targeted hypothesis is false; but it may instead (or in addition) be very small because the study protocols were violated, or because it was selected for presentation based on its small size. Conversely, a large p-value indicates only that the data are not unusual under the model, but does not imply that the model or any aspect of it (such as the targeted hypothesis) is correct; it may instead (or in addition) be large because (again) the study protocols were violated, or because it was selected for presentation based on its large size.
>
> The general definition of a p-value may help one to understand why statistical tests tell us much less than what many think they do: Not only does a p-value not tell us whether the hypothesis targeted for testing is true or not; it says nothing specifically related to that hypothesis unless we can be completely assured that every other assumption used for its computation is correct – an assurance that is lacking in far too many studies.

Upon reflection, most CA analysts would agree that P_c, as computed in practice, is conditioned on imperfect observations and assumed models of the orbit; for example, it does not tell us anything about the quality of the space weather data used in our models and hence may fall quite short of fully describing the total probability of collision.

Another way to think of this is the following. The only way uncertainty enters into the usual CA process is via errors in the tracking data, and the *a priori* uncertainty assumed for the two space objects' orbits, including parameters related to the orbit fits such ballistic coefficients, atmospheric density, etc. If one were to conduct a Monte Carlo simulation of a conjunction, each trial would make random draws on the tracking errors and the initial orbit conditions. One could then compute the relative frequency across the ensemble of trials for which a collision occurred as a result of the expected variation across the random numbers drawn. The value of P_c that the CA process computes is simply an analytic computation that the Monte Carlo relative frequency will approach "in the long run" as the number of trials increases. But in the real world, there are many other sources of variation, so that if we could repeat the actual conjunction over and over again, then "in the long run" we would see a greater variability in the results.

Greenland, et al. go on to list 25 "misconceptions" regarding p-values, confidence intervals, and statistical power that arise from an imprecise understanding of this definition. While we find many of these to be overly dogmatic and/or repetitive, we list those few of them that appear most relevant to CA here, modified as above for the CA context, and encourage the reader to consult the reference for explanations.

Misconception 1 The p-value [P_c] is the probability that the test hypothesis [the space objects will collide] is true; for example, if a test of the null hypothesis [that the objects will collide] gave $[P_c] = 0.01$, the null hypothesis [a collision] has only a 1% chance of being true; if instead it gave $[P_c] = 0.40$, the null hypothesis [a collision] has a 40% chance of being true.

Misconception 2 The p-value [P_c] for the null hypothesis [that the objects will collide] is the probability that chance alone produced the observed association; for example, if the p-value [P_c] for the null hypothesis is 0.08, there is an 8% probability that chance alone produced the association.

Misconception 3 A significant test result ($p \leq 0.05$) [$P_c \leq 1 \times 10^{-4}$] means that the test hypothesis [a collision] is false or should be rejected.

Misconception 4 A nonsignificant test result ($p > 0.05$) [$P_c \geq 1 \times 10^{-4}$] means that the test hypothesis [a collision] is true or should be accepted.

Misconception 9 The p-value is the chance of our data occurring if the test hypothesis [a collision] is true; for example, $p = 0.05$ [$P_c = 1 \times 10^{-4}$] means that the observed association [of tracking data with a predicted collision] would occur only 5% [0.01%] of the time under the test hypothesis [a collision will occur].

Misconception 10 If you reject the test hypothesis because ($p \leq 0.05$) [$P_c \leq 1 \times 10^{-4}$] , the chance you are in error (the chance your "significant finding" is a false positive) [the chance of a missed detection of a collision, or Type I error] is 5% [0.01%].

Misconception 19 The specific 95% confidence interval presented by a study has a 95% chance of containing the true effect size.

Misconception 20 An effect size outside the 95% confidence interval has been refuted (or excluded) by the data.

Misconception 21 If two confidence intervals overlap, the difference between two estimates or studies is not significant.

Misconception 22 An observed 95% confidence interval predicts that 95% of the estimates from future studies will fall inside the observed interval.

SUGGESTIONS

To address some of the concerns that motivated the publication of Reference 1, the ASA offered some recommendations:

> In view of the prevalent misuses of and misconceptions concerning p-values, some statisticians prefer to supplement or even replace p-values with other approaches. These include methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates. All these measures and approaches rely on further assumptions, but they may more directly address the size of an effect (and its associated uncertainty) or whether the hypothesis is correct.

A number of the commentaries accompanying the publication of Reference 1 offer more specific guidance, and we find Greenland, et al.[2] to be especially cogent in this regard. For example, Reference 2 points out that "... many authors agree that confidence intervals are superior to p-values because they allow one to shift focus away from the null hypothesis, toward the full range of effect sizes compatible with the data – a shift recommended by many authors and a growing number of journals." However, they also issue a "... caution that confidence intervals provide only a best-case measure of the uncertainty or ambiguity left by the data, insofar as they depend on an uncertain statistical model."

While a great deal of research in the CA field has focused on model improvement, such as improved density estimation, enhanced covariance realism, relaxation of simplifying assumptions such as linear relative motion and Gaussian distributions, etc.[need some references here], much of this work is only slowly migrating to operational usage. Some work has also been done to utilize a likelihood ratio in CA analysis using Bayesian methods,[3] but the availability in practice of reliable information concerning the underlying *a priori* statistics on the miss distance has proved challenging except in somewhat limited contexts.

