Invariance, Optimality and a 1-Observation Confidence Interval for a Normal Mean

Stephen Portnoy[1]

Dedicated to the memory of Charles Stein (1920 - 2016)

Abstract

In a 1965 Decision Theory course at Stanford University, Charles Stein began a digression with "an amusing problem": is there a proper confidence interval for the mean based on a single observation from a normal distribution with both mean and variance unknown? Stein introduced the interval with endpoints $\pm c|X|$ and showed indeed that for c large enough, the minimum coverage probability (over all values for the mean and variance) could be made arbitrarily near one. While the problem and coverage calculation were in the author's hand written notes from the course, there was no development of any optimality result for the interval. Here the Hunt-Stein construction plus analysis based on special features of the problem provides a "minimax" rule in the sense that it minimizes the maximum expected length among all procedures with fixed coverage (or, equivalently, maximizes the minimal coverage among all procedures with a fixed expected length). The minimax rule is a mixture of two confidence procedures that are equivariant under scale and sign changes, and are uniformly better than the classroom example or the natural interval $X \pm c|X|$.

[1] Professor, Department of Statistics, University of Illinois at Urbana-Champaign
corresponding email: sportnoy@illinois.edu

1 Introduction and basic result

After learning that t-statistics and t intervals for normal samples require estimates of the variance, most students would find it rather surprising to learn that it is possible to find a confidence interval for a normal mean with a single observation when the variance is unknown. In fact, the existence of such a confidence interval appears to have been published first in Abbott and Rosenblatt (1963), where the interval $X \pm c|X|$ is introduced in a side comment and shown to be a proper (conservative) confidence interval. Subsequent papers introduced the idea to the engineering field and developed various generalizations (see Machol and Rosenblatt (1966, and citations), Edelman (1990), Rodriguez (1996), Wall, Boen and Tweedie (2001), among others). However, none of the earlier work considered any optimality or general comparisons in terms of interval length. This seems rather odd, especially since the interval $\pm c|X|$ was introduced in the 1965 Decision Theory Course, where Charles Stein proved it was a proper confidence interval, but failed to provide any optimality properties.

A natural objective for a confidence interval is to minimize the expected length subject to fixed coverage (that is, coverage greater than a fixed lower bound over all parameter values). This is equivalent to maximizing the (minimal) coverage probability subject to fixed expected length. This criterion has the flavor of a "minimax" procedure, and so a procedure satisfying this criterion will be called "minimax". To someone familiar with the theory of invariant statistical decision problems, this suggests the value of trying to apply invariance and results like the Hunt-Stein theorem, which shows that a procedure minimax among invariant procedures is minimax among all procedures.

Invariance greatly simplifies statistical problems. Instead of intervals of the form $[b_1(x), b_2(x)]$ with two functions $b_1(x) < b_2(x)$, it will be shown that scale-sign invariant rules have the form $[c_1x, c_2x]$ if x is positive and $[c_2x, c_1x]$ if x is negative, and so are defined by two constants that do not depend on x. Though the Hunt-Stein argument (see Hunt and Stein (1945) and Lehmann (1959)) requires randomized procedures, invariant rules are still much simpler: a randomized invariant rule is given by a fixed bivariate distribution function, $G(c_1, c_2)$ (independent of x), while a non-invariant rule requires a distribution function for each x-value. The main aim of this paper is to identify an invariant rule that is minimax among all rules.

The following contributions are developed here:

1. The simplicity of the model provides an especially clear and accessible example for the application of general principles. Section 1 introduces a natural set of confidence intervals, which are later shown to be invariant. A simple formula for coverage probabilities is developed.

2. The aim of Section 2 is to apply the invariance principle of statistical decision theory to characterize invariant confidence intervals and to use a version of the Hunt-Stein construction (Hunt and Stein (1945)) to show that there is a (randomized) invariant rule that is minimax. After a brief exposition of the paradigm of invariant statistical decision problems, the basic ideas are generalized to cover confidence intervals. Again, the simplicity of the problem here makes the application of abstract principles much more transparent and accessible. It also provides an example where non-trivial randomization is needed.

3. Section 3 finds a non-trivial mixture of two non-randomized invariant confidence intervals that is minimax. The mathematical details are sufficiently complicated to be relegated to a supplemental paper (see Portnoy (2017)), but some general principles that exploit special features of the example are developed.

