Measuring Severity in Statistical Inference

Ruobin Gong (*Rutgers University*)

PSA Baltimore - November 12, 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 one may accord to the claim depends on the data and the strength of the "$\to$" relation. **Severity** [3, 2] concerns the measurement of strength in its contrapositive:

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

2 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 a positive (ill) diagnosis is

$$W(\Theta_+; \mathtt{T}(x) = \Theta_+) = \inf_{x \in \Theta_-} P_x(X \in \Theta_-) = Pr(X \in \Theta_- \mid x \in \Theta_-) = \text{Specificity}(\mathtt{T}).$$

The severity of a negative (healthy) diagnosis is

$$W(\Theta_-; \mathtt{T}(x) = \Theta_-) = \inf_{x \in \Theta_+} P_x(X \in \Theta_+) = Pr(X \in \Theta_+ \mid x \in \Theta_+) = \text{Sensitivity}(\mathtt{T}).$$

3 Definitions.

***Definition* 3.1 (inferential procedure).** Let $x \in \mathcal{X}$ be the observable data, and $\{\Theta_0, \Theta_1\}$ be a partition of $\boldsymbol{\Theta}$. The function $\mathtt{T} : \mathcal{X} \to \{\Theta_0, \Theta_1\}$ is a (binary) *inferential procedure* based on x, where

$$\mathtt{T}(x) = \begin{cases} \Theta_1 & \text{if } x \in C_1, \\ \Theta_0 & \text{if } x \in C_0, \end{cases} \tag{3.1}$$

where the *critical regions* $\mathbf{C} = \{C_1, C_0\}$ partition $\mathcal{X}$.

***Definition* 3.2 (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 procedures

$$\mathcal{T} = \{\mathtt{T}_a : \mathcal{X} \to \{\Theta_0, \Theta_1\}, a \in \mathcal{A}\}$$

is an *inferential strategy*, where each $\mathtt{T}_a$ is an inferential procedure accompanied by $\mathbf{C}_a$.

***Definition* 3.3 (warrant and severity).** The *warrant* accorded to an assertion $\Theta \subseteq \boldsymbol{\Theta}$ by the data x through an inferential procedure $\mathtt{T} \in \mathcal{T}$, is

$$W(\Theta; \mathtt{T}(x) = \Theta_i) = \inf_{\theta \notin \Theta} P_\theta(X \notin C_i^x), \qquad i = 0, 1, \tag{3.2}$$

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

$$C_i^x = \cap_{a \in \mathcal{A} : \mathtt{T}_a \in \mathcal{T}, \mathtt{T}_a(x) = \Theta_i} C_i(a) \tag{3.3}$$

is the *attained critical region* for Θ_i by $\mathcal{T}$. Furthermore,

$$W(\Theta = \Theta_i; \mathtt{T}(x) = \Theta_i)$$

is termed the *severity* with which Θ_i is inferred by $\mathtt{T} \in \mathcal{T}$ based on x.

***Remark* 3.1.** *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* 3.2.** $X \sim P_\theta$:

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

4 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* 4.1 (coherence).** *For fixed $x \in \mathcal{X}$ and a pair of assertions $\Theta, \tilde{\Theta} \subseteq \mathbf{\Theta}$ satisfying $\Theta \subseteq \tilde{\Theta}$,*

$$W(\Theta; \mathtt{T}(x) = \Theta_i) \leq W\big(\tilde{\Theta}; \mathtt{T}(x) = \Theta_i\big) \tag{4.1}$$

for $i \in \{0,1\}$.

***Definition* 4.1 (data informativeness).** Let $x, x' \in \mathcal{X}$, and $\mathcal{T}$ be an inferential strategy whose procedures 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 $i = 0,1$.

In words, x is more informative about a conclusion Θ_i if more inferential procedures from the same strategy $\mathcal{T}$ conclude Θ_i with x.