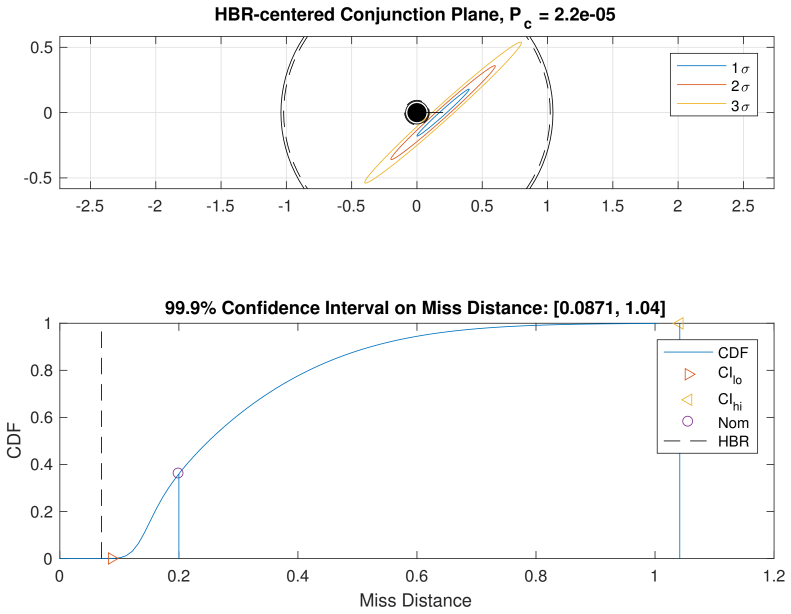

Figure 1. Confidence Intervals on Miss Distance for a Conjunction.

By contrast, it would be relatively easy to modify current CA practice to compute confidence intervals (CI) by merely varying the combined hard-body radius from a lower limit that captures $\alpha/2$ of the probability mass to an upper limit of $1 - \alpha/2$ of the mass. Figure 1 shows an example of such a calculation, for $\alpha = 1 \times 10^{-3}$. The upper subplot shows the conjunction plane, which is normal to the relative velocity at closest approach a solid black disk black at the center encloses the combined hard-body radius (HBR) of the primary and secondary objects. The colored elliptical contours are centered at the nominal miss distance, and indicate the probability masses in units of standard deviation, σ, specified by the legend. To compute CI, we increased the HBR to the thresholds indicated by the dashed/solid line boundaries. The lower subplot illustrates the cumulative distribution function (CDF) for the miss distance, with the miss distances associated with the HBR, lower CI (CI_{lo}, upper CI (CI_{hi}), and the nominal (Nom) all noted. By giving decision-makers both a P_c value and a confidence interval, CA analysts provide them with a measure of the precision of the estimates that a single number like P_c can never convey. However, confidence intervals are no panacea for issues with covariance realism or prediction modeling errors.

Any discussion of p-values would not be complete without mention of the concept of statistical power. The power of a hypothesis test is the probability of correctly rejecting the null hypothesis. In the CA context, it is the probability of dismissing the conjunction when the miss distance truly is greater than the combined HBR. The complement of the power is the probability of a Type II error (false alarm), that is, deciding to perform risk mitigation when it was not required. In the CA context, this may occur when P_c exceeds an action threshold because the relative position covariance is large in comparison to the relative position vector, as often occurs early in the CA process. As has been discussed in Reference 4, when the true outcome of a CA will be a safe miss, it will often be the case that as the time until closest approach decreases, P_c will reach a maximum and then decline. In this case, the same P_c values may occur on either side of the maximum, but a decision procedure based on the earlier value would have less statistical power than a decision procedure that uses the later value. Thus making a decision solely based on P_c, without regard to the size of the covariance and the time remaining until close approach, is more likely to result in a Type II error (false alarm). A difficulty with statistical power is that it can be difficult to quantitatively estimate, since this requires making assumptions about the likelihood of the alternative hypothesis. Nonetheless, in the CA context, analysts can increase the power of the CA hypothesis test by waiting as long as possible before the time of closest approach to make a decision, since both the increased data availability and decreased propagation time would be expected to reduce the relative covariance*. It is however important that analysts apply some degree of rigor to this process; otherwise the temptation to "p-hack" may intrude on the analysis. In particular, decisions concerning how long to wait, how small the covariance should be, etc. are tantamount to decisions concerning the design of a statistical experiment and must be decided in advance of conducting the experiment, i.e. in advance of analyzing a specific CA, in order to avoid biasing the results. Here again, use of confidence intervals on the miss distance may provide useful insight, since a confidence interval that is large in comparison to the estimated miss distance is a good indicator of low statistical power.

Another area in which the CA process might benefit from insight derived from a hypothesis testing perspective would be the application of additional rigor to the determination of thresholds for decisions based on P_c. In current practice, P_c thresholds sometimes appear to be based on the assumption that they correspond to limits on Type I error (missed detection) rates. However, to

*The degree to which it is possible to wait is often highly constrained by the capabilities of the spacecraft, ground systems, and Flight Operations Team to safely plan and execute a CA risk mitigation maneuver.

rigorously choose a decision threshold so that it guarantees a specified Type I error rate, an analyst needs to know the properties of the underlying probability distribution from which she draws her samples, as in the examples in the appendix. This is the same issue that has limited the acceptance of CA based on likelihood ratio methods. More work to characterize such prior densities, perhaps based on debris flux analysis, as was done in Reference 5, would seem to be in order.