4. The question of whether there are broader potential applications is broached in Section 4. A sample (for example, a time series) with a rather arbitrary dependence structure is a multivariate sample of size 1. Section 4 shows how to find a proper confidence set for the mean based on a single observation from a multivariate normal with arbitrary covariance structure. This offers a path to potential applications and, at least, provides a benchmark on what is possible with the most minimal assumptions on the covariance structure.

The source of the example is not clear. Statements attributing the example to Herb Robbins (in Rodriguez (1995) and in a personal communication from Persi Diaconis) suggest that the example was known to theoretical statisticians before Abbott and Rosenblatt (1963). This is partly corroborated by the fact that Stein's classroom interval differed from the published version.

To specify the problem, consider a single observation $X \sim \mathcal{N}(\mu, \sigma^2)$. Recognizing the importance of invariance, let $\lambda = \mu/\sigma$ and define

$$Y = X/\sigma \sim \mathcal{N}(\lambda, 1)\,. \tag{1}$$

Consider the following generalization of the intervals introduced above:

let $c_1 < c_2$ and define the interval

$$CI^* \equiv CI^*(X\,;\; c_1,\, c_2) \;=\; \begin{cases} c_1 X \;\le\; \mu \;\le\; c_2 X & X > 0 \\ c_2 X \;\le\; \mu \;\le\; c_1 X & X < 0 \end{cases} \tag{2}$$

Such rules will be shown to be scale and sign equivariant, and the primary aim is to find a mixture of two intervals of form CI^* that satisfies "minimax" optimality.

Begin by computing the coverage probability, $P\{\mu \in CI^*(X)\}$, which is the same as $P\{\lambda \in CI^*(Y)\}$ and is given by the function $P_0(\lambda;\, c_1,\, c_2)$:

Theorem 1. *The probability of coverage for the interval, CI^* for $\lambda > 0$ is:*

$$P_0(\lambda;\, c_1,\, c_2) \;=\; \begin{cases} \Phi\left(\lambda\left(1-\frac{1}{c_2}\right)\right) \;+\; 1 \;-\; \Phi\left(\lambda\left(1+\frac{1}{c_1}\right)\right) & c_1 \le 0\,;\; c_2 \ge 0 \\ \Phi\left(\lambda\left(1-\frac{1}{c_2}\right)\right) \;-\; \Phi\left(\lambda\left(1+\frac{1}{c_1}\right)\right) & c_1 > 0\,;\; c_2 > 0\;. \end{cases} \tag{3}$$

Note that the first line above holds for $c_1 = 0$ and/or $c_2 = 0$ by taking limits as $c_1 \nearrow 0$ and/or $c_2 \searrow$. The coverage probability for other cases is given from these results by symmetry, but it will be shown later that only cases with $c_2 > 0$ are needed when $\lambda \ge 0$.

Proof. To compute the coverage probability: take $\lambda > 0$, $c_1 < 0$, and $c_2 > 0$. Let $Y \sim \mathcal{N}(0,\, 1)$. Then (dividing by σ),

$$\begin{aligned} P_0(\lambda;\, c_1,\, c_2) \;&\equiv\; P\{Y \ge -\lambda,\, c_1(Y+\lambda) \;\le\; \lambda \;\le\; c_2\,(Y+\lambda)\} \\ &\qquad + P\{Y \le -\lambda,\, c_2(Y+\lambda) \;\le\; \lambda \;\le\; c_1\,(Y+\lambda)\} \\ &=\; P\{Y \ge -\lambda,\, Y \ge -\lambda\left(1-\frac{1}{c_1}\right),\, Y \ge -\lambda\left(1-\frac{1}{c_2}\right)\} \\ &\qquad + P\{Y \le -\lambda,\, Y \le -\lambda\left(1-\frac{1}{c_2}\right),\, Y \le -\lambda\left(1-\frac{1}{c_1}\right)\} \\ &=\; \Phi\left(\lambda\left(1-\frac{1}{c_2}\right)\right) \;+\; 1 \;-\; \Phi\left(\lambda\left(1+\frac{1}{c_1}\right)\right)\;. \end{aligned}$$

where the penultimate step uses the facts that $Y \ge -\lambda(1-1/c_2)$ is the weakest inequality in the first probability and $Y \le -\lambda(1-1/c_1)$ is the weakest in the second probability (and the symmetry of the normal distribution is applied). The result for $c_1 > 0$, and $c_2 > 0$ follows analogously, and those for $c_i = 0$ follow immediately, since the confidence intervals are closed. Results for $c_2 < 0$ could be obtained by symmetry, but in fact will not be needed here (see Lemma 1). □

