Measuring Severity in Statistical Inference

Ruobin Gong (*Rutgers University*)

version: November 8, 2021

1 Introduction: *modus tollens* in statistical inference.

Statistical inference mimics the logical form

$$x \text{ (data)} \to \Theta \text{ (claim)}.$$

The degree of warrant enjoyed by the claim depends on the data and the strength of the "$\to$" relation. **Severity** [3, 2] measures the strength of its contrapositive:

$$\neg\Theta \to \neg x.$$

Large severity affirms the support of x for Θ.

2 Quantifying severity in statistical inference.

Let $x \in \mathcal{X}$ be the observable data, $\mathbf{C} = \{C_1, C_0\}$ be a pair of *critical regions* that partition $\mathcal{X}$, and $\{\Theta_0, \Theta_1\}$ be a pair of *inferential conclusions* that partition $\mathbf{\Theta}$.

***Definition* 2.1 (inferential claim).** The function $\mathtt{T}(\cdot;\mathbf{C}) : \mathcal{X} \to \{\Theta_0, \Theta_1\}$ is a binary *inferential claim* based on x, where

$$\mathtt{T}(x;\mathbf{C}) = \begin{cases} \Theta_1 & \text{if } x \in C_1 \\ \Theta_0 & \text{if } x \in C_0. \end{cases} \tag{2.1}$$

***Definition* 2.2 (inferential strategy).** Suppose for every $a \in \mathcal{A}$, $\mathbf{C}_a = \{C_1(a), C_0(a)\}$ is a partition of $\mathcal{X}$ such that for $a, a' \in \mathcal{A}$ where $a < a'$, $C_1(a) \subseteq C_1(a')$. Then, the collection of inferential claims

$$\mathcal{T} = \{\mathtt{T}(\cdot;\mathbf{C}_a) : \mathcal{X} \to \{\Theta_0, \Theta_1\}, a \in \mathcal{A}\} \tag{2.2}$$

is called an *inferential strategy*.

An inferential strategy provides a sense of *informativeness* conveyed by the data. A data value is more informative about a conclusion if more inferential claims from the same strategy arrive at that conclusion with that data.

***Definition* 2.3 (data informativeness).** Let $x, x' \in \mathcal{X}$, and $\mathcal{T}$ be an inferential strategy whose claims are indexed by $\mathcal{A}$. Say that x is *more informative* than x' about Θ_i with respect to $\mathcal{T}$, if there exists some $a \in \mathcal{A}$ such that $\mathtt{T}_a(x) = \Theta_i$ and $\mathtt{T}_a(x') = \overline{\Theta}_i$, for $i = 0, 1$.

***Corollary* 2.1.** *If x is more informative than x' about Θ_i w.r.t $\mathcal{T}$, then x' is more informative than x about $\overline{\Theta}_i$.*

***Definition* 2.4 (severity).** The *severity* of an assertion $\Theta \subseteq \mathbf{\Theta}$, as substantiated by an inferential claim $\mathtt{T}$ (Def. 2.1) derived from the inferential strategy $\mathcal{T}$ (Def. 2.2) and based on data x, is

$$S(\Theta; \mathtt{T} = \Theta_i) = \inf_{\theta \notin \Theta} P_\theta(X \notin C_i^x), \qquad i = 0, 1, \tag{2.3}$$

where $X \sim P_\theta$, $\theta \in \mathbf{\Theta}$, and

$$C_1^x = \cap_{a:\mathtt{T}(x,\mathbf{C}_a)=\Theta_1} C_1(a), \tag{2.4}$$

$$C_0^x = \overline{C}_1^x \tag{2.5}$$

are the respective *attained* critical regions.

***Remark* 2.1.** *Modus tollens*:

- $X \notin C_i^x$: hypothetical or future realization of X results in an inferential claim that is the opposite of the one drawn from the current evidence x;
- $\theta \notin \Theta$: as the parameter varies in the range that is complement to Θ.

***Remark* 2.2.** $X \sim P_\theta$:

- (*Frequentist*) sampling distribution indexed by $\theta \in \mathbf{\Theta}$;
- (*Bayes*) conditional prior or posterior predictive distribution marginalized over nuisance parameters.

3 Severity: properties.

A measure of evidential support should satisfy (see [4]):
- *Coherence*: the larger the hypothesis is, the more support there is;
- *Informativeness*: the farther into the hypothesis the data are, the more support there is.

***Theorem* 3.1 (coherence).** *For a pair of assertions $\Theta, \tilde{\Theta}$ satisfying $\Theta \subseteq \tilde{\Theta}$,*
$$S(\Theta; \mathtt{T} = \Theta_i) \le S\big(\tilde{\Theta}; \mathtt{T} = \Theta_i\big) \tag{3.1}$$
for all $\Theta \in \boldsymbol{\Theta}$ and $i \in \{0,1\}$.