CONCLUSION

This paper has argued that the CA community uses P_c in a manner that approximates a classical statistical hypothesis test. To wit, the community effectively uses P_c as it were a p-value, and compares it to a significance level which is sometimes thought to limit the Type I error (missed detection) rate. As such, the CA process is susceptible to the same kinds of critiques that the overall scientific community has been incurring since the publication of Reference 6 and similar articles over the past decade. To assist the CA community in recognizing the potential applicability of the pitfalls that have affected the wider community, this paper has cast the ASA's best practices for use of p-values and related hypothesis testing constructs into a form that should be more familiar to CA practitioners. This paper has further offered specific suggestions for improving the existing CA process, derived from recognition of correspondences with statistical hypothesis testing methods. These suggestions include more rigorously constraining the CA process in advance of analyzing actual CAs, using confidence intervals on the miss distance as a supplement to P_c in order to more clearly communicate the power of the decision process, and performing more effort into characterizing the underlying densities from which miss vectors for various types of conjunctions may presumed to have been drawn.

APPENDIX: SIGNIFICANCE TESTING AND CONFIDENCE INTERVALS

This appendix provides two examples of hypothesis tests involving p-values and confidence intervals that compare and contrast a classic textbook example of a test on the mean of a normal distribution with a test that approximates a CA decision. At the end of each example is a figure illustrating the concepts each discusses; readers may find it helpful to consult the figures periodically while pondering each example.

Simple Example

The following simple example has been adapted from Chapter 8 of Reference 7. Suppose that an analyst has to decide whether to accept the null hypothesis that the mean of a normally-distributed random variable, with known standard deviation of σ, is less than some specified threshold, μ_o. The analyst takes the view that μ is not a random variable; it simply has a value that is unknown to him. Denoting the set for which $\mu \leq \mu_o$ as Ω_o, we can write the null hypothesis as $\mathcal{H}_o : \mu \in \Omega_o$. He will analyze a random sample of n observations, denoted by the set $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$, drawn from the aforementioned distribution. The set of all possible observations may be divided into two disjoint subsets; the subset for which the analyst will reject the null hypothesis is the known as the *critical region*, which we denote as C. For each possible value of the true but unknown mean, μ, one can specify the probability that the analyst's test procedure will lead him to reject the null hypothesis. This probability, expressed as a function of the unknown mean, $\pi(\mu)$, is called the *power function* of the procedure, and it is given by

$$\pi(\mu) = \Pr\left(\mathcal{X} \in C | \mu\right) \tag{1}$$

The ideal power function would be a step function for which $\pi(\mu) = 0$ for all $\mu \in \Omega_o$, and $\pi(\mu) = 1$ otherwise.

If the analyst rejects the null hypothesis when the true mean is less than μ_o, the analyst considers this to be a *missed detection.* The analyst may specify an upper bound, called the *level of significance,* α_o, on the probability of such an error, and only consider procedures for which $\pi(\mu) \leq \alpha_o$. A related concept is the *size* of the procedure, defined to be the least upper bound on the power among all values of $\mu \in \Omega_o$:

$$\alpha = \sup_{\mu \in \Omega_o} \pi(\mu) \tag{2}$$

Thus, to achieve a specified level of significance, the analyst would choose procedures of sufficient size that $\alpha \leq \alpha_o$.

Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ denote the estimator of the mean based on n sample observations; the associated sample standard deviation is $\sigma/\sqrt{n}$. Note that $\bar{X}_n$ is a random variable, while for a particular set of observed samples, $\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$, the realized estimate of the mean, $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, is just a real number. Chapter 8 of Reference 7 shows that there exists some value c such that if the analyst rejects $\mathcal{H}_o$ when $\bar{x}_n \geq c$, then the power of the test is as large as possible, subject to the constraint that $\pi(\mu) \leq \alpha_o$. The level of significance for this procedure is then

$$\alpha_o = \Pr\left(\bar{X}_n \geq c | \mu = \mu_o\right) = 1 - G_{\bar{X}_n}(c \,|\, \mu_o, \sigma/\sqrt{n}) \tag{3}$$

where $G_X(x \,|\, \mu, \sigma)$ denotes the cumulative Gaussian distribution with mean μ and standard deviation σ evaluated at some value x. It will be convenient to normalize the distribution, so let $Z = \frac{X-\mu}{\sigma}$. Now let $\zeta_{\alpha o}$ denote the value of the inverse standard normal distribution such that $\Pr(Z \geq \zeta_{\alpha o}) = \alpha_o$. Then,

$$\alpha_o = \Pr\left(Z \geq \frac{c - \mu_o}{\sigma/\sqrt{n}}\right) \tag{4}$$

Thus, $\zeta_{\alpha o} = \frac{\sqrt{n}}{\sigma}(c - \mu_o)$ implies that $c = \mu_o + \zeta_{\alpha o}\frac{\sigma}{\sqrt{n}}$. In this context, the p-value associated with a particular estimate $\bar{x}_n$ is given by the probability that an estimate equal to or more extreme than the observed sample mean could have occurred,

$$p = \Pr\left(\bar{X}_n \geq \bar{x}_n | \mu = \mu_o\right) = 1 - G_{\bar{X}_n}(\bar{x}_n \,|\, \mu_o, \sigma/\sqrt{n}) = \Pr\left(Z \geq \frac{\bar{x}_n - \mu_o}{\sigma/\sqrt{n}}\right) \tag{5}$$

so that $\bar{x}_n \geq c$ is equivalent to $p \geq \alpha_o$, and either of these conditions equivalently leads to rejection of the null hypothesis that $\mu \leq \mu_o$ at the α_o level of significance.

