1 The Likelihood Principle

Likelihood principle concerns foundations of statistical inference and it is often invoked in arguments about correct statistical reasoning.

Let $f(x|\theta)$ be a conditional distribution for X given the unknown parameter θ. For the observed data, $X = x$, the function $\ell(\theta) = f(x|\theta)$, considered as a function of θ, is called the *likelihood function.*

The name likelihood implies that, given x, the value of θ is more likely to be the true parameter than θ' if $f(x|\theta) > f(x|\theta')$.

> **Likelihood Principle.** In the inference about θ, after x is observed, all relevant experimental information is contained in the likelihood function for the observed x. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other.

Remark. The maximum-likelihood estimation does satisfy the likelihood principle.

Figure 1: Leonard Jimmie Savage; Born: November 20, 1917, Detroit, Michigan; Died: November 1, 1971, New Haven, Connecticut

The following example quoted by Lindley and Phillips (1976) is an argument of Leonard Savage discussed at Purdue Symposium 1962. It shows that the inference can critically depend on the likelihood principle.

Example 1: Testing fairness. Suppose we are interested in testing θ, the unknown probability of heads for possibly biased coin. Suppose,

$$H_0 : \theta = 1/2 \quad v.s. \quad H_1 : \theta > 1/2.$$

An experiment is conducted and 9 heads and 3 tails are observed. This information is not sufficient to fully specify the model $f(x|\theta)$. A rashomonian analysis follows:

• **Scenario 1:** Number of flips, $n = 12$ is predetermined. Then number of heads X is binomial $\mathcal{B}(n, \theta)$, with probability mass function

$$P_\theta(X = x) = f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{12}{9}\theta^9(1-\theta)^3 = 220 \cdot \theta^9(1-\theta)^3.$$

For a frequentist, the p-value of the test is

$$P(X \geq 9|H_0) = \sum_{x=9}^{12}\binom{12}{x}(1/2)^x(1-1/2)^{12-x} = \frac{1+12+66+220}{2^{12}} = 0.073,$$

and if you recall the classical testing, the H_0 is **not rejected** at level $\alpha = 0.05$.

• **Scenario 2:** Number of tails (successes) $\alpha = 3$ is predetermined, i.e, the flipping is continued until 3 tails are observed. Then, X - number of heads (failures) until 3 tails appear is Negative Binomial[1] $\mathcal{NB}(3, 1-\theta)$,

$$f(x|\theta) = \binom{\alpha+x-1}{\alpha-1}(1-\theta)^\alpha[1-(1-\theta)]^x = \binom{3+9-1}{3-1}(1-\theta)^3\theta^9 = 55 \cdot \theta^9(1-\theta)^3.$$

For a frequentist, large values of X are critical and the p-value of the test is

$$P(X \geq 9|H_0) = \sum_{x=9}^{\infty}\binom{3+x-1}{2}(1/2)^x(1/2)^3 = 0.0327.$$

since $\sum_{x=k}^{\infty}\binom{2+x}{2}\frac{1}{2^x} = \frac{8+5k+k^2}{2^k}$.

The hypothesis H_0 is **rejected,** and this change in decision not caused by observations.

According to Likelihood Principle, all relevant information is in the likelihood $\ell(\theta) \propto \theta^9(1-\theta)^3$, and Bayesians could not agree more!

Edwards, Lindman, and Savage (1963, 193) note: The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience.

2 Sufficiency

Sufficiency principle is noncontroversial, and frequentists and Bayesians are in agreement. If the inference involving the family of distributions and parameter of interest allows for a sufficient statistic then the sufficient statistic should be used. This agreement is non-philosophical, it is rather a consequence of mathematics (measure theoretic considerations).

[1]Let p be the probability of success in a trial. The number of failures in a sequence of trials until rth success is observed is Negative Binomial $\mathcal{NB}(r, p)$ with probability mass function

$$P(X = x) = \binom{r+x-1}{r-1}p^r(1-p)^x, \; x = 0, 1, 2, \ldots$$

For $r = 1$ the Negative Binomial distribution becomes the Geometric distribution, $\mathcal{NB}(1, p) \equiv \mathcal{G}(p)$.

Suppose that a distribution of random variable X depends on the unknown parameter θ. A statistics $T(X)$ is *sufficient* if the conditional distribution of X given $T(X) = t$ is free of θ.

