Statistical Inference

Frequentist v.s. Bayesian

Junxiang Zheng

ADI Research

March 16, 2022

Overview

Background: Frequentist

- Frequentist statistics is also known as classical statistics, and orthodox statistics, which are what we learn in undergraduate classes.
- Famous frequentist terms: "Law of Large Numbers", "Central Limit Theorem", "sampling distribution", "standard error", "MLE", "bootstrap", "sufficient statistics", "UMVUE", "UMP", "Type I error", "Type II error", "confidence interval",...
- Famous frequentist applications: experimental design, data analysis, FDA, business intelligence,...

Background: Bayesian

- Bayesian statistics is useful in the situation of machine learning, but it is excluded from many statistic books or classes.
- Famous Bayesian terms: "Bayesian theorem", "prior distribution", "posterior distribution", "evidence", "variational inference", "MC sampling", "MAP", "Bayes risk", "graphic models", "naive Bayes classifier",...
- Famous Bayesian applications: reinforcement learning, AutoML, topic model, decision making,...

Interpretation of Probability: Frequentist

- The probability of a random event denotes the relative frequency of occurrence of an experiment's outcome when the experiment is repeated infinitely [wikipedia].

Weak Law of Large Numbers

Let X_1, X_2,...,X_n be a sequence of independent random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\bar{X}_n \xrightarrow{P} \mu$.

- Some understandings:
  - Measurement value = **fixed parameter** + **random error**
  - It depends on repeated trials

Interpretation of Probability: Bayesian

- Probability is treated as a degree of belief. The degree of belief has been interpreted as "the price at which you would buy or sell a bet that pays 1 unit of utility if E, 0 if not E" [wikipedia].

Bayes Theorem

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

prior: $P(\theta)$, likelihood: $P(X|\theta)$, evidence: $P(X)$, posterior: $P(\theta|X)$

- Some understandings:
  - Do not need repeated trials to interpret probability
  - It depends on the choice of prior probability

Statistical Inference: Frequentist

- Regard $x_1, \ldots, x_n$ as a realization of random variables $X_1, \ldots, X_n$, and assign $X_1, \ldots, X_n$ a joint distribution: joint cdf $F(\cdot|\theta)$, joint pdf $f(\cdot|\theta)$, joint pmf $P(\cdot|\theta)$, where $\theta = (\theta_1, \ldots, \theta_k)$ are **fixed constants**, but their values are **unknown**.
- Core: sampling distribution of statistics.
- Ways to compute sampling distribution of statistics:
  - analytic: need to know relationship between difference distributions.
  - numeric: bootstrap.
  - asymptotic: e.g. asymptotic normal.

Statistical Inference: Bayesian

- Regard $x_1, \ldots, x_n$ as a realization of random variables $X_1, \ldots, X_n$, and assign $X_1, \ldots, X_n$, and θ a joint distribution: joint cdf $F(\cdot, \theta)$, joint pdf $f(\cdot, \theta)$, joint pmf $P(\cdot, \theta)$, where $\theta = (\theta_1, \ldots, \theta_k)$ are **unobserved random variables**.
- Core: posterior distribution $P(\theta|X)$.
- Ways to compute posterior distribution:
    - exact inference
    - approximate inference:
        - deterministic: variational inference
        - stochastic: MCMC

Point Estimation: MLE

Examples (Normal MLEs, μ and σ unknown)

Let $X_1, X_2, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, with both μ and σ^2 unkown. Then

$$L(\mu, \sigma^2|X) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\Sigma_{i=1}^{n}(X_i-\mu)^2/\sigma^2}$$

and

$$\log L(\mu, \sigma^2|X) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\Sigma_{i=1}^{n}(X_i - \mu)^2/\sigma^2$$

Setting partial derivatives with respect to μ and σ^2 equal to 0,

$$\frac{\partial}{\partial\mu}\log L(\mu, \sigma^2|X) = \frac{1}{\sigma^2}\Sigma_{i=1}^{n}X_i - \theta = 0 \Rightarrow \hat{\mu} = \bar{X} = \frac{1}{n}\Sigma_{i=1}^{n}X_i$$

Point Estimation: MLE

Examples (Normal MLE, μ and σ unknown)

and

$$\frac{\partial}{\partial \sigma^2} \log L(\mu, \sigma^2 | X) = \frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \Sigma_{i=1}^{n} (X_i - \mu)^2 = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \Sigma_{i=1}^{n} (X_i - \bar{X}).$$

According to Example 7.2.12 of [Casella, 2001], we also need to check the following calculus conditions: 1) at least one second-order partial derivative is negative; 2) the Jacobian of the second-order partial derivatives is positive.