The power function of this test is the probability of rejecting $\mathcal{H}_o$ as a function of a given μ:

$$\pi(\mu) = \Pr\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \geq \zeta_{\alpha o} + \frac{\mu_o - \mu}{\sigma/\sqrt{n}}\right) \tag{6}$$

For this example, the test is specified by the assumed standard deviation, σ, the threshold μ_o, and the significance level, α_o. With these parameters specified, the power function of the test depends only on the number of observations, although the actual power of the test will depend on the actual value of μ, which is unknown. But it is clear that as n increases for fixed σ, or as σ decreases for fixed n, c approaches μ_o from above; thus, the distance between c and μ_o gives some indication of the power the test. If the analyst were to exclusively report the p-value, both he and his audience would be missing key insight into the power of the test.

As either an alternative or a supplement to the hypothesis test just described, the analyst may wish to study probabilities of the form

$$\gamma = \Pr\left(\omega(X_1, X_2, \ldots, X_n) \leq \mu\right) \tag{7}$$

That is, if the analyst were to conduct a large number of studies, and compute a different value of $\omega(X_1, X_2, \ldots, X_n)$ from the realized values $\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$ from each study, the portion of those studies for which $\omega(X_1, X_2, \ldots, X_n) \leq \mu$ would be γ. For example, rather than forming a hypothesis about a threshold for μ that is fixed *a priori*, the analyst may wish to choose a threshold that is based on the estimated sample mean. Because the sample mean has a Gaussian distribution when the variance is known, and since $\mathrm{E}(\bar{X}_n) = \mu$, it is common to choose the threshold based on a value from the inverse normal distribution corresponding to the desired probability, in a manner similar to choosing the critical value for the significance test:

$$\gamma = 1 - \alpha = \Pr\left(\bar{X}_n - \zeta_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu\right) \tag{8}$$

Once the analyst computes the estimate, $\bar{x}_n$, the inequality

$$\bar{x}_n - \zeta_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \tag{9}$$

is no longer a statement of probability (recall that μ is not a random variable, just a fixed but unknown quantity), so in this context it is customary to refer to $1 - \alpha$ as a *level of confidence* rather than a probability, and the interval $[\bar{x}_n - \zeta_\alpha \frac{\sigma}{\sqrt{n}}, \infty)$ as a *confidence interval* [*]. The analyst might report, if he takes $\alpha = 0.05$, that he has 95% confidence that the mean is greater than or equal to $\bar{x}_n - \zeta_\alpha \frac{\sigma}{\sqrt{n}}$.

Note that $\mu_o \notin [\bar{x}_n - \zeta_{\alpha o} \frac{\sigma}{\sqrt{n}}, \infty)$ will incur if the analyst has rejected the null hypothesis according to the significance testing procedure, since $p \geq \alpha_o$ is equivalent to $\mu_o \leq \bar{x}_n - \zeta_{\alpha o} \frac{\sigma}{\sqrt{n}}$. Thus, as Chapter 8 of Reference 7 discusses in greater detail, a confidence interval corresponds to a continuum of significance tests, with each test corresponding to a value of $\mu_o \in [\bar{x}_n - \zeta_{\alpha o} \frac{\sigma}{\sqrt{n}}, \infty)$, all of which have size α. Although his conclusions could be the same with either approach, by using a confidence interval the analyst and his audience are forced to confront the precision of the estimate in a manner that they could not have if the analyst had reported only a p-value without information concerning the power the test.

Figure 2 illustrates the concepts this example has discussed. Realized values of the sample mean to the right of the yellow line would provide evidence at the α_o level of significance for rejecting the null hypothesis that $\mu \leq \mu_o$. The critical value of the test is the argument of the probability contained under the heavy red line and to the right of the intersection of the yellow and heavy red lines. The p-value corresponding to a realized sample mean $\bar{x}_n$ is the probability contained under the heavy red line and to the right of the intersection of the green and heavy red lines, which for the value shown, would indicate rejection of the null. The cyan line indicates the one-sided confidence interval associated with $\bar{x}_n$, whose confidence level is set *a posteriori* to correspond with a lower limit equal to μ_o, as will be discussed at the end of this appendix.

[*] The analyst may similarly specify an upper bound, $\mu \leq \bar{x}_n + \zeta_\alpha \frac{\sigma}{\sqrt{n}}$. The combination of the upper and lower bound produces a two-sided confidence interval, for which one would typically allocate $\alpha/2$ confidence to each side.