The Fisher-Neyman factorization lemma states that the likelihood can be represented as

$$\ell(\theta) = f(x|\theta) = f(x)g(T(x), \theta),$$

Example. Let $X_1, \ldots, X_n$ is a sample from uniform $\mathcal{U}(0, \theta)$ distribution with the density $f(x|\theta) = \frac{1}{\theta}\mathbf{1}(0 \leq x \leq \theta)$. Then

$$\ell(\theta) = \prod_{i=1}^{n} f(X_i|\theta) = \frac{1}{\theta^n}\mathbf{1}(0 \leq \min_i\{X_i\})\mathbf{1}(\max_i\{X_i\} \leq \theta)$$

The statistics $T = \max_i\{X_i\}$ is sufficient. Here, $f(x) = \mathbf{1}(0 \leq \min_i\{x_i\})$ and $g(T, \theta) = \frac{1}{\theta^n}\mathbf{1}(\min_i\{x_i\} \leq \theta)$.

If the likelihood principle is adopted, all inference about θ should depend on sufficient statistics since $\ell(\theta) \propto g(T(x), \theta)$.

> **Sufficiency Principle.** Let the two different observations x and y have the same values $T(x) = T(y)$, of a statistics sufficient for family $f(\cdot|\theta)$. Then the inferences about θ based on x and y should be the same.

3 Conditionality Perspective

Conditional perspective concerns reporting data specific measures of accuracy. In contrast to the frequentist approach, performance of statistical procedures are judged looking at the observed data. The difference in approach is illustrated in the following example.

Example 2. Consider estimating θ in the model

$$P_\theta(X = \theta - 1) = P_\theta(X = \theta + 1),\ \theta \in \mathbb{R},$$

on basis of two observations, X_1 and X_2.

The procedure suggested is

$$\delta(X) = \begin{cases} \frac{X_1+X_2}{2}, & \text{if } X_1 \neq X_2 \\ X_1 - 1, & \text{if } X_1 = X_2 \end{cases}$$

To a frequentist, this procedure has confidence of 75% for all θ, i.e., $P(\delta(X) = \theta) = 0.75$.

The conditionalist would report the confidence of 100% if observed data in hand are different (easy to check!) or 50% if the observations coincide. Does it make sense to report the preexperimental accuracy which is known to be misleading after observing the data?

> **Conditionality Principle.** If an experiment concerning the inference about θ is chosen from a collection of possible experiments, independently of θ, then any experiment not chosen is irrelevant to the inference.

Example: [From Berger (1985), a variant of Cox (1958) example.] Suppose that a substance to be analyzed is to be sent to either one of two labs, one in California or one in New York. Two labs seem equally equipped and qualified and a coin is flipped to decide which one will be chosen. The coin comes up tails, denoting that California lab is to be chosen. After the results are returned back and report is to be written, should report take into account the fact that coin did not land up heads and that New York laboratory could have been chosen. Common sense and conditional view point say NO, but the frequentist approach calls for averaging over all possible data, even the possible New York data.

The conditionality principle makes clear the implication of the likelihood principle that any inference should depend only on the outcome observed and not on any other outcome we might have observed and thus sharply contrasts with the method of likelihood inference from the Neyman-Pearson, or more generally from a frequentist, approach. In particular, questions of unbiasedness, minimum variance and risk, consistency, the whole apparatus of confidence intervals, significance levels, and power of tests, etc., violate the conditionality principle.

Example 1 (continued): Testing fairness. Here is yet another scenario that will not impress a conditionalist:

• **Scenario 3:** Another coin (not the coin for which θ is to be tested) with probability of heads equal to ξ (known) was flipped. If heads were up, an experiment as in the **Scenario 1** was performed and if tail was up, the **Scenario 2** was used. The number of heads in experiment was 9 and the number of tails observed was 3.

Can you design ξ so that p-value of the test matches exactly 5%?

However, even the conditionalist agrees that the following scenario yields different evidence about θ than the Scenarios 1-3. The selection of the experiment depends on the parameters θ which is in violation of the conditionality principle.

• **Scenario 4:** The coin for which θ is to be tested was pre-flipped to determine what kind of experiment is to be performed. If the coin was heads up, an experiment as in the **Scenario 1** was performed and if it was tails up, the **Scenario 2** was used. The number of heads in the subsequent experiment was 9 and the number of tails observed was 3, and the initial flip to specify the experiment was not counted.

> **Birnbaum (1962)** Sufficiency Principle + Conditionality Principle $\equiv$ Likelihood Principle

Berger (1985), Berger and Wolpert (1988), Robert (2001) have additional discussion and provide more examples.

4 Some Sins of Being non-Bayesian

> **Thou shalt not** integrate with respect to sample space. A perfectly valid hypothesis can be rejected because the test failed to account for unlikely data that had not been observed...

The Lindley Paradox.[2]

[2] Lindley, D. V. (1957). A Statistical Paradox, *Biometrika*, **44**, 187–192.

Suppose $\bar{y}|\theta \sim N(\theta, 1/n)$. We wish to test $H_0 : \theta = 0$ vs the two sided alternative. Suppose a Bayesian puts the prior $P(\theta = 0) = P(\theta \neq 0) = 1/2$, and in the case of alternative, the 1/2 is uniformly spread over the interval $[-M/2, M/2]$.

