Confidence Limits

Philip B. Stark
Department of Statistics
University of California, Berkeley

Progress on Statistical Issues in Searches
SLAC National Accelerator Laboratory, Stanford, CA
4–6 June 2012

Confidence sets

- Datum $\mathbf{X} \in \mathcal{X}$ drawn from $\mathbb{P}_{\boldsymbol{\mu}}$, $\boldsymbol{\mu} \in \Theta$.
- If the random set $S(\mathbf{X})$ satisfies $\mathbf{P}_\theta\{S(\mathbf{X}) \ni \theta\} \geq 1 - \alpha$ for all $\theta \in \Theta$, S is a $1 - \alpha$ confidence set.
- Probability meaningful only before the datum is observed: If $\mathbf{X} = \mathbf{x}$, either $S(\mathbf{x}) \ni \theta$ or not.
- Connected confidence set for real parameter: confidence interval (CI).

What are Confidence Intervals Good For?

- Express uncertainty in estimates of parameters
- Also allow inferences about signs of parameters: positive, indeterminate, negative. Tukey's "three decisions."
- Short intervals desirable to minimize uncertainty, but not necessarily for sign determination: don't maximize the number of correct decisions
- Cf. 1-sided versus 2-sided hypothesis tests

Hypothesis testing

- Decision between two theories about the world: null and alternative hypotheses.
- Null: $\boldsymbol{\mu} \in \Theta_0$. Alternative: $\boldsymbol{\mu} \in \Theta_a$.
- Datum $\mathbf{X}$ drawn from $\mathbf{P}_{\boldsymbol{\mu}}$, $\boldsymbol{\mu} \in \Theta$.
- If set $\mathcal{A}_{\Theta_0} \subset \mathcal{X}$ satisfies $\mathbf{P}_\theta\{\mathbf{X} \notin \mathcal{A}_{\Theta_0}\} \leq \alpha$ for all $\theta \in \Theta_0$, $\mathcal{A}_{\Theta_0}$ is a *(significance) level* α *test*.
- The *power* of $\mathcal{A}_{\Theta_0}$ against the alternative $\boldsymbol{\mu} \in \Theta_a$ is $\inf_{\theta \in \Theta_a} \mathbf{P}_\theta\{\mathbf{X} \notin \mathcal{A}_{\Theta_0}\}$.
- Nearly always some tradeoff between level and power.
- $\mathcal{A}_\theta$ is *unbiased* if $\forall \boldsymbol{\eta} \in \Theta$, $\mathbf{P}_\theta\{\mathbf{X} \in \mathcal{A}_\theta\} \geq \mathbf{P}_{\boldsymbol{\eta}}\{\mathbf{X} \in \mathcal{A}_\theta\}$.

Duality between Tests and Confidence Sets

- *Simple* hypothesis completely specifies distribution of $\mathbf{X}$.
- Suppose have family $\{\mathcal{A}_\theta\}_{\theta\in\Theta}$ of level-α tests, one for each simple hypothesis $\boldsymbol{\mu} = \theta \in \Theta$.
- Then $S(\mathbf{X}) \equiv \{\theta \in \Theta : \mathbf{X} \in \mathcal{A}_\theta\}$ is a $1-\alpha$ confidence set for $\boldsymbol{\mu}$:

$$\mathbf{P}_\theta\{S(\mathbf{X}) \ni \theta\} \geq 1-\alpha, \quad \forall\theta \in \Theta.$$

Standard result, extremely powerful!

Univariate Location Model, Nonnegative Parameter

- Datum $\mathbf{X}$.
- $\mathbf{X} - \mu$ has cdf F.
- F has a symmetric, continuous, unimodal density $f(\mathbf{x})$, strictly decreasing for $\mathbf{x} \geq 0$ in the support of f.
- Want to learn about μ.
- Know *a priori* that $\mu \geq 0$. I.e., $\Theta = [0, \infty)$.

Nonnegative Univariate Location: Conventional intervals

- Conventional approach: make acceptance regions as small as possible for 2-sided, or as powerful as possible for 1-sided.
- E.g., take $F \sim N(0, 1)$; $\alpha = 0.05$ (95% CL).
- 2-sided interval for $\boldsymbol{\mu}$ is $[\mathbf{X} - 1.96, \mathbf{X} + 1.96] \cap [0, \infty)$.
- 1-sided upper interval is $[0, \mathbf{X} + 1.64] \cap [0, \infty)$
  (actually a 2-sided interval, but never "separates" from 0).
- 2-sided is empty if $\mathbf{X} < 1.96$; one-sided is empty if $\mathbf{X} < 1.64$.