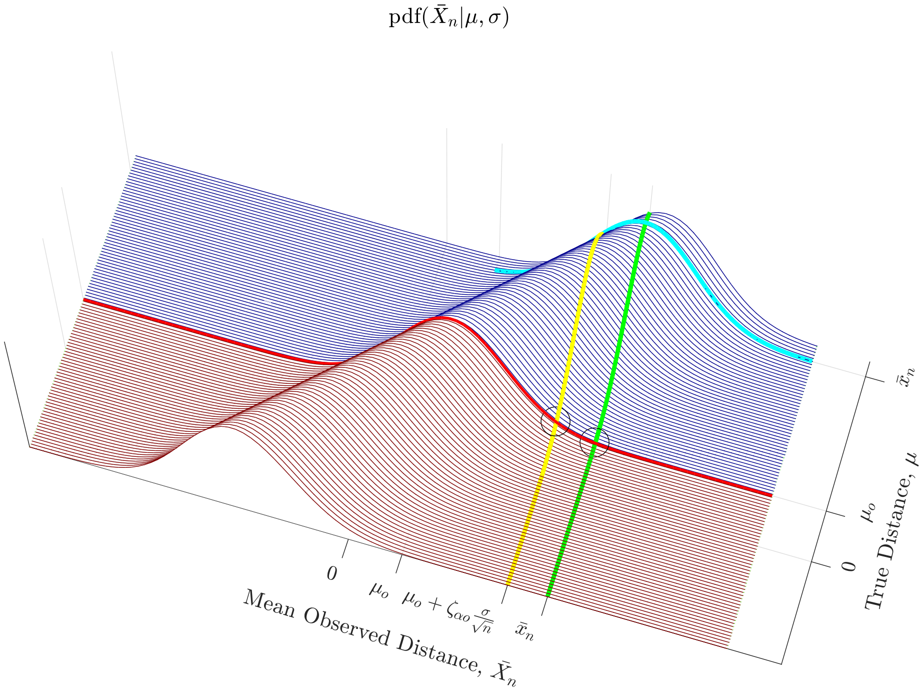

Figure 2. Illustration of p-values and confidence intervals for the hypothesis that $\mu \leq \mu_o$ for observations drawn from a Gaussian distribution with known variance.

Generalization to a Simplified CA Example

Suppose that a CA analyst has to decide whether to accept the null hypothesis that the true miss distance between two space objects, ρ, is less than some value, ρ_o, which corresponds to the combined hard-body radius of the two objects. She will analyze a random sequence of observations of the relative position vector, $\boldsymbol{\mathcal{X}} = \{\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n\}$, drawn from a Gaussian distribution with known covariances $P_1, P_2, \ldots, P_n$ and unknown mean $\vec{\mu}$, corresponding to the true miss vector, which she derives from predictions of the states of the two objects from the times $t_1, t_2, \ldots, t_n$ to the time of closest approach. Note also that $\rho = \|\vec{\mu}\|$, and hence there is some mean vector $\vec{\mu}_o$ such that $\rho_o = \|\vec{\mu}_o\|$.

She assumes that for each observation, the cumulative probability that the relative position vector predicted from time t_i is within a region defined by a set $\mathcal{D}_r$, which defines a circular disk centered on the origin such as that depicted in the upper subplot of Figure 1, where the miss distance is less than or equal to some specified value r, is given by

$$
\begin{aligned}
\Pr\left(R_i \le r\right) &= F_{Ri}(r \,|\, \vec{\mu}, P_i) && (10)\\
&= \frac{1}{\sqrt{|2\pi P_i|}} \int_{\vec{X}\in\mathcal{D}_r} \mathrm{e}^{-\frac{1}{2}(\vec{X}-\vec{\mu})' P_i^{-1} (\vec{X}-\vec{\mu})} \, \mathrm{d}^2\vec{X} && (11)\\
&= \frac{1}{\sqrt{8\pi}\sigma_{ib}} \int_{-r}^{r} \exp\left(-\frac{(\mu_b + X_b)^2}{2\sigma_{ib}^2}\right) \\
&\quad \times \left\{ \operatorname{erf}\left(\frac{\mu_a + \sqrt{r^2 - X_b^2}}{\sqrt{2}\sigma_{ia}}\right) - \operatorname{erf}\left(\frac{\mu_a - \sqrt{r^2 - X_b^2}}{\sqrt{2}\sigma_{ia}}\right) \right\} \mathrm{d}X_b && (12)
\end{aligned}
$$

where (12) follows from References 8 and 9. In (12), a and b denote the major and minor axes of the ellipse associated with the covariance P_i, σ_{ia}^2 and σ_{ib}^2 are the associated eigenvalues of P_i, μ_a and μ_b are the coordinates of $\vec{\mu}$ along the corresponding axes, and the variable of integration, X_b, is the coordinate of the vector of integration $\vec{X}$, in (11), along the minor axis.

The subset of all possible observations for which the analyst will reject the null hypothesis is the critical region, C. For each possible value of the miss distance, one can specify the probability that the analyst's test procedure will lead her to reject the null hypothesis. The power function for this procedure, expressed as a function of the unknown miss distance, is given by

$$\pi(\rho) = \Pr\left(\boldsymbol{\mathcal{X}} \in C \,|\, \rho\right) \qquad (13)$$

The analyst specifies a level of significance, α_o, on the probability of a Type I error (missed detection), and only considers procedures for which the size of the procedure,

$$\alpha = \sup_{\rho \le \rho_o} \pi(\rho) \qquad (14)$$

is sufficient that $\alpha \le \alpha_o$.