Stein's classroom interval has the form $CI^*(X\,;\ -c,\, c)$; while the interval $X \pm c|X|$ has the form $CI^*(X\,;\ 1-c,\, 1+c)$. Thus, their coverage probabilities are immediate consequences of Theorem 1. In fact, it is not hard to minimize this coverage probability over λ for the Stein interval in closed form to show that a proper confidence interval exists. While these confidence intervals are known, apparently, CI^* has not been previously studied. Note that since the length of an interval of form CI^* is $(c_2 - c_1)\,|X|$, it suffices to consider the length solely in terms of $(c_2 - c_1)$ for comparing intervals of form CI^*. Numerical calculations give the following (approximate) values for c_1, c_2, and the length $(c_2 - c_1)$ for a 95% interval: Stein: -9.68, 9.68, 19.36; Abbott-Rosenblatt: -8.68, 10.68, 19.36; best CI^*: -9.15, 10.15, 19.30. The intervals are remarkably similar in length, with CI^* (perhaps surprisingly) located about halfway between the two earlier intervals.

2 Optimality of intervals CI^*: Hunt-Stein construction

The formal setting for discussing invariance is the paradigm of statistical decision theory. This posits an observation $X \in \mathcal{X}$ (generally multidimensional), a parameter space Θ, an action space $\mathcal{A}$, a family of distributions $P_\theta(x)$ with $X \sim P_\theta$, and a loss function $L(\theta,\, a)$ $(\theta \in \Theta,\ a \in \mathcal{A})$. A non-randomized decision rule is given by a function $d(x)$ from $\mathcal{X}$ to $\mathcal{A}$, and is evaluated in terms of the risk function: $R(\theta,\, d) \equiv E_\theta L(\theta,\, d(X))$.

The mathematical analysis of a statistical decision problem requires the powerful tools of convex analysis and regularization. Unfortunately, the set of non-randomized decision rules fails to satisfy needed convexity properties. Thus, randomized rules are introduced: the set of decision functions is taken to be the set of measures on the non-randomized rules (or nearly equivalently) the set of functions from $\mathcal{X}$ to the probability distributions on $\mathcal{A}$. The set of randomized decision functions thus becomes a sub-compact affine space (as a set of probability distributions), and so permits the use of convex analysis.

Note that the confidence interval problem here provides an elementary example where non-trivial randomization is needed. In classical estimation problems with convex loss, Jensen's inequality always allows randomized rules to be replaced by non-randomized ones; and in hypothesis testing, the randomization is only needed in a rather trivial manner to obtain exact level in discrete problems. The optimal rule found in Section 3 is a mixture of two non-randomized intervals that is strictly better than either of the components (at least if the coverage probability is not too large).

Invariance requires the introduction of a group, $\mathcal{G}$, of transformations on each of the spaces $(\mathcal{X},\, \Theta,\, \mathcal{A})$ with composition of transformations being the group operation. Denoting the operation corresponding to $g \in \mathcal{G}$ by $g \circ \cdot$, the problem is said to be invariant if for all $g \in \mathcal{G}$,

$$g \circ X \sim P_{g \circ \theta} \quad \text{and} \quad L(g \circ a,\, g \circ \theta) = L(a,\, \theta) \tag{4}$$

Given an invariant decision function, it seems reasonable to consider invariant decision rules; that is, decision functions satisfying:

$$d(g \circ x) = g \circ d(x) \qquad \text{or, equivalently} \qquad d(x) = g^{-1} \circ d(g \circ x)\,. \tag{5}$$

Such rules are sometimes called “equivariant”, since g appears on both sides of (5); but the word “invariant” is also reasonable since it applies to the space of functions (graphs) induced by the transformations on $\mathcal{X}$ and $\mathcal{A}$.

Invoking invariance tends to simplify the problem greatly. Rather than being general functions, invariant rules often have a fixed form depending on only a few constants. In some cases, there is a unique choice of the constants giving a uniformly best invariant rule. The question is: does this simplification come at a cost? The answer can be yes, definitely. Stein (1956, 1961) showed famously that best location invariant estimates of a multivariate location parameter tend to be inadmissible in 3 or more dimensions. In fact, decision rules tend to be inadmissible when both location and scale are unknown (for example, see Stein (1964)). This suggests that the rules here are not admissible, though substantial improvement seems unlikely.