Properties of MLE

- **Invariance property of MLE**: If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$.
- **Consistency of MLE**: Under some regularity conditions, the MLE from an iid sample is consistent: $\lim_{n\to\infty} P_\theta(\left|\tau(\hat{\theta}), \tau(\theta)\right| \leq \epsilon) = 0$.
- **Efficiency of MLE**: Under some regularity conditions, $\sqrt{n}(\tau(\hat{\theta}) - \tau(\theta)) \to N(0, \nu(\theta))$, where $\nu(\theta)$ is the Cramér-Rao Lower Bound.

Bayes Estimator

Example (Normal Bayes estimator, μ is unknown)

There are n observations of $X \sim N(\mu, \sigma^2)$, where μ is unknown and σ^2 is known. $\mu \sim N(\mu_0, \sigma_0^2)$, where μ_0 and σ_0^2 is known. Let $\xi = 1/\sigma^2$, $\xi_0 = 1/\sigma_0^2$, and $\hat{\mu}$ is the estimate of MLE, the posterior distribution of μ is $N(\mu_1, \sigma_1^2)$, where

$$\mu_1 = \frac{\xi_0}{\xi_0 + n\xi}\mu_0 + \frac{n\xi}{\xi_0 + n\xi}\hat{\mu}$$

and

$$\xi_1 = n\xi + \xi_0, \sigma_1^2 = \frac{1}{\xi_1}.$$

Properties of Bayes Estimator

- $\xi = 1/\sigma^2$ is called precision. Note that $\xi_1 = n\xi + \xi_0$, which means $\sigma_1 < \sigma_0$.
- posterior mean is weighted average of prior mean and sample mean, where weights proportional to their precisions.
- **Connection with frequentist**: If n is large, then $\xi_1 \approx n\xi$, and $\mu_1 \approx \hat{\mu}$.
- **Noninformative prior**: If the prior distribution is "flat", e.g. constant, then the prior has little influence on the posterior distribution.

Hypothesis Testing

Definition of hypothesis testing

There are two complementary hypotheses: null hypothesis H_0: $\theta \in \Theta_0$ and alternative hypothesis H_1: $\theta \in \Theta_A$. A hypothesis testing is a rule that specifies that:

- For which sample values the decision is made to accept H_0 as true.
- For which sample values H_0 is rejected and H_1 is accepted.

Definition of likelihood ratio test

The likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_A$ is

$$\lambda(X) = \frac{\underset{\Theta_0}{sup}\, L(\theta|X)}{\underset{\Theta}{sup}\, L(\theta|X)}.$$

A likelihood ratio test (LRT) is any test that has a rejection region of the form: $\{x : \lambda(x) \leq c\}$, where c is any number satisfying $0 \leq c \leq 1$.

Likelihood Ratio Test

Example (Normal LRT)

Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, \sigma^2)$ population. Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. So the LRT statistic is

$$\lambda(X) = \frac{\underset{\Theta_0}{sup}\, L(\theta|X)}{\underset{\Theta}{sup}\, L(\theta|X)} = \frac{\underset{\Theta_0}{sup}\, L(\theta|X)}{L(\theta = \bar{X}|X)}$$

Then

$$\lambda(X) = \begin{cases} 1, & if\, \bar{X} \leq \theta_0 \\ \frac{L(\theta_0|X)}{L(\theta = \bar{X}|X)}, & if\, \bar{X} > \theta_0 \end{cases}$$

Likelihood Ratio Test

Example (Normal LRT)

If $\bar{X} > \theta_0$

$$\begin{aligned}
\lambda(X) &= \frac{L(\theta_0|X)}{L(\theta = \bar{X}|X)} \\
&= \frac{(2\pi)^{-n/2}\sigma^n exp[-\Sigma_{i=1}^n (X_i - \theta_0)^2/(2\sigma^2)]}{(2\pi)^{-n/2}\sigma^n exp[-\Sigma_{i=1}^n (X_i - \bar{X})^2/(2\sigma^2)]} \\
&= exp[(-\Sigma_{i=1}^n (X_i - \theta_0)^2 + \Sigma_{i=1}^n (X_i - \bar{X})^2)/(2\sigma^2)] \\
&= exp[-n(\bar{X} - \theta_0)^2/(2\sigma^2)]
\end{aligned}$$

The rejection region, $\{X : \lambda(X) \leq c\}$, can be written as $\{X : \bar{X} \geq \theta_0 + \sigma\sqrt{-2(\log c)/n}\}$, where $c \in (0, 1)$.