Flip-flopping

Scientific goal may change, depending on what can be said based on the data available.

Some practitioners make upper 1-sided CI if the results are “null” (i.e., consistent with zero) but a 2-sided CI if the results are “significant” (i.e., sufficiently larger than zero).

In other fields, common to make upper 1-sided CI if results are below zero and lower 1-sided CI if the results are above zero.

If you make the decision based on the data but use 95% CI either way, the composite procedure can have much less than 95% coverage for some θ.

Feldman & Cousins (1998) Complaints

- Flip-flopping overstates the true coverage.
- If **X** is sufficiently small, both 1-sided and 2-sided traditional CIs are empty. What then?
- CIs combine goodness of fit testing with parameter estimation; Feldman & Cousins prefer to separate those functions. (Introduces problems I won't discuss.)

Feldman-Cousins "Unified" Method

- Construct acceptance region *not* to make the region as small as possible, but to consist of points with highest likelihood ratio to constrained MLE (cMLE).
- E.g.,

$$f(\mathbf{x}||\mu_{\text{cMLE}}) = \begin{cases} (2\pi)^{-1/2}, & \mathbf{x} \geq 0 \\ (2\pi)^{-1/2} \exp(-\mathbf{x}^2/2), & \mathbf{x} < 0. \end{cases}$$

- Ratio for null θ is

$$R_\theta(\mathbf{x}) = \begin{cases} \exp((\mathbf{x}-\theta)^2/2), & \mathbf{x} \geq 0 \\ \exp(-\mathbf{x}\theta - \theta^2/2), & \mathbf{x} < 0. \end{cases}$$

- Calibrate to have right level for each θ; correct coverage then guaranteed.
- Does not separate from zero until later than "flip-flop."
- Does not give empty CIs, even for large negative **x**.

Feldman Cousins (1998) Figure 10

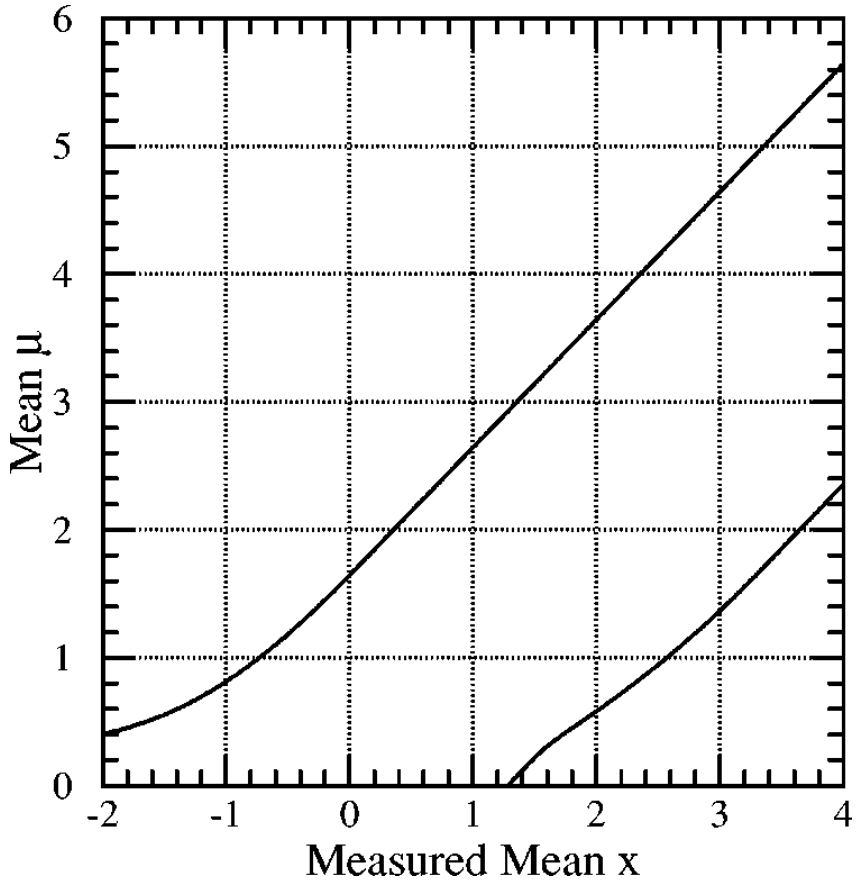

FIG. 10. Plot of our 90% confidence intervals for the mean of a Gaussian, constrained to be non-negative, described in the text.

