Creating a context-specific quantitative backdrop to support conclusions and decisions

Megan Higgs, PhD Statistics, Critical Inference LLC

DRAFT Last updated on 15 February, 2024

Contents

1 Introduction

Intervals representing some sources of quantified uncertainty/variability are now a typical output from statistical methods (e.g., confidence, credible, or more generally compatibility intervals). However, the valuable information contained in the intervals is often still ignored in favor of a focus on point estimates and p-values, a practice that contributes to conflating the magnitude of p-values with the practical meaning of implications of the results (i.e., treating statistical "significance" as if it represents practical importance). More effective statistical practice carefully assesses the range of values in the interval - which are those deemed most compatible with the data and all background assumptions. This process can be greatly facilitated by the development of an *a priori* established backdrop in front of which resulting intervals will be compared for more meaningful and justified interpretations prior to stating conclusions or making decisions. An important part of the process is also making an *a priori* plan for how intervals with different locations and widths will realistically be used to inform decision making. I use "decision" in this document, though terms such as conclusions, recommendations, plans for future research, etc. may seem more appropriate for some contexts.

2 Outline of the process

The process can be described by four phases: (I) Describe the context for the quantity of interest and decisions to be made based on a study to estimate the quantity, (II) Construct the quantitative backdrop assuming you can get the desired quantity (no uncertainty), (III) Consider different possible intervals that could be obtained for estimating the quantity and document a plan for what decision would accompany each, and (IV) consider limitations of the intervals to be used based on study design, data to be collected, and reasonableness of model and other background assumptions. An optional web applet can facilitate phases II and III, as

well as support collaboration with others in the process. This working document does not currently contain examples, but they will be added in the future.

2.1 Phase I - Describe the desired quantity and decision

1. Describe the desired quantity of interest in the context of the larger research problem. The desired quantity should be the quantity you would ideally base your conclusions/decisions on. For now, ignore any challenges in measurement or design that make actually obtaining the desired quantity difficult.
    - *Comment*: In general, the desired quantity is unknown and impossible to actually obtain directly. While the quantity will ultimately be estimated/predicted, we can ignore that challenge for Phase I and Phase II of this process.
2. Describe how the desired quantity, if it could be directly obtained, would be used to inform decisions.
    - *Comment*: Use a natural quantitative scale that is able to be interpreted by a wide range of stakeholders. Standardized effect sizes or percent changes are often difficult to attach to the context (see Phase II).

2.2 Phase II - Develop the quantitative backdrop picture

Phase II entails the actual construction of the backdrop and has six steps.

1. Consider plausible values of the quantity of interest described in Phase I, again on a scale that can facilitate interpretation within the problem context. Sketch a number line covering the entire range of values deemed at all possible; "zoom in" on the part of the number line relevant to the defined quantity and problem.
    - *Web applet*: Choose values for the minimum and maximum values displayed.
2. Spend some time considering how different values of the desired quantity would affect the decision described in Phase I. This part of the exercise will depend on knowledge and honest consideration of how the quantity would be used assuming the quantity is directly available with no uncertainty. The next steps will help formalize the outcome of this thought exercise.
    - *Comment*: Different stakeholders may have different responses for some values of the quantity depending on understanding of the system and values regarding different risks. This is expected and simply requires explanation by the individual.
3. Starting at the left side of the number line, sequentially consider larger values until you reach a magnitude where it would be difficult to make a decision one way or the other. Color the range of values that would *definitely* support the decision associated with small values. In other words, choose a largest value that would still clearly result in the decision associated with small values and color the region to the left.
    - *Web applet*: Enter the maximum value that would clearly lead to the decision associated with smaller values and then everything to the left will be colored blue. (See the screenshot in Section 2.3.)
    - *Comments*: Different stakeholders may choose different numbers depending on their understanding of the system and the values they hold regarding different risks. This is expected and simply requires explanation by the individual, which hopefully motivates important discussions. Realistically, the choice of value will always feel somewhat arbitrary and contrived because it does require specification of a single number.
4. Starting at the right side of the number line, sequentially consider smaller values until you reach a magnitude where it would be difficult to make a decision on way or the other. Color the range of values that would *definitely* support the decision associated with larger values. In other words, choose a smallest value that would still clearly result in the decision associated with large values and color the

region to the right using a different color than Step 4. There will be separation between the region specified in this step and that colored in Step 3.

- *Web applet*: Enter the minimum value that would clearly lead to the decision associated with larger values and everything to the right will be shaded green. There will be a region between the two colored regions that represents gray area between values that would clearly lead to different decisions. The choice of the cutoffs is necessary for creating the backdrop, but ultimately the backdrop is designed to be interpreted as a continuum from one color, through gray area, and into the the other color.