Fortunately, an earlier result (Hunt and Stein (1945)) shows that best invariant rules do tend to be minimax. While the Hunt-Stein Theorem applied only to hypothesis testing problems, rather broad generalizations are available; for example, see Kiefer (1957) and Bondar and Milnes (1981). These results require that the group be “amenable”. Though there are several formal characterizations defining amenability, equation (7) provides the definition applied here to prove the optimality result in Theorem 2. Amenable groups tend to be smaller and well-behaved. They include finite dimensional Euclidean spaces under addition (multivariate location shifts), the non-zero (or positive) reals under multiplication (scale changes), any compact group, the group of triangular non-singular $(n \times n)$ matrices, but not the full linear group of all non-singular $(n \times n)$ matrices.

This reinforces a fundamental idea: in applied mathematics the hypotheses of the theorem are of critical importance and almost invariably are more important than the conclusions (which are rarely very surprising). If the group is not amenable, then best invariant rules will typically fail to be minimax. For example, consider inference on multivariate covariance matrices, which will generally be invariant under both the triangular group and the full linear group. The best invariant rules tend to be different under these two groups, and so the best invariant rule under the full linear group, which is not amenable, will generally fail to be minimax.

Confidence-interval problems require some adjustment to be put in a general decision-theoretic framework (see, for example, Kiefer (1977) and Casella and Hwang, (1991)). The main complication is that there are two "loss" functions: the indicator function of non-coverage (giving a risk function that is the negative of the coverage probability), and the length of the interval (with risk equal to the expected length). While this could be addressed, say, by using a linear combination of coverage and length, or by taking a Bayesian approach, such methods would tend to violate the fundamental requirement that the coverage probability of the confidence interval be bounded below. Alternative methods that satisfy this requirement could be no better than the optimal rules found here. Fortunately, some decision theoretic principles (including Hunt-Stein) can be applied to the two loss functions separately, and thus do not require redefining the problem (though theorems may need to be restated to cover confidence procedures).

The simplicity of the 1-observation confidence interval problem here makes application of these ideas quite transparent. The distributions are clearly invariant under the multiplicative group of non-zero reals: for any $g \neq 0$, $gX \sim \mathcal{N}\,(g\mu,\, (|g|\sigma)^2)$. However, to deal with the action space, we will consider the (equivalent) product of the group of scale changes (multiplication by positive reals) and the two-point group of sign changes. Note that a sign change in μ will reverse the confidence interval inequalities. Specifically, writing g_r for the transformation that is a scale change if r is positive and is a sign change if $r = -1$, the indicator function of coverage will be invariant if (and only if) the transformation on the endpoints is

$$g_r \circ (c_1,\, c_2) = (rc_1,\, rc_2),\;\; r > 0\,; \qquad g_{-1} \circ (c_1,\, c_2) = (-c_2,\, -c_1)\,. \qquad (6)$$

To specify an invariant loss function for the loss induced by the interval length, take $L(c_1,\, c_2, \sigma) \equiv |c_2 - c_1|/\sigma$, which is clearly invariant under the group of scale-sign changes.

Characterization of the non-randomized invariant decision rules is also immediate. Invariance requires the endpoints to satisfy $c_i(g\,x) = g\,c_i(x)$ for any positive g. Setting $g = 1/(|x|)$, we have $c_i(x) = |x|\,c_i(\operatorname{sgn} x)$. The coefficients $c_i(\pm 1)$ can take on only two values, and so sign invariance leads immediately to an interval of form CI^*. Thus, a randomized invariant rule is given by a distribution function $G(c_1,\, c_2)$ and generates the random interval $CI^*(x;\, C_1,\, C_2)$ with $(C_1,\, C_2) \sim G$. The following theorem provides optimality of invariant confidence procedures. Given the simplicity of the 1-observation problem, an argument following the classical construction that appears in the first edition of Lehmann (1959) will be sketched. The Hunt-Stein version of amenability is given as follows in Lehmann (1959): there is a sequence of probability measures, $\{\nu_n\}$, on the group such that for any (measurable) subset, B, of the group

$$\lim_{n\to\infty} |\, \nu_n(g \circ B) - \nu_n(B)\, | \;=\; 0 \tag{7}$$

for any element, g, in the group. That is, the group has (approximately) a left-invariant mean. Quite generally in statistics, $\{\nu_n\}$ is the invariant (Haar) measure restricted to a sequence of compact subsets increasing to the whole group. With this version of the condition, we show the following:

Theorem 2. *Consider a single observation $X \sim \mathcal{N}(\mu,\, \sigma^2)$, and let $F(x;\, b_1, b_2)$ be a function that is measurable in x and is a distribution function on $b_1 < b_2$, and which generates a randomized confidence interval whose lower and upper endpoints are random variables $b_1 < b_2$ with distribution function $F(x;\, b_1,\, b_2)$. Then there is a scale and sign equivariant (randomized) confidence interval given by a distribution function $G(c_1,\, c_2)$ randomly choosing a (non-randomized) interval of the form $CI^*(x;\; C_1,\, C_2)$ (see equation (2)) with $(C_1,\, C_2) \sim G$ and satisfying the following "minimax" property:*

$$\begin{aligned} \inf_{\mu,\sigma}\, E_{\mu,\sigma} P_G\{\mu \in CI^*(X;\; c_1,\, c_2)\} \;&\geq\; \sup_F \inf_{\mu,\sigma}\, E_{\mu,\sigma} P_{F(X;\,\cdot)}\{\mu \in (b_1,\, b_2)\} \quad (8)\\ E_{\mu,\sigma}\, E_G[(c_2 - c_1)\, |X|\,] \;&=\; E_{\mu,\sigma}\, E_{F(X;\,\cdot)}[(b_2 - b_1)]\;. \quad (9)\end{aligned}$$

The proof appears in the Appendix. It begins by integrating over the group to define the optimal invariant procedure (see (A-1)):

$$F^*(x;\, b_1,\, b_2) \equiv \liminf \int_0^\infty F(gx;\, gb_1,\, gb_2)\, d\nu_n(g)\;. \tag{10}$$

The results in (8) and (9) use properties of limits of integrals and change-of-variable results for group transformations.

3 An optimal randomized rule

While non-randomized invariant rules are quite simple, the need for randomization leaves a complex problem with no obvious solution. Fortunately, special features of this problem make it possible to find an optimal rule among mixtures of no more than 2 of a small list of specific non-randomized invariant intervals. The details are relatively complicated and require rather extensive analysis and development, and so they are relegated to a supplemental paper on arXiv (Portnoy, (2017)). However, some of the basic ideas can be presented easily.

Since CI^* is invariant, let $Y \sim \mathcal{N}(\lambda,\, 1)$, and take $\lambda > 0$. Let F be a distribution function on the set $\{(c_1,\, c_2)\, :\, c_1 < c_2\}$. The coverage probability can be written

$$CP = \int P_0(\lambda;\, c_1,\, c_2)\, \mathrm{d}F(c_1,\, c_2)$$

where P_0 is given by (3).

To simplify notation, refer to the interval $CI^*(Y;\ c_1,\, c_2)$ as $[c_1,\, c_2]$ (but note that endpoints are multiplied by y). First, since λ is taken to be positive, Y is also more likely to be positive; and thus it seems reasonable that c_2 should be positive (and perhaps large), and that c_1 should be greater than $-c_2$. This takes substantial analysis of the function P_0, but the arXiv paper shows that the coverage probability is increased (uniformly in λ) by moving the interval $[c_1,\, c_2]$ with $c_2 < 0$ to $[c_1 - c_2,\, 0]$ and by moving the interval $[c_1,\, c_2]$ with $c_1 \leq -c_2$ to $[-c_2,\, -c_1]$. With some work this leads to the following:

Lemma 1. *The distribution F can be restricted to one putting probability 1 on the set $\{[c_1,\, c_2] :\ -c_2 \leq c_1 \leq 1$ and $c_2 > 0\}$.*

Now, fix λ. If $P_0(\lambda;\, c_1,\, c_2)$ were concave in c_1 or c_2, Jensen's inequality would imply that the maximum over F occurs when F puts probability 1 on a fixed value for c_1 or c_2. While $P_0(\lambda;\, c_1,\, c_2)$ is not concave, it is a difference of normal distribution functions, which are concave for positive arguments and convex for negative ones. Thus, $P_0(\lambda;\, c_1,\, c_2)$ has consecutive segments that are alternatively concave and convex in c_1 and c_2. Any probability on a concave segment can be put at a point mass at the conditional mean by Jensen's inequality. Probability mass on a convex segment can be put on a two-point mixture at the interval endpoints (again having the same conditional mean) The condition on the means ensures that the expected lengths will remain the same. This will generate a mixture on a small number of intervals, though the argument requires Lemma 1 and is complicated by the discontinuity in the derivative at $c_1 = 0$ or $c_2 = 0$.