Complaint about "Unified"

- Using a principled approach is nice, but why that principle?
- Heuristic "maximize likelihood ratio compared to cMLE" is not designed to accomplish the real goal. Why not go straight for that?
- Desirable properties of Unified Method are accidental consequences.
- Empty CIs are informative: Evidence that the model is wrong. Should not happen often if model is right.
- Unified Method never tells you the model is no good, no matter how bad it is.
  E.g., upper confidence bound at 90% is 0.4 when $\mathbf{X} = -2$.

Goals

Want to distinguish a parameter from zero, but also find a short CI for it.

One-sided tests and CIs can discriminate the parameter from zero for a smaller value of the observation than two-sided tests and CIs, but:

- must fix the sign you are looking for before looking at the data (c.f. flip-flopping)
- one-sided CIs are infinitely long—precision limited

Neyman's 1935 *three decision rule*

Observe $\mathbf{X} = \mu + Z$, where $Z \sim N(0, 1)$. Neyman's rule is

1. Decide $\mu > 0$ if $\mathbf{X} \geq z_\alpha$.
2. Decide $\mu < 0$ if $\mathbf{X} \leq -z_\alpha$.
3. Make no decision if $-z_\alpha < \mathbf{X} < z_\alpha$.

This rule controls the misclassification probability at level α:

$$\mathbf{P}_\theta\{\text{misclassifying the sign of } \theta\} \leq \alpha, \quad \forall \theta \in \mathbf{R}. \tag{1}$$

Neyman's goal was a sharper rule for classifying $\text{sgn}(\mu)$ than obtained by testing $H_0 : \mu = 0$ against a two-sided alternative and making a directional decision when $|\mathbf{X}| > z_{\alpha/2}$.

Tukey (1991)

Statisticians classically asked the wrong question—and were willing to answer with a lie . . .. They asked: “Are the effects of A and B different?” and they were willing to answer “no.” All we know about the world teaches us that the effects of A and B are always different in some decimal place . . .. What we should be answering first is “can we tell the direction in which the effects of A differ from the effects of B?” . . . The follow-up question is about how much—about what we are confident of concerning the numerical difference:

effect of A MINUS effect of B

which we abbreviate as A-B. If the first question was answered “direction uncertain” then the larger part of the answer to follow-up question is how big might A-B be... If the first question was answered “A-B positive” then the larger part of the answer to the follow-up question answers, usually: “what is the minimum size of A - B ?.” The smaller part, usually, answers: “What is the maximum of A-B?”

Distillation

- When can't classify sign, should give a short two-sided CI.
- When classify sign as positive, should have a lower endpoint larger than zero ("larger part" of the follow-up question)
- Upper endpoint should be finite ("smaller part" of the follow-up question)

Types of sign determinations

- "Sign exclusion" or "weak sign determination": CI contains values of only one sign, and possibly zero.
- (Strict) sign determination: CI contains values of only one sign, and does not contain zero.
  A strict implies weak, but not *vice versa.*
- "Separates from zero": CI contains values of only one sign, and its closure does not contain zero. If a CI separates from zero, it gives a strict sign determination.

Strategy

- Trade off *some* length for *some* observations; get sign determination for smaller observed values than CS.
- Sign exclusion almost as early as Neyman's three-decision rule.
- Exploit duality between tests and confidence sets to tailor CIs to have special properties.
- Construction analogous to Feldman-Cousins, but based on desired property of CI instead of likelihood ratio to cMLE.

Back to math

Recall $\mathbf{X} - \boldsymbol{\mu} \sim F$.

Let $c_p \equiv F^{-1}(1-p)$.

Conventional symmetric (CS) interval: $[\mathbf{X} - c_{\alpha/2}, \mathbf{X} + c_{\alpha/2}]$.