5. After the previous two steps, there should be a region with no color between the two colored regions because the shift from values definitely aligned with one decision to those clearly aligned with the opposite decision rarely manifests as a sharp threshold. This in-between region represents "gray area" illustrating the realistic gradual transition between the two decisions where it is too difficult to say that the values support one decision over the other. That is, if the value is in this region, it would be very hard to use the information to support a decision in one direction or the other. This region, while often ignored in practice, is important when considering how quantitative results will be used to inform the decision making process.

    - *Comments*: There may be situations when a strict numeric cutoff was already in place before doing this exercise. If this is the case, it is still useful to go through the process of specifying the gray area region and considering the region relative to how decisions will be made.

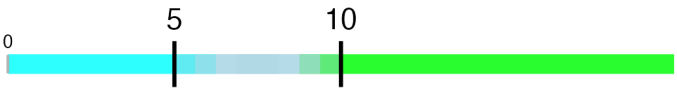

Figure 1: Example picture of a backdrop. This is a screenshot from the web applet.

6. If needed, iterate through the above steps after obtaining feedback from other experts and stakeholders. The iterating should be done before data analysis and interpretation.

    - *Comments*: As mentioned previously, differences in colored regions are expected from different stakeholders. These differences can motivate discussions and collaborations early on in the process, which is a valuable part of this approach. Ideally, general agreement can be reached, but if not then multiple options can be used with documentation and justification for the differences, as well as a plan to ultimately deal with differences in interpretation relative to decision making.

2.3 Phase III - Incorporate hypothetical uncertainty into the decision-making process

Phase III now acknowledges that the desired quantity will ultimately be estimated and thus will associated with uncertainty typically represented by an interval. This phase still appeals to a best-case scenario by assuming the intervals considered can be fully trusted for conclusions and decision making. Therefore, this phase adds one layer of challenge to the decision-making process without having to wrestle with trust or uncertainty in the interval itself.

The goal of this phase is to walk through the process of considering and justifying different conclusions and decisions under many hypothetical outcomes in the form of intervals *before* actually having the results. Different possible scenarios and associated conclusions/decisions can be documented and discussed among stakeholders before data are collected or analyzed. Having a plan for interpretation laid out before analysis can make the process after analysis easier and better justified - and can help protect against bias in interpretations. Ideally, stakeholders would agree on how the results will be used under different scenarios; if agreement isn't possible, at least there is time for discussion before results are available, including a path forward if quantified uncertainty is too large to support a specific decision.

Document different scenarios and a description of what the associated decision/conclusion would be, as well as ones that remain challenging and need further work. Screenshots of the app can be used to help with documentation of the scenarios.

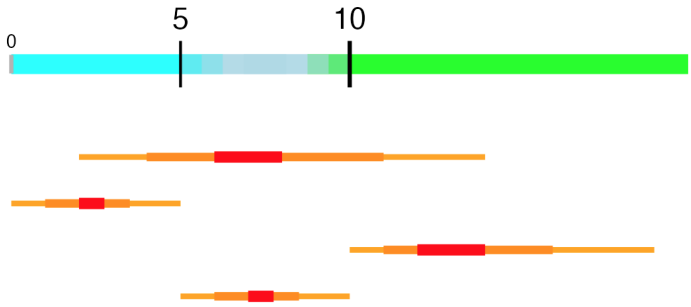

Figure 2: Example picture of a backdrop along with hypothetical intervals. This is a screenshot from the web applet.

2.4 Phase IV - Acknowledge limitations of intervals in real-life

Up until this point, the exercise has been mainly hypothetical. However, at some point the limitations that come from estimating the desired quantity and quantifying uncertainty must enter, and of course this adds challenges and discomfort. It can be tempting to wait until after data analysis when intervals are in-hand to engage with this part of the process, but considering it in the design phase provides valuable motivation for improving the design and/or analysis, or at least planning ahead for the challenges that will arise in interpretation and justifying results and decisions. This process should then be re-visited post-analysis before using the intervals obtained from a statistical method with the results of Phase III developed before analysis.