By considering the transformation $d = 1 - 1/c_2$, it is possible to show that the interval for c_2 is $c_2 \geq 1$; but the intervals for c_1 depend on λ. Thus, substantial further analysis is required to show that the mixture can be put on values independent of λ. Basically, this can be done by choosing $\lambda = \lambda^*$, the value at which the minimum over λ occurs. Finally, once the existence of an optimal mixture is shown, the endpoints of the mixture components can be fixed, and the problem of maximizing the coverage probability over the mixing probabilities subject to fixed expected length becomes a linear program. Thus only two non-zero mixing probabilities are needed. As proven in detail in Portnoy (2017), this provides:

Theorem 3. *Consider all randomized confidence intervals with expected length, $h > 0$. Then, there is a mixture of two of 8 specific intervals (see Portnoy (2017)) that maximizes the minimal coverage probability (over λ) among all rules with expected length h; that is, it is optimal in the minimax sense. Computational results described in Portnoy (2017) find constants $c_1 < a_1 \leq 0$ and a probability $p \in [0,\, 1]$ such that the p-mixture of $[c_1,\, c_2]$ and $[a_1,\, c_2]$ is numerically "minimax", where c_2 is chosen so that the mixture has length h (that is, c_2 satisfies $h = p\,(c_2 - c_1) + (1-p)\,(c_2 - a_1)$).*

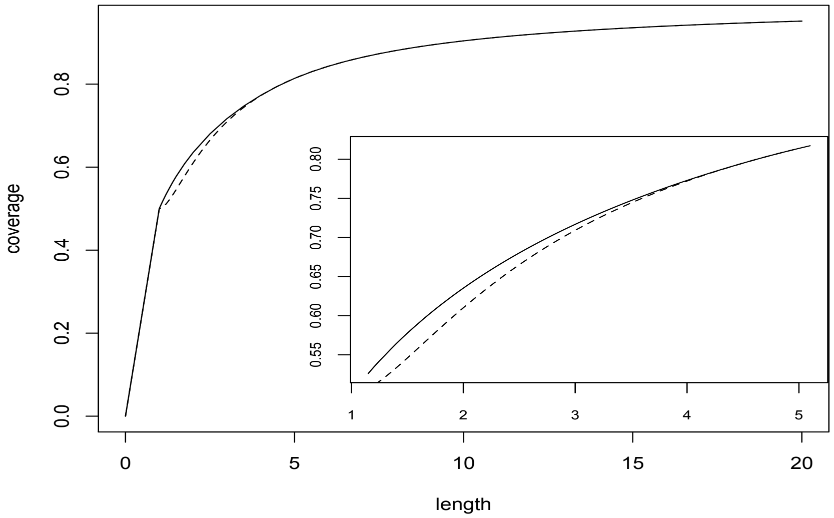

Figure 1: Best confidence interval: coverage versus expected length h Optimal mixture: solid; best non-randomized: dashed

It turned out to be surprisingly difficult to find the optimal rule by numerical optimization over the various two point mixtures given in Theorem 3. As described in Portnoy (2017), the optimal coverages were found numerically using the R-functions `optim` and `optimize` (R Core Team (2015)). After considerable refinement, the computer results appeared to be reliable, with accuracy of at least 4 decimal places. The coverages are plotted in Figure 1 as a function of interval length h, along with the coverage best interval of form CI^*. The inset focusses on $h < 5$, from which concavity of the optimal coverage for non-randomized rules can be seen in an interval above $h = 1$. It is clear that the coverages for the best CI^*'s are strictly smaller than the optimal coverages for $h < 4$; and so randomization is required here. It appears that for coverage larger than about .8, the best rule may be non-randomized; though numerical computation to 4 or 5 places cannot prove this. Finally note that the best CI^*'s have larger coverages than either the Stein intervals or the intervals of Abbott and Rosenblatt (1962) (since these are special cases).

4 A confidence set based on one multivariate normal vector

The desire to do inference with a sample of size one seems sufficiently rare to make the problem somewhat of a curiosity. However, allowing a somewhat general dependence structure within an arbitrary sample is not uncommon. For example, a time series can be considered as a multivariate sample of size one. While the imposition of a restrictive dependence structure can rescue classical rates of statistical inference, it may be quite interesting and even useful to ask what can be done under the most minimal assumptions on the dependence structure. Thus, finding a proper confidence set for the mean of a multivariate normal observation vector with arbitrary covariance matrix may be useful and potentially applicable.