Length of CS is $2c_{\alpha/2}$.

Allow CIs to be longer than this, to determine signs more often.

Deliberately use *biased* tests, to get other desirable properties.

MP and QC: Benjamini, Hochberg, Stark (1998)

Modified Pratt and Quasi-Conventional intervals
95% confidence and 20% increase in max length vs. CS:

- MP makes weak sign determination when $|\mathbf{x}| \geq 1.656$ (c.f. 1.645 for 1-sided)
- MP makes strict sign determination when $|\mathbf{x}| \geq 1.96$ (like CS)
- MP separates from zero when $|\mathbf{x}| > 3.048$.
- QC makes weak sign determination when $|\mathbf{x}| > 1.675$
- QC makes strict sign determination when $|\mathbf{x}| \geq 1.96$
- QC separates from zero when $|\mathbf{x}| > 2.744$
- QC same length as CS when $|\mathbf{x}| > 4.419$.
- CS makes weak and strict sign determinations when $|\mathbf{x}| \geq 1.96$
- separates from zero when $|\mathbf{x}| > 1.96$.

MP & QC give weak sign determination for $\mathbf{x} \approx 15\%$ smaller than CS, while *at most* 20% longer (not on average).

The Modified Pratt (MP) Acceptance Region

Under the restriction that the density f of $\mathbf{X} - \theta$ is unimodal and symmetric, the acceptance region of the most powerful test of $\mathbf{E}\mathbf{X} = \theta$ against the alternative $\mathbf{E}\mathbf{X} = 0$ is

$$\mathcal{A}_{\mathrm{MP}}(\theta) \equiv \begin{cases} (\theta - \tilde{c}, \theta + \bar{c}), & \theta < 0 \\ (\theta - \bar{c}, \theta + \tilde{c}), & \theta > 0, \end{cases} \tag{2}$$

where $\bar{c}$ is the smaller root of

$$F(2rc_{\alpha/2} - c) = 2 - \alpha - F(c), \text{ and} \tag{3}$$

$$\tilde{c} \equiv 2rc_{\alpha/2} - \bar{c}. \tag{4}$$

Define $\mathcal{A}_{\mathrm{MP}}(0) \equiv (-c_{\alpha/2}, c_{\alpha/2})$ for symmetry.

The Modified Pratt (MP) Interval

Inverting $\mathcal{A}_{\mathrm{MP}}$ gives

$$S_{\mathrm{MP}}(\mathbf{X}) = \begin{cases} (\mathbf{X} - \bar{c}, \mathbf{X} + \bar{c}), & 0 \le \mathbf{X} < \bar{c} \\ [0, \mathbf{X} + \bar{c}), & \bar{c} \le \mathbf{X} < c_{\alpha/2} \\ (0, \mathbf{X} + \bar{c}), & c_{\alpha/2} \le \mathbf{X} < \tilde{c} \\ (\mathbf{X} - \tilde{c}, \mathbf{X} + \bar{c}), & \mathbf{X} \ge \tilde{c}, \end{cases} \tag{5}$$

with $S_{\mathrm{MP}}(\mathbf{X}) = -S_{\mathrm{MP}}(-\mathbf{X})$ for $\mathbf{X} < 0$.

For $r = 1$, MP is CS; for $r = \infty$, MP is unbounded.

MP weakly determines the sign of θ for the largest possible set of values of $\mathbf{X}$, among CIs that are never longer than $2rc_{\alpha/2}$.

MP is longer than CS (by as much as the fraction $r - 1$) when $|\mathbf{X}| > 2c_{\alpha/2} - \bar{c}$.

Quasi-Conventional (QC) Confidence Intervals

CI that reverts to CS when $|\mathbf{X}|$ is large by penalizing the size of the acceptance region.

Earlier weak sign determinations than CS by penalizing the extent to which the acceptance region crosses the origin

Leads us to seek for each θ

$$\arg\min_{\mathcal{A}}\{\lambda|\mathcal{A}| + \sup_{\mathbf{x}\in\mathcal{A}:\mathrm{sgn}\mathbf{x}\neq\mathrm{sgn}\theta} |\mathbf{x}|\} \text{ s.t. } \mathbf{P}_\theta(\mathbf{X}\in\mathcal{A}) \geq 1-\alpha. \quad (6)$$

1st term controls the length of the CI.