***Theorem* 3.2 (informativeness).** *If the data x is more informative than x' about Θ_i with respect to $\mathcal{T}$ for some $i \in \{0,1\}$, then for all $\Theta \in \boldsymbol{\Theta}$,*
$$S(\Theta; \mathtt{T}(x) = \Theta_i) \ge S\big(\Theta; \mathtt{T}(x') = \Theta_i\big).$$

The severity enjoyed by an assertion Θ depends on the design of the inferential strategy $\mathcal{T}$, rather than the specific inferential claim derived from it.

***Theorem* 3.3 (strategic equivalence).** *For fixed data $x \in \mathcal{X}$, let $\mathtt{T}_a$ and $\mathtt{T}_{a'}$ be two inferential claims derived from the inferential strategy $\mathcal{T}$ such that $\mathtt{T}_a(x) = \mathtt{T}_{a'}(x) = \Theta_i$, where $i \in \{0,1\}$. Then for all $\Theta \in \boldsymbol{\Theta}$,*
$$S(\Theta; \mathtt{T}_a(x) = \Theta_i) = S(\Theta; \mathtt{T}_{a'}(x) = \Theta_i). \tag{3.2}$$

***Theorem* 3.4 (subadditivity).** *For fixed data $x \in \mathcal{X}$, let $\mathtt{T}_a$ and $\mathtt{T}_{a'}$ be two inferential claims derived from the inferential strategy $\mathcal{T}$ such that $\mathtt{T}_a(x) = \Theta_i$ and $\mathtt{T}_{a'}(x) = \overline{\Theta}_i$, where $i \in \{0,1\}$. Then for all $\Theta \in \boldsymbol{\Theta}$,*
$$S(\Theta; \mathtt{T}_a(x) = \Theta_i) + S\big(\Theta; \mathtt{T}_{a'}(x) = \overline{\Theta}_i\big) \le 1. \tag{3.3}$$
The equality attains if Θ is a singleton.

4 Severity in binary classification.

A classifier $\mathtt{T}$ aims to discern between the following two assertions:
$$\Theta_- : \text{patient } x \text{ is healthy}, \qquad vs. \qquad \Theta_+ : \text{patient } x \text{ is ill}.$$
The severity of the ill/positive diagnosis is ($\mathcal{X} = \boldsymbol{\Theta} = \{-,+\}$ in this case):
$$S(\Theta_+; \mathtt{T}(x) = \Theta_+) = \inf_{x \in \Theta_-} P_x(X \in \Theta_-) = Pr(X \in \Theta_- \mid x \in \Theta_-) = \text{Specificity}(\mathtt{T}).$$
On the other hand, the severity of the healthy/negative diagnosis is
$$S(\Theta_-; \mathtt{T}(x) = \Theta_-) = \inf_{x \in \Theta_+} P_x(X \in \Theta_+) = Pr(X \in \Theta_+ \mid x \in \Theta_+) = \text{Sensitivity}(\mathtt{T}).$$

5 Severity in frequentist inference.

Consider a family of tests for the pair of null and alternative hypotheses Θ_0 and Θ_1, characterized by the collection of rejection regions $\{R(\alpha)\}$, $\alpha \in (0,1)$. The *attained* power function is
$$\beta(\theta; x) = P_\theta(X \in R^x) \tag{5.1}$$
where $R^x = \cap_{\alpha : x \in R(\alpha)} R(\alpha)$ is the *attained* rejection region, and $p_x = \sup_{\theta \in \Theta_0} \beta(\theta; x)$ is the p-value of the test.

***Corollary* 5.1 (severity and p-value).** *The inferential strategy $\mathcal{T}$, characterized by critical regions that are rejection regions of a family of hypothesis tests: $C_1(\alpha) = R(\alpha)$, satisfies that for every $\mathtt{T} \in \mathcal{T}$,*
$$S(\Theta_1; \mathtt{T}(x) = \Theta_1) = 1 - p_x,$$
Furthermore, if the attained power function $\beta(\theta; x)$ satisfies $\sup_{\theta \in \Theta_0} \beta(\theta; x) = \inf_{\theta \in \Theta_1} \beta(\theta; x)$ for $x \in \mathcal{X}$, then
$$S(\Theta_0; \mathtt{T}(x) = \Theta_0) = p_x.$$

***Corollary* 5.2 (severity and the attained power function).** *Suppose $\boldsymbol{\Theta} = \mathbb{R}$, and consider one-sided assertions of interest $\tilde{\Theta}_0 = (-\infty, \tilde{\theta}]$ versus $\tilde{\Theta}_1 = (\tilde{\theta}, \infty)$. If the attained power function $\beta(\theta; x)$ is continuous and monotone increasing in θ for every x, we have that*
$$S\big(\tilde{\Theta}_1; \mathtt{T} = \Theta_1\big) = 1 - \beta(\tilde{\theta}; x),$$
and
$$S\big(\tilde{\Theta}_0; \mathtt{T} = \Theta_0\big) = \beta(\tilde{\theta}; x).$$

***Remark* 5.1.** Under the assumption of Corollary 5.2, the attained power function $\beta(\theta; x)$ is a **significance function** [1] (also called the p-value function), which is also a **confidence distribution** [5].