Clearly, by the Bonferroni inequality, using the interval here for each of the p coordinates (with coverage bound $(1-\alpha/p)$) would provide a confidence rectangle with coverage $(1-\alpha)$. However, this rectangle would have very large volume. Is it possible to find a confidence set of the form $||\mu|| \leq c\,||X||$ with much smaller volume? In fact, the answer is yes, and the following result is proven in the supplemental arXiv paper, Portnoy (2017):

Theorem 4. *Let* $X \sim \mathcal{N}_p(\mu,\, \Sigma)$ *. Then to achieve*

$$\inf_{\mu,\, \Sigma} P\{||\mu|| \leq c\, ||X||\} \geq 1 - \alpha$$

it suffices to take $c = 3.85\, \alpha^{-1/p}$ *.*

The proof uses the density for a non-central Chi-square distribution, and requires somewhat careful analysis and bounds on the terms of this infinite series. No optimality claims are made concerning this confidence set, and, in fact, the constant, 3.85, is not sharp. I conjecture that a constant much nearer 1 will suffice. Nonetheless, the ball clearly has much smaller volume, and simultaneous confidence intervals for specific linear combinations of means would be smaller using the ball (especially for large p). In fact, a Bonferroni rectangle based on the Stein interval $\{|\mu_i| \leq c(\alpha/p)|X_i|\,\}$ would imply a norm-bounded confidence set $\{||\mu|| \leq c(\alpha/p)||X||\}$. It can be shown that $c(\alpha/p)$ must grow at rate $\sqrt{\log\, p}$ as p increases, while the bound in Theorem 4 tends to a constant (since $\alpha^{-1/p} \rightarrow 1$ as p increases).

5 Conclusions

The problem of finding a confidence interval for a mean with variance unknown provides an amusing example of the application of statistical theory. The development here shows the value of applying statistical invariance in confidence interval problems and of using convex analysis to find an especially simple form for the minimax procedure. The theoretical development here is clarified by the simplicity of the example, but the use of invariance and convex analysis is applicable in a wide variety of problems. The development here and the technical arguments in Portnoy (2017) should apply in more general problems with smooth objective functions having alternate regions of convexity and concavity. Thus, these approaches should be useful in more general parametric confidence interval problems.

Finally, the development of a confidence set for the multivariate version of the problem may help to set valuable benchmarks in certain modern "big data" problems.

6 Acknowledgments

I must acknowledge the inspiration of Charles Stein, who was my teacher, mentor, colleague, and friend. Surely, this paper would have been impossible without the training, knowledge, and encouragement he gave me. I would also like to acknowledge Mahlon Day for greatly deepening my knowledge of amenable groups, and for assuring me that by defining an amenable group as one with a "mean", the pun was intended. I would also like to thank Roger Koenker for extensive conversation over many cups of coffee, and give special thanks to the reviewers, whose comments greatly improved the paper.

7 Appendix

The following is a proof of Theorem 2:

Proof. (Sketch.) Given F, define

$$F^*(x;\, b_1,\, b_2) \equiv \liminf \int_0^\infty F(gx;\, gb_1,\, gb_2)\, d\nu_n(g) \;. \tag{A-1}$$

Compactness of the closure of the set of probability measures on the set of intervals $\{b_1,\, b_2) \;:\; b_1 < b_2\}$ allows the limit to define a randomized confidence procedure, as follows from the argument in Lehmann (1959). The limit F^* is easily seen to be equivariant: for $g > 0$

$$F^*(x;\, gb_1,\, gb_2) = F^*(x/g;\, b_1,\, b_2)\ .$$

To show his, choose $g = |x|$. Then $F^*(x; b_1\,|x|,\, b_2\,|x|) = F^*(\pm 1\,;\, b_1,\, b_2)$, and thus is given by a randomization of intervals of the form $CI^*(x;\ c_1,\, c_2)$, though not necessarily satisfying the symmetry of CI^* in equation (2). To obtain the symmetry, apply sign-equivariance using a similar (simpler) average over two group elements $\{\pm 1\}$. This will provide the randomization distribution, $G(c_1\, c_2)$ such that

$$P_G\{\mu \in CI^*(x;\ c_1,\, c_2)\} = P_{F^*(x,\,\cdot)}\{\mu \in [b_1,\, b_2]\}$$

for all x, where b_1 and b_2 are the endpoints of the interval $CI^*(x;\ c_1\,, c_2)$, and so have the form $b_i = c_j\, x$ with j depending on the sign of x.