2nd term controls the range of values of $\mathbf{X}$ for which the CI includes parameter values with sign opposite to that of θ.

λ is a Lagrange multiplier for the constraint $|\mathcal{A}(\theta)| \leq C$.

With no penalty for the acceptance region crossing the origin, solution to the optimization problem is CS.

If choose λ so that $|\mathcal{A}| \leq C$, optimal acceptance regions are

$$\mathcal{A}_{\mathrm{QC}}(\theta) = \begin{cases} (-c_{\alpha/2}, c_{\alpha/2}), & \theta = 0 \\ (\theta - \bar{c}, \theta + \tilde{c}), & 0 < \theta \leq \bar{c} \\ (0, \theta + F^{-1}(2 - \alpha - \theta)), & \bar{c} < \theta \leq c_{\alpha/2} \\ (\theta - c_{\alpha/2}, \theta + c_{\alpha/2}), & \theta > c_{\alpha/2}, \end{cases} \tag{7}$$

with $\mathcal{A}_{\mathrm{QC}}(\theta) = -\mathcal{A}_{\mathrm{QC}}(-\theta)$ for $\theta < 0$,

$$\tilde{c} = (2r - 1)c_{\alpha/2}, \tag{8}$$

$$\bar{c} = F^{-1}(2 - \alpha - F(\tilde{c})). \tag{9}$$

Inverting the QC tests

Inverting these acceptance regions and taking the convex hull yields

$$S_{\mathrm{QC}}(\mathbf{X}) = \begin{cases} (-\bar{c}, \bar{c}), & \mathbf{X} = 0 \\ (\mathbf{X} - \bar{c}, \mathbf{X} + c_{\alpha/2}), & 0 < \mathbf{X} \leq \bar{c} \\ [0, \mathbf{X} + c_{\alpha/2}), & \bar{c} < \mathbf{X} < c_{\alpha/2} \\ (0, \mathbf{X} + c_{\alpha/2}), & c_{\alpha/2} \leq \mathbf{X} \leq \tilde{c} \\ (\mathbf{X} - \tilde{c}, \mathbf{X} + c_{\alpha/2}), & \tilde{c} < \mathbf{X} \leq \bar{c} + \tilde{c} \\ (\mathbf{X} - c_{\alpha/2}, \mathbf{X} + c_{\alpha/2}), & \mathbf{X} > \bar{c} + \tilde{c} \end{cases} \tag{10}$$

for $\mathbf{X} \geq 0$; for $\mathbf{X} < 0$, $S(\mathbf{X}) = -S(-\mathbf{X})$.

Maximum length is $\mathcal{L}(\mathcal{A}_{\mathrm{QC}}) = \tilde{c} + c_{\alpha/2} = 2rc_{\alpha/2}$.