6 Severity in Bayesian inference: an illustration.

Suppose the sampling distribution is $X \sim Bin(n,\theta)$. Consider

$$\Theta_0 : \theta \sim \delta_{(\theta=\frac{1}{2})} \qquad \text{versus} \qquad \Theta_1 : \theta \sim \text{Unif}(0,1).$$

Notice that both Θ_0 and Θ_1 are *singleton* sets. This is the setting known as the Lindley's paradox.

The marginal likelihood for the data under the two respective assertions are

$$P(X=x\,|\,\Theta_0) = \binom{n}{x}\left(\frac{1}{2}\right)^n \qquad \text{and} \qquad P(X=x\,|\,\Theta_1) = \int \binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta = \frac{1}{n+1}.$$

Under equal prior probabilities, $P(\Theta_0) = P(\Theta_1) = 0.5$, the Bayes factor

$$BF(x;n) = \frac{P(x\,|\,\Theta_0)}{P(x\,|\,\Theta_1)} = (n+1)\binom{n}{x}\left(\frac{1}{2}\right)^n. \tag{6.1}$$

The Bayesian inferential strategy. The Bayesian inferential strategy is to endorse Θ_1 if the attained Bayes factor $BF(x;n)$ falls below a certain pre-determined threshold. This creates critical regions of the form

$$C_1(a) = \left\{x \in [n] : \left|x - \frac{n}{2}\right| > a\right\} \tag{6.2}$$

for some $a \geq 0$, and $C_0(a) = \overline{C}_1(a)$. For illustration, suppose $n = 100$ and $x = 35$. The attained critical region for the inferential strategy is

$$C_1^{35} = \{x < 36 \text{ or } x > 64\},$$

and $C_0^{35} = \overline{C}_1^{35}$.

Case 1: A ten-fold Bayes factor. Adopt a specific inferential claim that a Bayes factor less than **0.1** is considered strong evidence towards Θ_1. This effectively sets $a = 14$ in (6.2). The attained Bayes factor $BF(x=35;n=100) = 8.72 \times 10^{-2}$, and the inferential claim results in the conclusion $\mathtt{T}_{14}(35) = \Theta_1$, with severity

$$S(\Theta_1; T_{14}(35) = \Theta_1) = \inf_{\theta \notin \Theta_1} P_\theta\left(X \notin C_1^{35}\right) = Pr\left(36 \leq X \leq 64\,|\,\theta = \frac{1}{2}\right) \doteq 99.65\%.$$

On the other hand, the severity for the unendorsed claim Θ_0 is

$$S(\Theta_0; T_{14}(35) = \Theta_1) = \inf_{\theta \notin \Theta_0} P_\theta\left(X \notin C_1^{35}\right) = \sum_{x=36}^{64} \int \binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta = \frac{29}{101}.$$

Case 2: A hundred-fold Bayes factor. Adopt a more stringent inferential claim, that only a Bayes factor less than **0.01** would be considered strong evidence towards Θ_1. This is equivalent to setting $a = 18$. The inferential claim concludes $\mathtt{T}_{18}(x=35) = \Theta_0$, with severity

$$S(\Theta_0; T_{18}(35) = \Theta_0) = \inf_{\theta \notin \Theta_0} P_\theta\left(X \notin C_0^{35}\right) = 1 - \sum_{x=36}^{64} \int \binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta = \frac{82}{101},$$

whereas the severity for the unendorsed claim Θ_1 is

$$S(\Theta_1; T_{18}(35) = \Theta_0) = \inf_{\theta \notin \Theta_1} P_\theta\left(X \notin C_0^{35}\right) = 1 - P\left(36 \leq X \leq 64\,|\,\theta = \frac{1}{2}\right) \doteq 0.35\%.$$

References

[1] D. Fraser. Statistical inference: Likelihood to significance. *Journal of the American Statistical Association*, 86(414):258–265, 1991.

[2] D. G. Mayo. *Statistical inference as severe testing*. Cambridge: Cambridge University Press, 2018.

[3] D. G. Mayo and A. Spanos. Severe testing as a basic concept in a neyman–pearson philosophy of induction. *The British Journal for the Philosophy of Science*, 57(2):323–357, 2006.

[4] M. J. Schervish. P values: what they are and what they are not. *The American Statistician*, 50(3):203–206, 1996.

[5] M.-g. Xie and K. Singh. Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81(1):3–39, 2013.