The intervals obtained in real-life are theoretically valid *if* all background assumptions hold, though of course that is rarely, if ever, the case in real life. In an ideal, though unrealistic, situation uncertainties in assumptions would be quantified and then propagated through the model to fully capture all sources of uncertainty in the intervals. Clearly this is impossible in real life, but the thought exercise can make it clear that intervals we do create are necessarily too narrow if we acknowledge the underlying assumptions. Unfortunately, there is no magic formula to widen the intervals by an appropriate amount, as the degree to which uncertainty is understated depends on the specific violations and the severity of those violations. At some point, we end up in the situation of having to qualitatively assess unquantified uncertainty through extra information about the design and system and judgments about how much to trust the results based on assessment of the background assumptions. Adding a dose of reality is needed to avoid over-relying on the intervals for conclusions and decision-making, and this generally comes in the form of recognizing intervals are too narrow and/or centered in a misleading location.

Here are some leading questions that can help start the difficult process of Phase IV:

- What are reasons the interval obtained from available data and methods/models used may *not* adequately represent plausible values of the desired quantity? What are some reasons that would lead you to question how heavily you rely on the interval as a basis for decision making?
- What are some reasons an interval coming straight from statistical methods might not reflect important sources of variability or uncertainty (i.e. reasons it may be considered too narrow)? How substantial are the missing sources? How much will loosely assuming the intervals should be wider potentially change decisions?
- What are some reasons the intervals coming straight from statistical methods might be misleading in their location (i.e. reasons they may be shifted relative to the desired quantity)? This can come from

such challenges as measurement, sampling bias, or simply limitations in translating the desired quantity into the parameter of a statistical model. Do you have enough information to go through the thought exercise of how specific shifts would potentially change decisions?

The implications for interpretation and decisions will depend on where the obtained intervals lie relative to the backdrop from Phase II - is there plenty of wiggle room before conclusions change or are things already on the cusp with the obtained intervals?

Part of what may come out of this exercise is the awareness that you're unsure about how to go forward given these questions. If this is the case, it may help to engage with those who have more expertise in the modeling strategies used or more information about design and data collection relative to the system being studied and the desired quantity. It can be very difficult to translate theoretical assumptions into the practical context to assess how sensitive results may be to assumption violations, but this should be something statisticians or other quantitative scientists can help with.

- *Comments*:
  - There are technically no restrictions on how the intervals are created, though typically they will represent sources of uncertainty quantified through use of a statistical model in the form of confidence, credible, or more generally, compatibility intervals. As previously mentioned, each of these rely on a large set of background assumptions that should be acknowledged and checked when possible. Ultimately, the goal is obtain an interval that provides reasonable information about values of the desired quantity such that it is trusted by stakeholders with different interests and/or values. Distributions can also be used in this process.
  - Relying on the concept of compatibility is recommended to provide greater flexibility in terms of acknowledging background assumptions Confidence and credible intervals can be interpreted as compatibility intervals – the range of values most compatible with the model and all other background assumptions given the available data. In general, intervals are interpreted as a range of values compatible with the data *and* all background assumptions used to construct it. See Rafi and Greenland (2020) and draft of Greenland et al. (2023) for more detail.
  - Part of using a statistical model to estimate the quantity of interest is translation of the desired quantity to a parameter of a statistical model. The potential simplifications and assumptions made in this translation are often ignored later in the process, but are included in the reference to "background assumptions" and affect how trustworthy intervals are to support the decision process.

3 Web app and drawings

A web applet provides a quick way to create a visual representation of the quantitative backdrop, simply depicted as a colored number line (Phase II). Then, up to four hypothetical intervals conveying plausible values of the quantity of interest can also be plotted below the backdrop (Phase III). The exercise can just as easily be carried out on paper, but the app may be helpful initially, for collaboration among stakeholders, and/or at the end to save a picture of the backdrop and different scenarios for intervals (easily done with a screenshot).

The app can be found here. Note the app is still under construction and will not be live while edits are being made. Default values may also change depending on the context I'm currently working with.

Inputs:

- As described in Phase II, two values are needed to delineate regions associated with different decisions and to define the gray area region between them. The lower value is the largest value that would definitely lead to one decision and the larger value is smallest value that would definitely lead to the other decision.
- Hypothetical intervals representing a quantified range of values deemed consistent with the data and methods/models used can be added below the number line. As described in Phase III, this facilitates the

process of thinking through decisions and wording under different realistic outcomes early in the research process. The number of quantiles is an input to allow for a more flexible display beyond a simple line defined with hard endpoints. A traditional interval represented by a segment with constant weight is obtained using 2 quantiles (e.g. 0.025 and 0.975 to display the ends of a 95% interval). Including more quantiles/intervals allows more information about the distribution underlying an interval and helps avoid the misinterpretation of a uniform distribution within the interval. Multiple intervals based on different interquantile regions are distinguished by different line weights and colors to represent the collection of several intervals under different criteria (i.e., include the 99% interval, the 95% interval, the 80% interval, and the 50% interval).