Benjamini, Hochberg, Stark (1998) Figure 1

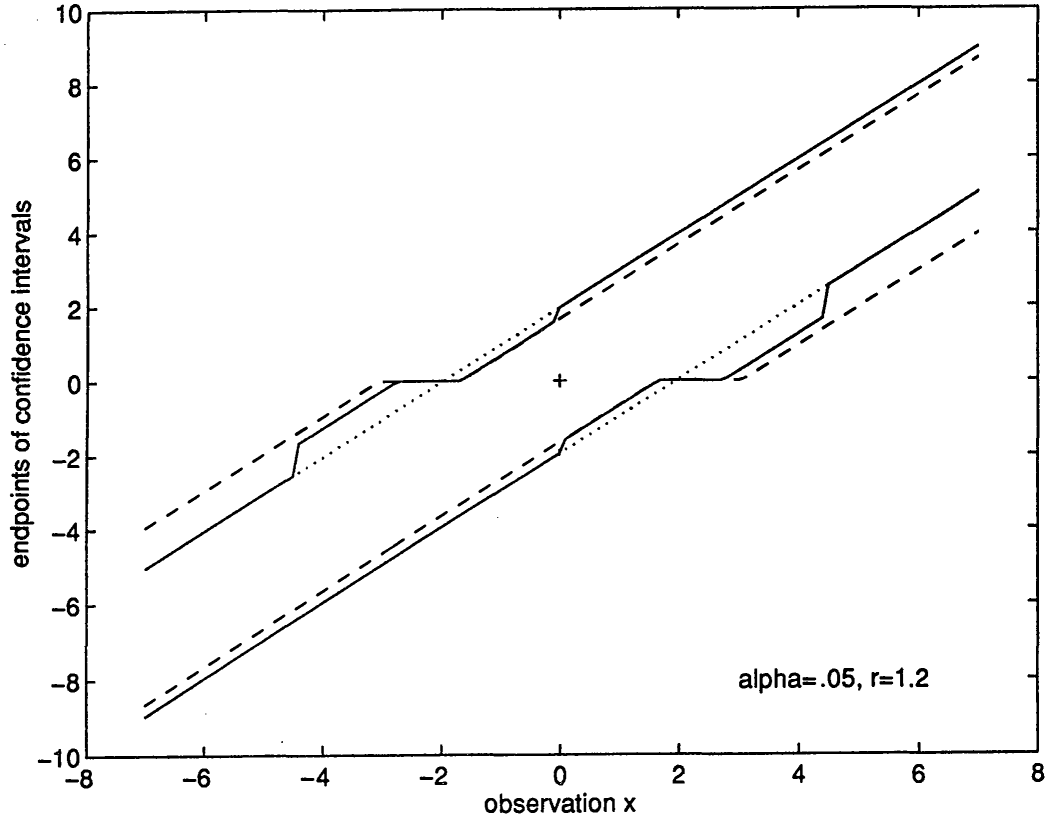

Figure 1. CS (· · ·), MP (– – –), and QC (——) 95% Confidence Intervals for a Normal Mean as a Function of the Observed Value $X = x$. The MP and QC intervals are constrained to have a length that does not exceed 1.2 times the length $2z_{\alpha/2}$ of the CS interval. For each abscissa x, the two dotted ordinates are the endpoints of the CS interval, and the dashed and solid ordinates are those of the MP and QC intervals. The MP and QC intervals are open at 0 when $|x| \geq z_{\alpha/2}$ but reach 0 for smaller values of x than the CS interval; compared to the CS interval, they have enhanced ability to exclude parameters of one sign.