***Theorem* 4.2 (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 $\Theta \in \mathbf{\Theta}$,*

$$W(\Theta; \mathtt{T}(x) = \Theta_i) \geq W\big(\Theta; \mathtt{T}(x') = \Theta_i\big).$$

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

***Theorem* 4.3 (strategic equivalence).** *For fixed data $x \in \mathcal{X}$, let $\mathtt{T}_a$ and $\mathtt{T}_{a'}$ be two inferential procedures 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 $\Theta \in \mathbf{\Theta}$,*

$$W(\Theta; \mathtt{T}_a(x) = \Theta_i) = W(\Theta; \mathtt{T}_{a'}(x) = \Theta_i). \tag{4.2}$$

***Theorem* 4.4 (subadditivity).** *For fixed data $x \in \mathcal{X}$, let $\mathtt{T}_a$ and $\mathtt{T}_{a'}$ be two inferential procedures 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 $\Theta \in \mathbf{\Theta}$,*

$$W(\Theta; \mathtt{T}_a(x) = \Theta_i) + W\big(\Theta; \mathtt{T}_{a'}(x) = \overline{\Theta}_i\big) \leq 1. \tag{4.3}$$

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 $\mathtt{T} \in \mathcal{T}$,*

$$W(\Theta_1; \mathtt{T}(x) = \Theta_1) = 1 - p_x,$$

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

$$W(\Theta_0; \mathtt{T}(x) = \Theta_0) = p_x.$$

***Corollary* 5.2 (severity and the attained power function).** *Suppose $\mathbf{\Theta} = \mathbb{R}$, 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*

$$W\big(\tilde{\Theta}_1; \mathtt{T} = \Theta_1\big) = 1 - \beta(\tilde{\theta}; x), \qquad \text{and} \qquad W\big(\tilde{\Theta}_0; \mathtt{T} = \Theta_0\big) = \beta(\tilde{\theta}; x).$$

***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 that underlies the Lindley's paradox. The marginal likelihood under the two assertions are

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

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

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

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

$$C_1(a) = \left\{ x \in [n] : \left| x - \frac{n}{2} \right| > a \right\}, \quad C_0(a) = \overline{C}_1(a), \quad a \geq 0. \tag{6.2}$$

For illustration, suppose $n = 100$ and $x = 34$, with a Bayes factor $BF(34;100) = 4.63 \times 10^{-2}$. The attained critical regions for the inferential strategy are

$$C_1^{34} = \{x \leq 34 \text{ or } x \geq 66\} \qquad \text{and} \qquad C_0^{34} = \{34 \leq x \leq 66\}.$$

Case 1: A ten-fold Bayes factor. Adopt a specific inferential procedure that a Bayes factor less than **0.1** is considered strong evidence towards Θ_1. This effectively sets $a = 14$ in (6.2). The observation $x = 34 \in C_1(14)$, and the inferential procedure concludes $\mathtt{T}_{14}(x=34) = \Theta_1$. The severity accorded to the inference is

$$W(\Theta_1; T_{14}(34) = \Theta_1) = \inf_{\theta \notin \Theta_1} P_\theta\left(X \notin C_1^{34}\right) = Pr\left(35 \leq X \leq 65 \mid \theta = \frac{1}{2}\right) \doteq 99.89\%,$$

whereas the warrant accorded to the unendorsed claim Θ_0 is

$$W(\Theta_0; T_{14}(34) = \Theta_1) = \inf_{\theta \notin \Theta_0} P_\theta\left(X \notin C_1^{34}\right) = \sum_{x=35}^{65} \int \binom{n}{x} \theta^x (1-\theta)^{n-x} d\theta = \frac{31}{101}.$$

Case 2: A hundred-fold Bayes factor. Adopt a more stringent inferential procedure, that only a Bayes factor less than **0.01** would be considered strong evidence towards Θ_1. This is equivalent to setting $a = 18$ in (6.2). The observation $x = 34 \in C_0(18)$, and the inferential procedure concludes $\mathtt{T}_{18}(x=34) = \Theta_0$, with severity

$$W(\Theta_0; T_{18}(34) = \Theta_0) = \inf_{\theta \notin \Theta_0} P_\theta\left(X \notin C_0^{34}\right) = 1 - \sum_{x=34}^{66} \int \binom{n}{x} \theta^x (1-\theta)^{n-x} d\theta = \frac{68}{101},$$

whereas the warrant accorded to the unendorsed claim Θ_1 is

$$W(\Theta_1; T_{18}(34) = \Theta_0) = \inf_{\theta \notin \Theta_1} P_\theta\left(X \notin C_0^{34}\right) = Pr\left(X \leq 33 \text{ or } X \geq 67 \mid \theta = \frac{1}{2}\right) = 0.087\%.$$

Select 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.