Let $\hat{\vec{X}}_n$ denote the estimator of $\vec{\mu}$ based the observations up to t_n, with covariance P_n. For a particular set of observed samples, $\{\vec{X}_1 = \vec{x}_1, \vec{X}_2 = \vec{x}_2, \ldots, \vec{X}_n = \vec{x}_n\}$, the realized estimate is $\hat{\vec{x}}_n$. Letting $\hat{R}_n = \|\hat{\vec{X}}_n\|$, then per the example above, if there exists some value c such that if the analyst rejects $\mathcal{H}_o$ when $\hat{r}_n = \|\hat{\vec{x}}_n\| \ge c$, then the power of the test is as large as possible, subject to the constraint that $\pi(\rho) \le \alpha_o$. The level of significance for this procedure is then

$$\alpha_o = \Pr\left(\hat{R}_n \ge c \,|\, \rho = \rho_o\right) \qquad (15)$$

Unlike for the previous example, it is not immediately obvious how to choose c to maximize the power of the test, subject to $\alpha \le \alpha_o$, since $\rho_o = \|\vec{\mu}_o\|$ can correspond to any point on the boundary of the combined hard body disk. Consulting Figure 1, it becomes clear that choosing $\vec{\mu}_o$ to point in the direction of the major axis of the error ellipse corresponding to P_n gives the correct value*. Any other vector would violate the constraint $\alpha \le \alpha_o$. Denoting this vector as $\rho_o \vec{u}_n^a$, where $\vec{u}_n^a$ is a unit vector along the major axis of P_n, the level of significance is

$$\alpha_o = 1 - F_{Rn}(c \,|\, \rho_o \vec{u}_n^a, P_n) \qquad (16)$$

*Recall that in this simplified example, P_n is known in advance of collecting any random observations.

and the critical value is

$$c = r_{\alpha o} = F_{Rn}^{-1}(1-\alpha_o|\rho_o \vec{u}_n^a, P_n) \tag{17}$$

The p-value corresponding to a particular estimate is then the probability that an estimate equal to or more extreme than the observed value could have occurred,

$$p = \Pr\left(\hat{R}_n \geq \hat{r}_n \,|\, \rho = \rho_o\right) = 1 - F_{Rn}(\hat{r}_n \,|\, \rho_o \vec{u}_n^a, P_n) \tag{18}$$

so that $\hat{r}_n \geq c$ is equivalent to $p \geq \alpha_o$, and either of these conditions equivalently leads to rejection of the null hypothesis that $\rho \leq \rho_o$ at the α_o level of significance. For the scenario Figure 1 depicts, $p = 47\%$, indicating that the null hypothesis could not be rejected at a common level of significance, such as $\alpha_o = 5\%$.

The power function of this test is the probability of rejecting $\mathcal{H}_o$ as a function of ρ:

$$\pi(\rho) = \Pr\left(\hat{R}_n \geq c \,|\, \rho\right) = 1 - F_{Rn}(c \,|\, \rho\, \vec{u}_n^a, P_n) \tag{19}$$

For this example, the test is specified by the covariance, P_n, the threshold ρ_o, and the significance level, α_o. As the norm of P_n decreases, c approaches ρ_o from above; thus, the distance between c and ρ_o gives some indication of the power the test. If the analyst were to exclusively report the p-value, both she and her audience would be missing key insight into the power of the test. For the scenario Figure 1 depicts, the power associated with the particular estimated miss distance the figure shows is $\pi(\hat{r}_n) = 8.6\%$, which is considerably lower than the $80\% - 90\%$ typically associated with reliable hypothesis testing results.

As either an alternative or a supplement to the hypothesis test just described, the analyst may wish to study probabilities of the form

$$\gamma = \Pr\left(\vec{\mu}_o \in \omega(\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n)\right) \tag{20}$$

A common choice for $\omega(\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n)$ might be an error ellipsoid derived from P_n, corresponding to a probability of $1 - \alpha$ that $\vec{\mu}_o$ is contained within it. Once the analyst has computed the estimate $\hat{\vec{x}}_n$, she can center the error ellipsoid on $\hat{\vec{x}}_n$ to define a *confidence region*. Another option would be to define a confidence interval for the miss distance, without regard for direction, as Figure 1 depicts, such as

$$\gamma = 1 - \alpha = \Pr\left(\hat{R}_n - r_\alpha \leq \rho\right) \tag{21}$$

Once the analyst computes the estimate, $\hat{r}_n$, the inequality

$$\hat{r}_n - r_\alpha \leq \rho \tag{22}$$

defines the $1-\alpha$ confidence interval $[\hat{r}_n - r_\alpha, \infty)$. The analyst might report, if hse takes $\alpha = 0.05$, that she has 95% confidence that the true miss distance is greater than or equal to $\hat{r}_n - r_\alpha$.

Note that $\rho_o \notin [\hat{r}_n - c, \infty)$ will incur if the analyst has rejected the null hypothesis according to the significance testing procedure, since $p \geq \alpha_o$ is equivalent to $\rho_o \leq \hat{r}_n - c$. Although her conclusions regarding the risk of the conjunction would be the same with either approach, by using a confidence interval the analyst and her audience are forced to confront the precision of the estimate in a manner that they could not have if the analyst had reported only a p-value without information concerning the power the test.