Benjamini, Hochberg, Stark (1998) Figure 2

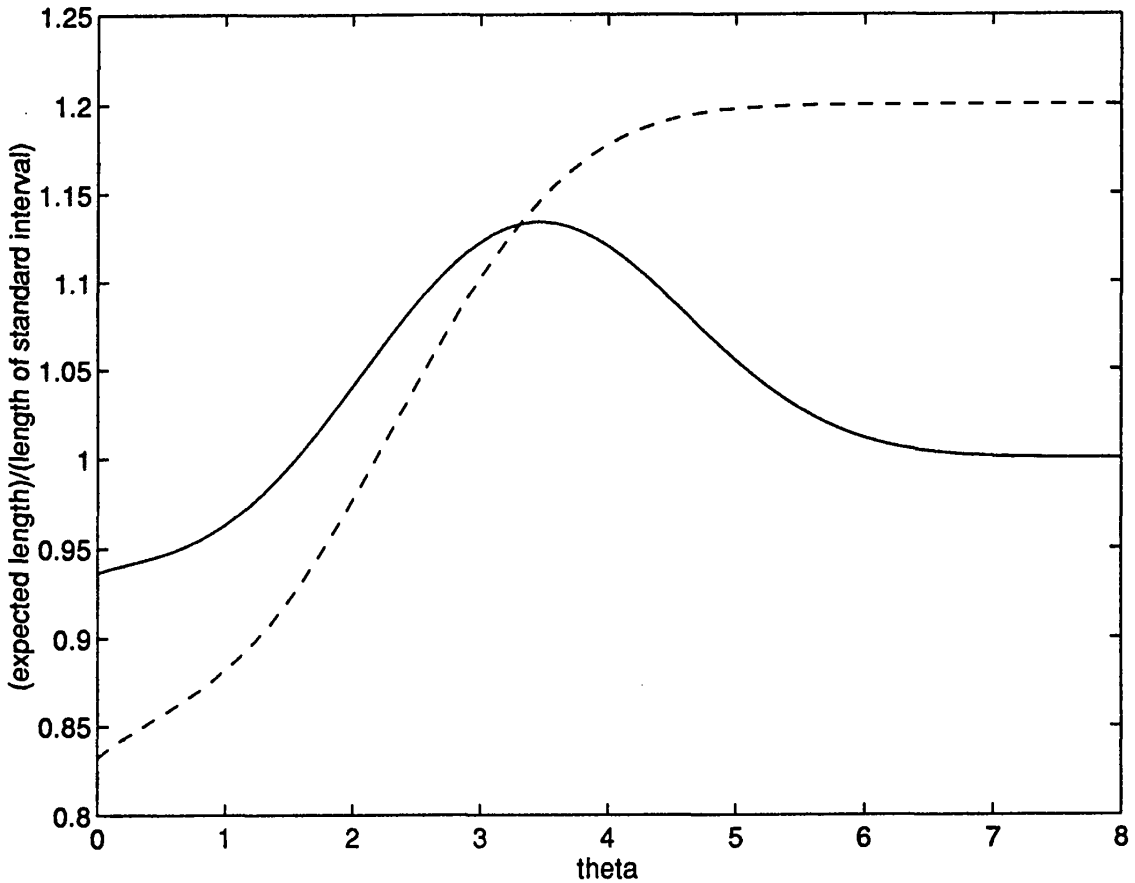

Figure 2. Expected Lengths of MP (– – –) and QC (——) 95% Confidence Intervals, Relative to the Length of the Conventional Symmetric Interval, as a Function of the True Parameter Value θ, for Observations With a Normal Distribution. The MP and QC intervals are constrained to have maximum length not exceeding 1.2 times the length of the conventional symmetric interval.

MP and QC improve on CS for this problem

MP and QC make earlier sign exclusions than CS.

MP makes earliest sign exclusions, at the cost of being longer than the CS and QC on a set of infinite measure.

QC sacrifices some power against zero and some length when $|\mathbf{X}|$ is small, but has a big length advantage over MP elsewhere.

The values of $\bar{c}$ for MP and QC can be effectively indistinguishable, but still the values of $\tilde{c}$ for the two methods differ noticeably.

Then, QC separate from zero much sooner than MP, and are ultimately much shorter. This results particularly when $\mathbf{X} - \theta$ has thin tails.

Most of the benefit from MP and QC comes with an increase in the maximum possible length over CS of 10%–20%.

With that increase, MP and QC make weak sign determinations almost as early as a one-sided test: for $\alpha = 0.05$ and $r = 1.2$, one sign is excluded for $|\mathbf{X}|$ beyond about $1.01 z_\alpha$.

For $r = 1.5$ and reasonable confidence levels, MP and QC exclude one sign for essentially the same values of $\mathbf{X}$ as one-sided tests would, but give finite-length CIs.

Compared with CS, QC gives up length exactly where it buys an earlier sign exclusion, and in the region $z_{\alpha/2} \leq |\mathbf{X}| \leq \bar{c} + \tilde{c}$, where QC still yields a strict sign determination, but one endpoint of QC is open at zero.

In contrast, MP gives up length on an infinite set.

The "cost" of MP and QC in terms of expected CI length is even less than the cost in maximum length.

Summary, further work

Can do the equivalent of MP or QC for parameters constrained to be of one sign; analogous to Feldman-Cousins: Simply intersect MP or QC with $[0, \infty)$.

Protects against flip-flopping but reveals when there is strong evidence that the model is wrong, and gives shorter expected lengths (at least for many μ).

Claim: It more sense to optimize the desired criterion—make a sign determination as soon as possible and keep length under control—rather than to use an ad hoc principle such as likelihood ratio to cMLE.

Multivariate extension for simultaneous CIs: Benjamini, Madar, Stark (2012).

References

Benjamini, Y., Y. Hochberg, and P.B. Stark, 1998. Confidence Intervals with more Power to determine the Sign: Two Ends constrain the Means, Journal of the American Statistical Association, 93, 309–317.

Benjamini, Y., V. Madar, and P.B. Stark, 2011. Simultaneous Confidence Intervals with more Power to Determine Signs. Submitted to Biometrika. Preprint: `http://statistics.berkeley.edu/~stark/Preprints/qc11.pdf`

Feldman, G.J. and R.D. Cousins, 1998. Unified approach to the classical statistical analysis of small signals, Phys. Rev. D, 57, 3873–3889.