How to Evaluate Tests

Definition of power function

The power function of a hypothesis test with rejection region R is the function of θ defined by $\beta(\theta) = P_\theta(X \in R)$

Example (Normal power function)

From last example, an LRT of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ is a test that rejects H_0 if $(\bar{X} - \theta_0)/(\sigma/\sqrt{n}) > c$ (please ignore the difference with $\sqrt{-2(\log c)}$). The power function of this test is

$$\beta(\theta) = P_\theta\left(\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c\right) = P_\theta\left(\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)$$
$$= P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)$$

How to Evaluate Tests

Definition of two types errors

If $\theta \in \Theta_0$ but the test incorrectly reject H_0, then the test has made a Type *I* Error.
If $\theta \in \Theta_A$ but the test incorrectly reject H_A, then the test has made a Type *II* Error.

Example (Continuation of normal power function)

Suppose the experimenter wishes to have a maximum Type *I* Error probability of 0.1 (level $\alpha = 0.1$ test), and have a maximum Type *II* Error probability of 0.2 if $\theta \geq \theta_0 + \sigma$. We now show how to choose c and n to achieve these goals using the test in last example. Because $\beta(\theta)$ is increasing in θ, the requirements will be met if

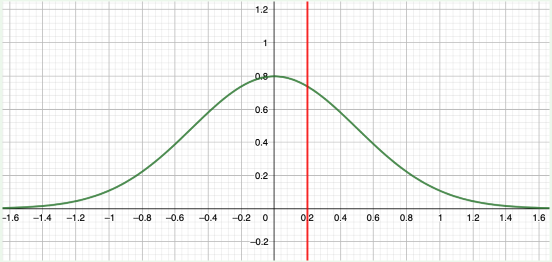
$$\beta(\theta_0) = 0.1 \text{ and } \beta(\theta_0 + \sigma) = 1 - 0.2 = 0.8.$$

By choosing $c = 1.28$ we achieve $\beta(\theta_0) = P(Z > 1.28) = 0.1$, regardless of n. Then $\beta(\theta_0 + \sigma) = P(Z > 1.28 - \sqrt{n}) = 0.8$, because $P(Z > -0.84) = 0.8$, so $n = \lceil 4.49 \rceil = 5$.

Bayesian test

Example (Normal Bayesian Test)

Using Bayesian approach to solve last example, we just need to compare two posterior probabilities $P(\theta \in \Theta_0|X)$ and $P(\theta \in \Theta_A|X)$.
If we decide to accept H_0 if and only if $P(\theta \in \Theta_0|X) \geq P(\theta \in \Theta_A|X)$, then we will accept H_0 if and only if $\frac{1}{2} \leq P(\theta \in \Theta_0|X) = P(\theta \leq \theta_0|X)$, which means if and only if posterior mean $\hat{\theta}_{\text{Bayes}} \leq \theta_0$.



Interval Estimation

Definition of interval estimation

An interval estimate of a real-valued parameter θ is any pair of functions, $L(x_1, \ldots, x_n)$ and $U(x_1, \ldots, x_n)$, of a sample that satisfy $L(x) \leq U(x)$ for all $x \in \mathcal{X}$. If $X = x$ is observed, the inference $L(x) \leq \theta \leq U(x)$ is made. The random interval $[L(X), U(X)]$ is called interval estimator.

Definition of coverage probability & confidence coefficient

The coverage probability of $[L(X), U(X)]$ is the probability that the random interval covers the true parameter θ. It is denoted by either $P_\theta(\theta \in [L(X), U(X)])$ or $P(\theta \in [L(X), U(X)] | \theta)$. The confidence coefficient is the infimum of the coverage probabilities, $\inf_\theta P_\theta(\theta \in [L(X), U(X)])$.

Interval Estimation

Example (Inverting a normal test)

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ and considering testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. For a fixed α level, a reasonable test (in fact, the UMPU test) has rejection region $\{x : |\bar{x} - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}\}$.
Since the test has size α, $P(H_0 \text{ is accepted}|\mu = \mu_0) = 1 - \alpha$. Then we can write

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}|\mu = \mu_0) = 1 - \alpha.$$

But this statement is true for every μ_0. Hence, the statement

$$P_\mu(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

is true.