Outputs:

- Screenshots of the output included

4 Other comments and considerations

This section contains additional comments that may be helpful in making sense of the approach and it connects to other aspects of statistical inference.

- Connection to sample size calculations: The work done in Phases I, II, and III can directly inform sample size decisions that align with the context and decisions to be made. This builds off of precision-based (e.g. confidence interval based) methods, as opposed to those based on Type I and Type II error rates (and power). See for example Chapter 24 in Ramsey & Schafer 2013, Higgs 2019, Greenland 1988, and/or Rothman & Greenland 2018. The quantitative backdrop framework described here explicitly adds the quantitative backdrop development and provides a larger framework built around preparing for interpretation in the face of uncertainty after analysis by doing work before the study begins – which inherently involves study design considerations such as sample size calculations.
- There are essentially three situations that could lead to the conclusion that the interval is not useful for supporting one decision or conclusion over the other: (1) The interval is trusted, but covers a wide range of values spanning those indicative of both decisions and the gray area between; (2) the interval is trusted but falls completely or mostly within the gray area region that was specified before data analysis to not support one decision over the other; or (3) the interval is not trusted given identified limitations or other questions about the background assumptions and potential adjustments to the obtained intervals to reflect the limitations would lead to situations such as those described in (1) and (2).
- At an extreme end, it is possible the process leads to the realization that there won't be (or isn't) an interval that is trustworthy enough to be relied on for decision making in front of the developed backdrop. While frustrating, this is clearly valuable information and the process of creating the backdrop remains helpful with future efforts to obtain more precise and/or trustworthy intervals. In general, the backdrop can be completely reasonable and useful, as well as the work in Phase III relative to hypothetical intervals, but the inferential process ultimately relies on an interval to compare to the backdrop that can be trusted and relied upon.
- Unfortunately, intervals are typically presented with hard ends – which can inadvertently interpreted *as if* there is no chance the quantity of interest may fall outside the interval and equal probability of it falling anywhere in the interval. This is, of course, a mis-interpretation in multiple ways and the intervals are gross over-simplifications used make things easier for the user in terms of presentation and interpretation. I encourage envisioning of a distribution behind each interval and caterpillar-like displays that better acknowledge while still providing the conveniences associated with intervals. In general, it is helpful to relax the rigidity implied by how intervals are represented. The web applet provided with this document allows the user to include multiple intervals based on different quantiles, rather than just single endpoints (see Figure 2 for an example screenshot).
- As previously described, but worth repeating here, intervals quantified through statistical models are *conditional* on many assumptions and decisions, and these are often difficult to assess adequately.

"Background assumptions" include everything the approach relies on for results to be considered valid, trustworthy, and thus useful. This includes many things beyond the recognizable lists of model assumptions provided in textbooks or tutorials, such as distributional assumptions, constant variance, linearity, and independence of residuals. Background assumptions are related to other aspects of design, data collection, and analysis including no selection bias, omitted confounding variables, negligible remaining measurement error or model selection uncertainty, appropriateness of included and omitted variables, and no selective reporting of results based on statistical summaries. A yet deeper level includes such assumptions such as no intentional or unintentional mistakes in data collection, data entry, analysis, documentation, or interpretation. This is not an exhaustive list, but just meant to expand the typical idea of all that is conditioned upon. Given this list, it is clear that the uncertainty quantified via a statistical model is necessarily a simplification and understatement of all uncertainty in the ultimate quantity of interest. This does not mean the intervals or distributions should not be used, but language should reflect the conditionality and limitations and then that should be carried over into how they are used in practice.

5 Related references

Higgs, M.D. (2019). Critical Inference Blog post. https://critical-inference.com/sample-size-without-power-yes-its-possible/

Rafi Z and Greenland S. (2020). Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Med Res Methodol*, 20:244.

Rothman KJ and Greenland S (2018). Planning study size based on precision rather than power. *Epidemiology*, 29, pp 599–603.

Greenland S. (1988). On sample-size and power calculations for studies using confidence intervals. *Am J Epidemiol,* 128, pp 231–7.

Greenland S, Rafi Z., Matthews, R., and Higgs, M.D. (in progress). To aid scientific inference, emphasize unconditional descriptions of statistics. ArXiv190908583 StatME. 2020; https://arxiv.org/ abs/1909.08583.

Ramsey, F., and Schafer, D. (2013). *The Statistical Sleuth*, third edition. Brooks/Cole. Chapter 23.