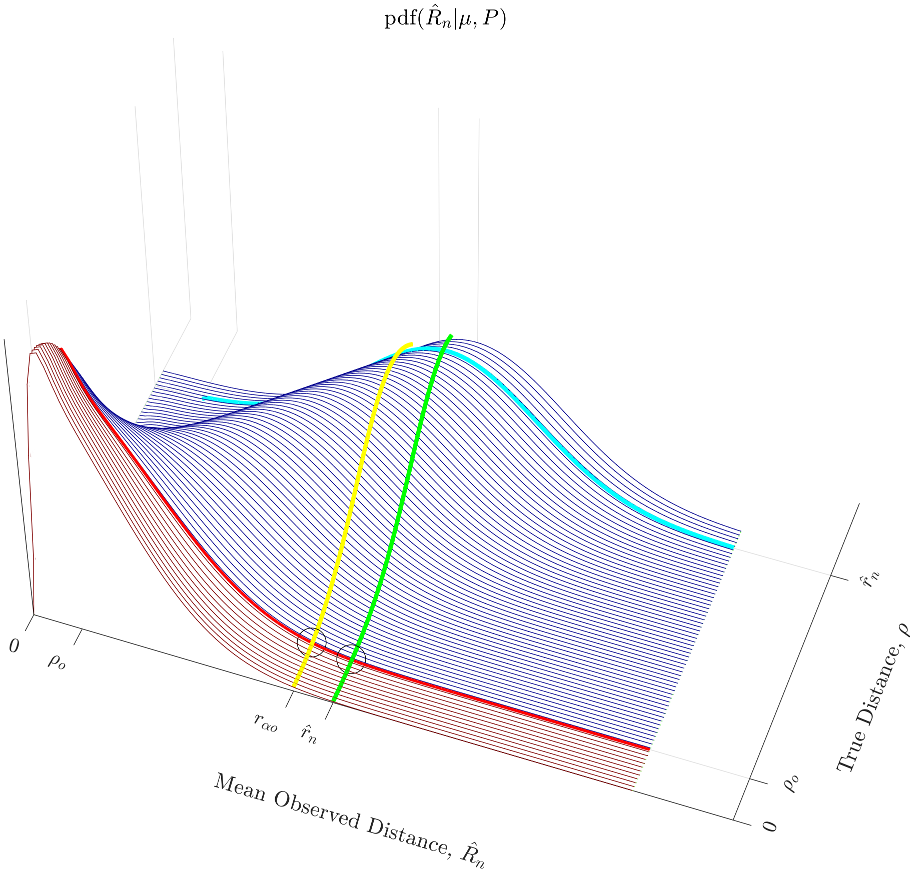

Figure 3. Illustration of p-values and confidence intervals for the hypothesis that $\rho \leq \rho_o$ for miss vector observations drawn from a Gaussian distribution with known covariance.

Figure 3 illustrates the concepts this example has discussed. Realized values of the estimated miss distance to the right of the yellow line would provide evidence at the α_o level of significance for rejecting the null hypothesis that $\rho \leq \rho_o$. The critical value of the test is the argument of the probability contained under the heavy red line and to the right of the intersection of the yellow and heavy red lines. The p-value corresponding to a realized estimate $\hat{r}_n$ is the probability contained under the heavy red line and to the right of the intersection of the green and heavy red lines, which for the value shown, would indicate rejection of the null. The cyan line indicates the one-sided confidence interval associated with $\hat{r}_n$, whose confidence level is set *a posteriori* to correspond with a lower limit equal to ρ_o, indicating a confidence of $1 - P_c$ that the secondary object will not penetrate the combined hard-body disk surrounding the primary, as will be discussed below.

Figure 4 revisits the particular result Figure 1 depicts, adding the CDF as well as the probability density function (PDF) associated with the null hypothesis, which the figure indicates with dashed lines, which are blue for the CDF and red for the PDF. This PDF, along with the PDF associated with the estimate, shown as a solid red curve, correspond to two of the infinity of PDFs that Figure 3 depicts. The dashed vertical line at a miss distance of approximate 0.55 is the value c_o for which the null hypothesis could be rejected at a 5% level of significance, if the estimated miss distance had been equal to or greater than c_o. The proximity of the dashed and solid blue curves is the reason the p-value is quite large, and the large extent of the curves in relation to the difference between the nominal estimate and c_o is the reason the power is low.

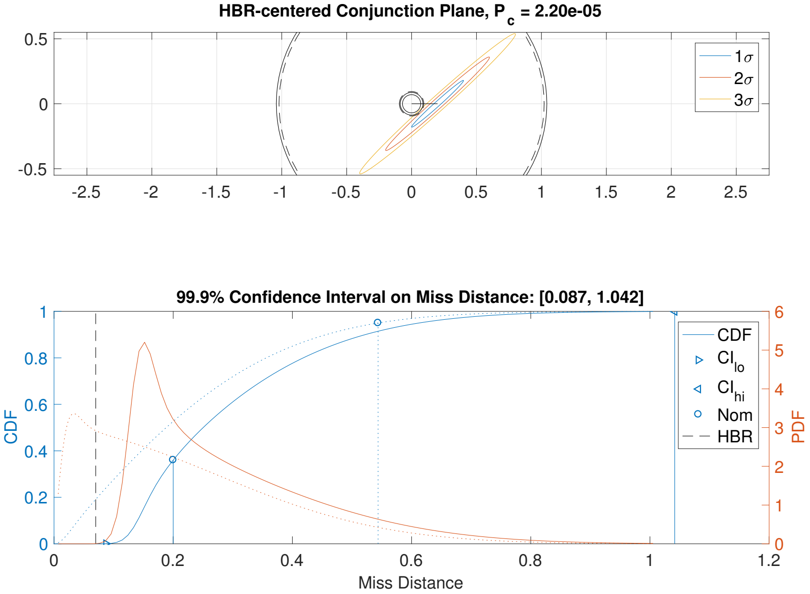

Figure 4. Update to Figure 1, with CDF and PDF associated with null hypothesis overlaid as dashed lines.