Interval Estimation

Example (Inverting a normal test)

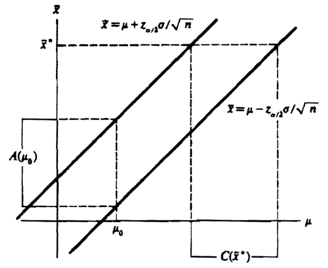


Figure: Relationship between confidence intervals and acceptance regions for tests.

Accept region: $A(\mu_0) = \{(x_1, \ldots, x_n) :$
$\mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\}$.

Confidence interval:
$C(x_1, \ldots, x_n) = \{\mu : \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\}$.

They are connected to each other by the tautology:
$(x_1, \ldots, x_n) \in A(\mu_0) \iff \mu_0 \in C(x_1, \ldots, x_n)$.

Bayesian Interval Estimation

Example (Credible interval of normal)

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, and let μ have the prior pdf $N(\mu_0, \sigma_0^2)$, where μ_0, σ, σ_0 are all known. From previous results, we know that posterior pdf of μ is $N(\mu_1, \sigma_1^2)$, where

$$\mu_1 = \frac{\xi_0}{\xi_0 + n\xi}\mu_0 + \frac{n\xi}{\xi_0 + n\xi}\hat{\mu}$$

and

$$\xi_1 = n\xi + \xi_0, \sigma_1^2 = \frac{1}{\xi_1}.$$

Therefore, a $(1-\alpha)$ credible set for μ is given by

$$\mu_1 - z_{\alpha/2}\sigma_1 \leq \mu \leq \mu_1 + z_{\alpha/2}\sigma_1.$$

Confidence Interval vs. Credible Interval

- **For confidence interval, we need to say that interval covers the parameter, not the parameter is inside the interval**. For example, a 90% confidence interval for θ is $[0.262, 1.184]$, the statement "the probability is 90% that θ is in $[0.262, 1.184]$" is invalid in frequentist statistics since the parameter is assumed fixed. $[0.262, 1.184]$ is a realized value of the random interval $L(X), U(X)$, θ is in the realized interval $[0.262, 1.184]$ with probability 0 or 1. When we say that the realized interval $[0.262, 1.184]$ has 90% chance of coverage, we mean that 90% sample points of the random interval cover the true parameter.
- In contrast, Bayesian allows us to say that θ is inside $[0.262, 1.184]$ with probability between 0 and 1. Note that the coverage probability of 90% credible interval not be 90%.

Decision Theory

Concepts of decision theory

- **sample space** $\mathcal{X}$: the set of all data values.
- **action space** A: the set of all possible actions.
- **decision rule** δ: a map from sample space to action space, i.e. $a = \delta(X)$.
- **parameter space** Θ: the set of all possible values of θ.
- **loss function** $loss(\theta, a)$: is a real function defined on $\Theta \times A$.
- **risk function** $R(\theta, \delta)$: The expected loss function of a decision d, formalized as $R(\theta, \delta) = E_X[loss(\theta, \delta(X))]$.

- We are going to find a best decision rule δ which has the minimal risk.
- There are many ad-hoc evaluation methods for estimation and testing. However, decision theory is a general theory which can unify estimation and testing.

Minimax Rule

Since the true value of θ is unknown, as a frequentist, we want to find a decision rule δ which has a small risk for all values of θ.

Definition of minimax rule

$$\delta_{MM} \triangleq \underset{\delta}{\operatorname{argmin}} \max_{\theta} R(\theta, \delta)$$

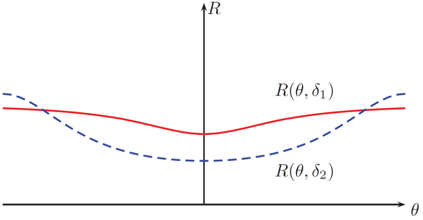


Figure 6.2 Risk functions for two decision procedures, δ_1 and δ_2. Since δ_1 has lower worst case risk, it is the minimax estimator, even though δ_2 has lower risk for most values of θ. Thus minimax estimators are overly conservative.

Bayes Rule

Definition of Bayes rule

When θ has a prior distribution $p(\theta)$, we define Bayes risk as

$$R_{Bayes} = \int_{\Theta} R(\theta, \delta) p(\theta) d\theta.$$

Bayes rule is the decision rule δ_{Bayes} which minimize the Bayes risk.