Suppose $n = 40000$ and $\bar{y} = 0.01$ are observed, so $\sqrt{n}\,\bar{y} = 2$.

• Classical statistician rejects H_0 at level $\alpha = 0.05$.

• We will show that posterior odds in favor of H_0 are 11 if $M = 1$, so Bayesian statistician strongly favors H_0.

Ten in a row. Consider testing θ - the unknown probability of getting a correct answer. $H_0 : \theta = 1/2$ corresponds to guessing. The alternative $H_1 : \theta > 1/2$ corresponds to a "possession of knowledge, ability, or ESP."

• A lady who adds milk to her tea, claims to be able to tell whether the tea or the milk was poured to the cup first. In all 10 trials conducted to test this she is correct in determining what was poured first.

• A music expert claims to be able to distinguish a page of Haydn score from a page of Mozart score. In all 10 trials conducted to test this he makes correct determination each time.

• A drunken friend claims to be able to predict result of a fair coin flip. In all 10 trials conducted to test this he is correct each time.

In all 3 situations the one-sided p-value is 2^{-10} and the hypothesis $H_0 : \theta = 1/2$ is rejected. We will return to this bullet later with a Bayesian alternative.

Probability of heads. A coin is biased and the probability of heads θ is of interest. The coin is flipped 4 times and 4 tails are obtained. The frequentist estimate is $\hat{\theta} = 0$. More sensible Bayes solution will be proposed later.

More sins to follow:).

5 Exercises

1. Let $X_1, \ldots, X_n$ be Bernoulli $\mathcal{B}er(p)$ sample. Show that $T(X) = \sum_i X_i$ is sufficient by demonstrating that $X|T = t$ is discrete uniform on space of all n-tuples $x = (x_1, \ldots, x_n)$, $x_i = 0$ or 1, such that $\sum_i x_i = t$ with probabilities $\binom{n}{t}^{-1}$, and thus independent of p.

2. Let $X_1, \ldots, X_n$ be a sample from the exponential distribution $\mathcal{E}(\theta)$. Show that $\sum X_i$ is sufficient by demonstrating that the conditional distribution of $X_1, \ldots, X_n$ given $\sum X_i = t$ is free of θ. Use the fact that $\sum X_i$ is Gamma.

3. The Two-Envelope Paradox. Here is "paradox" coming from a subtle use of conditional reasoning at the place where unconditional reasoning is needed.

> I am presented with two envelopes A and B, and told that one contains twice as much money as the other. I am given envelope A, and offered the options of keeping envelope A or switching to B. What should I do? I reason: (i) For any x, if I knew that A contained x, then the odds are even that B contains either $2x$ or $x/2$, so the expected amount in B would be $5x/4$. So (ii) for all x, if I knew that A contained x, I would have an expected gain in switching to B. So (iii) I should switch to B.

Discuss. Discuss the following scenarios as well:

Scenario 1: After selecting the envelope A, I am told that one of the envelopes A or B contains 20\$. Should I switch?

Scenario 2: After selecting the envelope A, I am told that the envelope A contains 20\$. Should I switch?

Scenario 3: After selecting the envelope A, I am told that the envelope B contains 20\$. Should I switch?

References

[1] G.A. Barnard, G.A., G.M. Jenkins, G.M., and C.B. Winsten, C.B. (1962). Likelihood Inference and Time Series, *J. Royal Statistical Society, Series A*, **125**, 321–372.

[2] Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, Second Edition, Springer.

[3] Berger, J. O. and Wolpert, R. L. (1984), The Likelihood Principle. Hayward, CA: Institute of Mathematical Statistics.

[4] Birnbaum, A. (1962). On the Foundations of Statistical Inference. *Journal of the American Statistical Association*, **57**, 269–306.

[5] Cox, D. R. (1958). Some problems connected with the statistical inferece. *Ann. Math. Statistics,***29**, 357–372.

[6] Edwards A.W.F. (1974). The history of likelihood. *Int. Statist. Rev.* **42**, 9–15.

[7] Edwards, A.W.F. *Likelihood.* 1st edition 1972 (Cambridge University Press), 2nd edition 1992 (Johns Hopkins University Press).

[8] Edwards, W., Lindman, H., and Savage, L. J. (1963), Bayesian Statistical Inference for Psychological Research, *Psychological Review*, **70**, 193–242.

[9] Lindley, D. V. and Phillips, L. D. (1976). Inference for a Bernoulli process (a Bayesian view). *Amer. Statist.***30,** 112-119.

[10] Robert, C. (2001)*Bayesian Choice*, Second Edition, Springer Verlag.