Relations among P_c, p-values, and Confidence Intervals

In operational practice, analysts compute the collision probability by taking the mean miss vector to be some nominal value, which results from the latest estimate of the predicted relative position vector, $\hat{\vec{x}}_n$. As the previous examples have shown, this differs from the computation of a p-value, in which one would assume a mean vector that corresponds to the null hypothesis of an unsafe conjunction, such as $\vec{\mu} = \rho_o\,\vec{u}_n^a$. Instead, the operational CA practice more closely resembles the computation of a confidence interval, in being based on the latest realized estimate. But as the examples have shown, an interval based on a realized estimate is not a statement of probability, which is why statisticians insist on the term confidence value for such intervals. However, rather than basing the confidence interval on a confidence value of $\gamma = 1 - \alpha$ that is fixed *a priori*, the CA practice is to fix the endpoint of the confidence limit at the combined hard-body radius, ρ_o, and let α denote the resulting collision probability. Thus, for a $1 - \alpha$ confidence interval derived from

$\Pr(\hat{R}_n - \rho_o \leq \rho)$, the "$P_c$ confidence value" would be

$$
\begin{aligned}
P_c = \Pr\left(\hat{R}_n \leq \rho_o\right) &= F_{Rn}(\rho_o \,|\, \vec{\mu} = \hat{\vec{x}}_n, P_n) && (23)\\
&= \frac{1}{\sqrt{|2\pi P_n|}} \int_{\vec{X} \in \mathcal{D}_{\rho_o}} \mathrm{e}^{-\frac{1}{2}(\vec{X} - \hat{\vec{x}}_n)' P_n^{-1} (\vec{X} - \hat{\vec{x}}_n)} \, \mathrm{d}^2 \vec{X} && (24)\\
&= \frac{1}{\sqrt{8\pi}\sigma_{ib}} \int_{-\rho_o}^{\rho_o} \exp\left(-\frac{(\mu_b + X_b)^2}{2\sigma_{ib}^2}\right) && \\
&\quad \times \left\{ \operatorname{erf}\left(\frac{\mu_a + \sqrt{\rho_o^2 - X_b^2}}{\sqrt{2}\sigma_{ia}}\right) - \operatorname{erf}\left(\frac{\mu_a - \sqrt{\rho_o^2 - X_b^2}}{\sqrt{2}\sigma_{ia}}\right) \right\} \mathrm{d}X_b && (25)
\end{aligned}
$$

So, given a realized estimate for the miss vector, an analyst could state that she had $1 - P_c$ confidence that the true miss distance is greater than the combined hard-body radius, and as with the previous examples, there exists an equivalent significance test in which P_c functions in similar, but not identical manner, as a p-value.

REFERENCES

[1] R. L. Wasserstein and N. A. Lazar, "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, Vol. 70, No. 2, 2016, pp. 129–133, 10.1080/00031305.2016.1154108.

[2] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman, "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations," *European Journal of Epidemiology*, Vol. 31, No. 4, 2016, pp. 337–350, 10.1007/s10654-016-0149-3.

[3] J. R. Carpenter and F. L. Markley, "Wald Sequential Probability Ratio Test for Space Object Conjunction Assessment," *Journal of Guidance, Control, and Dynamics*, 2014/07/15 2014, pp. 1–12, 10.2514/1.G000478.

[4] S. Alfano, "Relating Position Uncertainty to Maximum Conjunction Probability," *Astrodynamics 2003*, Vol. 116 of *Advances in the Astronautical Sciences*, Univelt, 2004, 10.1.1.372.8578.

[5] J. L. Foster, Jr., "The Analytical Basis for Debris Avoidance Operations for the International Space Station," *Third European Conference on Space Debris*, Noordwijk, The Netherlands, ESA Publications Division of the European Space Research and Technology Centre, 1997, pp. 441–445.

[6] J. P. A. Ioannidis, "Why Most Published Research Findings Are False," *PLOS Medicine*, Vol. 2, 08 2005, 10.1371/journal.pmed.0020124.

[7] M. H. DeGroot, *Probability and Statistics*. Addison–Wesley, 1975.

[8] K. T. Alfriend, M. R. Akella, J. Frisbee, J. L. Foster, Jr., D.-J. Lee, and M. Wilkens, "Probability of Collision Error Analysis," *Space Debris*, Vol. 1, No. 1, 1999, pp. 21–35.

[9] S. Alfano, "A Numerical Implementation of Spherical Object Collision Probability," *Journal of the Astronautical Sciences*, Vol. 53, Jan–Mar 2005, pp. 103–109.