Definition of posterior expected loss

$$R_{Bayes} = \int_{\Theta} \left[\int_{\mathcal{X}} loss(\theta, \delta(x)) p(x|\theta) dx \right] p(\theta) d\theta \quad = \int_{\mathcal{X}} \left[\int_{\Theta} loss(\theta, \delta(x)) p(\theta|x) d\theta \right] p(x) dx,$$

where $\int_{\Theta} loss(\theta, \delta(x)) p(\theta|x) d\theta$ is called posterior expected loss.

Bayes Rule

- In fact, all minimax rules are equavilent to Bayes rules under a least favorable prior [Murphy, 2012].
- **Method to find Bayes rule**: The posterior expected loss is a function of x, and not a function of θ. Thus, if we choose the action $\delta(x)$ to minimize the posterior expected loss, we will minimize the Bayes risk.

Application of Decision Theory: Point Estimation

- When applied in point estimation, action space $A =$ parameter space Θ, a decision rule is an estimator.
- Three Bayes rules:
  - For squared error loss, the posterior expected loss is $E_{\theta|x}((\theta - a)^2|X = x)$, which is minimized by $\delta_{Bayes} = E(\theta|x)$. So the Bayes rule is the mean of the posterior distribution.
  - For absolute error loss, the posterior expected loss is $E_{\theta|x}(|\theta - a||X = x)$, which is minimized by choosing $\delta_{Bayes} =$ median of $p(\theta|x)$.
  - If θ only have two values to choose, for 0-1 error loss, the posterior expected loss is $p(a \neq \theta|x) = 1 - p(\theta|x)$, which is minimized by choosing $\delta_{Bayes} = \arg\max_{\theta \in \Theta} p(\theta|x)$.

Application of Decision Theory: Testing

- In hypothesis testing, there are only two actions allowed: "accept H_0" or "reject H_0". We denote them as a_0 and a_1 respectively. The set $\{x : \delta(x) = a_0\}$ is the acceptance region and the set $\{x : \delta(x) = a_1\}$ is the rejection region.
- Testing can be reduced to classification in Bayesian approaches.
- We can use simple 0-1 loss which is defined by

$$loss(\theta, a_0) = \begin{cases} 0, & if\, \theta \in \Theta_0 \\ 1, & if\, \theta \in \Theta_A \end{cases} \text{ and } loss(\theta, a_1) = \begin{cases} 1, & if\, \theta \in \Theta_0 \\ 0, & if\, \theta \in \Theta_A \end{cases}$$

- We can use generalized 0-1 less which is defined by

$$loss(\theta, a_0) = \begin{cases} 0, & if\, \theta \in \Theta_0 \\ c_{II}, & if\, \theta \in \Theta_A \end{cases} \text{ and } loss(\theta, a_1) = \begin{cases} c_I, & if\, \theta \in \Theta_0 \\ 0, & if\, \theta \in \Theta_A \end{cases}$$

Application of Decision Theory: Interval Estimation

- We use C to denote action which means $\theta \in C$.
- There are two quantities in loss function of interval estimation:
  - correctness: we use 0-1 loss.

$$loss(\theta, C) = \begin{cases} 0, & if \theta \in C \\ 1, & if \theta \notin C \end{cases}$$

  - length of C
- Therefore, one of such loss functions is:

$$loss(\theta, C) = b\text{Length}(C) - loss(\theta, C).$$

Why Isn't Everyone a Bayesian? [Efron, 1986]

- Fisher's theory (MLE) has automatic nature, and statistician does not have to think a lot about the specific situation in order to get on toward its solution.
- Bayesian theory concentrates on inference, but Fisher paid a lot of attention to the earlier steps of the data analysis.
- Bayesian solution cannot be applied in some situations. For example, bootstrap analysis (much like a cross-validation) can give an unbiased estimate of the decision trees. Bayesian cannot handle this problem.
- Bayesian are subjective, but strict objectivity is one of the crucial factors separating scientific thinking from wishful thinking. Objective Bayesian are difficult in some cases.

Comments to "Why Isn't Everyone a Bayesian? [Efron, 1986]"

- Freqentist is lack of theory and coherency.
- Decisions need to be made when data are lacking, e.g., acid rain, the safety of nuclear power.
- Agreement with frequentist theories may be interesting but is no justification.
- The objective element is the data, interpretation of data is subjective, as anyone who has interacted with scientists knows.

References

- https://en.wikipedia.org/wiki/Probability
- George Casella and George Berger. Statistical Inference, 2001.
- Kevin P. Murphy. Machine learning: A Probabilistic Perspective, 2012.
- Christopher M. Bishop. Pattern Recognition and Machine Learning, 2006.
- B. Efron. Why Isn't Everyone a Bayesian?, 1986.

The End