It remains to consider the coverage and length of G. The coverage probability is

$$\begin{aligned}
& \int\!\!\int (I\{\mu \in [c_1 x,\, c_2 x]\,,\ x > 0\} + I\{\mu \in [c_2 x,\, c_1 x]\,,\ x < 0\}) \\
& \qquad dG(c_1,\, c_2)\, \varphi_{\mu,\sigma}(x)\, dx \\
= & \int\!\!\int I\{\mu \in [b_1,\, b_2]\,\}\, dF^*(x;\, b_1,\, b_2)\, \varphi_{\mu,\sigma}(x)\, dx \\
= & \liminf_{n\to\infty} \int\!\!\int\!\!\int_0^\infty I\{g\mu \in [gb_1,\, gb_2]\,\}\, dF(gx;\, gb_1,\, gb_2)\, \varphi_{\mu,\,\sigma}(x)\, dx\, d\nu_n(g) \\
= & \liminf_{n\to\infty} \int\!\!\int\!\!\int_0^\infty I\{g\mu \in [b_1',\, b_2']\,\} dF(gx;\, b_1',\, b_2')\, \varphi_{\mu,\,\sigma}(x)\, dx\, d\nu_n(g) \\
= & \liminf_{n\to\infty} \int\!\!\int\!\!\int_0^\infty I\{g\mu \in [b_1',\, b_2']\,\} dF(z;\, b_1',\, b_2')\, \varphi_{g\mu,\, g\sigma}(z)\, dz\, d\nu_n(g) \\
\geq & \inf_{\mu',\sigma'} E_{\mu',\,\sigma'} \int I\{\mu' \in [b_1',\, b_2']\}\, dF(Z;\, b_1',\, b_2')\ . \qquad \text{(A-2)}
\end{aligned}$$

Since the last value is just the coverage probability of F, the equivariant rule G has at least the minimal coverage of F.

To deal with the length, note that under the correspondence between G and F^*, $|b_2 - b_1| = |c_2 - c_1|\,|X|$, and so (9) follows by Fubini's Theorem. □

References

[1] Abbott, J. H. and Rosenblatt, J. I. (1962). Two stage estimation with one observation in the first stage, *Annals of the Institute of Statistical Mathematics, 14*, 229-235.

[2] Bondar, J.V. and Milnes, P. (1981). Amenability: A survey for statistical applications of Hunt-Stein and related conditions on groups, *Z. Wahrscheinlichkeitstheorie, 57*, 103-128.

[3] Casella, G., Hwang, J.T. (1991). Evaluating confidence sets using loss functions, *Statistica Sinica, 1*, 159-173.

[4] Edelman, D. (1990), A Confidence Interval for the Center of an Unknown Unimodal Distribution Based on a Sample Size 1, *The American Statistician, 44*, 285287.

[5] Hunt, G. and Stein, C. (*c.* 1945). Most stringent tests of composite hypotheses, unpublished.

[6] James, W. and Stein, C. (1961). Estimation with quadratic loss, in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, Vol. 1*, 361-379.

[7] Kiefer, J. (1957). Invariance, minimax sequential estimation and continuous time processes, *Ann. Math. Statist. 28*, 573-601.

[8] Kiefer, J. (1977). Conditional confidence statements and confidence estimators, *J. Amer. Statist. Assoc., 72*, 789-808.

[9] Lehmann, E. (1959). *Testing Statistical Hypotheses*, Wiley, New York.

[10] Machol, R. E., and Rosenblatt, J. (1966). Confidence Interval Based on Single Observation, *Proceedings of the Institute of Electrical and Electronics Engineers, 54*, 1087-1088.

[11] Portnoy, S. (2017). Some Theorems on Optimality of a Single Observation Confidence Interval for the Mean of a Normal Distribution, arXiv: 1702.05545 [math.ST].

[12] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL www.R-project.org.

[13] Rodriguez, C. C. (1996), Confidence Intervals From One Observation, in *Maximum Entropy and Bayesian Methods: Cambridge, England, 1994 Proceedings of the Fourteenth International Workshop on Maximum Entropy and Bayesian Methods*, Springer-Netherlands, 175-182.

[14] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, in *Proceedings of the Third Berkeley symposium on mathematical statistics and probability, vol. 1*, 197-206.

[15] Stein, C. (1964) Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean, *Annals of the Institute of Statistical Mathematics,16*, 155-160.

[16] Wall, M. M., Boen, J. and Tweedie, R. (2001) An Effective Confidence Interval for the Mean With Samples of Size One and Two, *The American Statistician, 55*, 102 - 105.