Report: Basic Concept of Statistical Inference

sitao min

Basic Concept

Statistical inference is the process of using data to infer the distribution that generated the data. The aim of statistical inference is to make certain determinations with regard to the unknown **parameters** figuring in the underlying distribution.

- **Population** is any large collection of objects or individuals, such as Americans, students, or trees about which information is desired.
- **Parameter** is any summary number, like an average or percentage, that describes the entire population. The **population mean** μ and the **population proportion** p are two different population parameters. The problem is that 99.9... % of the time, we cannot know the real value of a population parameter. The best we can do is estimate the parameter. This is where **samples** and **statistics** come in to play.
- **Sample** is a representative group drawn from the population.
- **Statistic** is any summary number, like an average or percentage, that describes the sample. The sample mean, $\bar{x}$, and the sample proportion $\hat{p}$ are two different **sample statistics**.

Statistical model $\mathfrak{F}$ is a set of distributions (or densities or regression functions), which has two types:

- **Parametric model** is a set $\mathfrak{F}$ that can be parameterized by a **finite** number of parameters. $\mathfrak{F} = \{f(x, \theta),\ \theta \in \Theta\}$
- **Nonparametric mode**l is a set F that **cannot** be parameterized by a **finite** number of parameters.

  > **Example**: Suppose we observe pairs of data$(X_1, Y_1), \dots (X_n, Y_n)$. Perhaps X_i is the blood pressure of subject i and Y_i is how long they live. **X** is called a **predictor** or regressor or **feature** or **independent variable**. **Y** is called the outcome or the **response variable** or the **dependent variable**. We call$r(x) = E(Y|X = x)$ the regression function. If we assume that $r \in F$ where F is **finite dimensional** —the set of straight lines for example — then we have a **parametric regression model**. If we assume that $r \in F$ where F is not finite dimensional then we have a **nonparametric regression model**. The goal of predicting Y for a new patient based on their X value is called **prediction**. If Y is **discrete** (for example, live or die) then prediction is instead called classification. If our goal is to estimate the function r, then we call this **regression** or curve estimation.

The **two dominant approaches** for statistical inference are called **frequentist inference** and **Bayesian inference**.

- **Frequentist inference** draws conclusions from sample data by emphasizing the **frequency or proportion** of the data. An alternative name is **frequentist statistics**. This is the inference framework in which the well-established methodologies of statistical **hypothesis testing** and **confidence intervals** are based.
- **Bayesian inference** is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayesian inference derives the posterior probability as a consequence of two antecedents, a prior probability and a "likelihood function" derived from a statistical model for the observed data. Bayesian inference computes the posterior probability according to Bayes' theorem:

  $P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$

Parameter Estimation

In parametric model, the main aim of statistical inference is to get the estimation of parameters for specific statistical distribution from practice data. Within the frame-work of parametric statistical inference, there are three types of parameter estimation: ***Point Estimation*** , ***Interval Estimation*** (**Confidence Interval**) , ***Testing Hypotheses***.

- **Point Estimation**

  Point estimation is to estimate the value of an unknown parameter.

  In practical situation, we first draw a random sample of size n, $X1, \ldots, Xn$, from the underlying distribution, and on the basis of it to construct a point estimate (or estimator) for parameter θ. The estimated parameter statistic is denoted as $\hat{\theta} = \hat{\theta}(X1, \ldots, Xn)$, where a statistic is a known function of the random sample $X1, \ldots, Xn$. The simple idea is to find a funcion $f(\theta, X_i)$ and find the optimal value of θ to minimize the expectation of the $f_i(\theta, X)$. So the basic problem of point estimation is to formed as following formula:

  $\hat{\theta} = argmin \frac{1}{n} \sum_{i=1}^{n} f(\theta; X_i)$

  The most popular method for choosing function f is using Likelyhoond function and this method is called **MLE** (***Maximum Likelyhood Estimation***). The likelyhood function is defined by $\mathcal{L}(\theta) = \Pi_{i=1}^{n} f(\hat{\theta}, X_i)$ and Log-Likelyhood function is defined by $l(\theta) = log\mathcal{L}(\theta)$ and MLE is denoted by $\hat{\theta}$ which is maximize the $\mathcal{L}(\theta)$. It can be proved that MLE converges in probability to the true value.

- **Confidence Interval**

  Confidence Interval is the interval that contains unknown parameter with high prescribed probability.

> **Example** We can be 95% confident that the proportion of Penn State students who have a tattoo is between 5.1% and 15.3%.

To be more precise and in casting the problem in a general setting, let $X1, \ldots, Xn$be a random sample from the $p.d.f.\ f(\cdot;\theta), \theta \in \Omega \subseteq \mathbb{R}$, and let $L = L(X1, \ldots, Xn)$ and $U = U(X1, \ldots, Xn)$ be two statistics of the $Xi's$ such that $L < U$. Then the interval with end-points L and U, $[L, U]$, is called a random interval. Let α be a small number in $(0, 1)$, such as $0.005, 0.01, 0.05$, and suppose that the random interval $[L, U]$ contains θ with probability equal to $1 - \alpha$ (such as $0.995, 0.99, 0.95$) no matter what the true value of θ in Ω is.
$P_\theta(L \leq \theta \leq U) = 1 - \alpha \ for\ all\ \theta \in \Omega.$

If this relation holds, then we say that the random interval $[L, U]$ is a confidence interval for θ with confidence coefficient $1 - \alpha$.

For example, suppose we want to estimate an actual population mean μ. As you know, we can only obtain $\bar{x}$, the mean of a sample randomly selected from the population of interest. We can use $\bar{x}$ to find a range of values:

$$Lower\ value\ (L) < population\ mean\ \mu < Upper\ value\ (U)$$

that we can be really confident contains the population mean μ. The range of values is called a "**confidence interval**." Once we've obtained the interval, we can claim that we are really confident that the value of the population parameter is somewhere between the value of **L** and the value of **U**.

> For example, a newspaper report (ABC News poll, May 16-20, 2001) was concerned whether or not U.S. adults thought using a hand-held cell phone while driving should be illegal. Of the 1,027 U.S. adults randomly selected for participation in the poll, 69% thought that it should be illegal. The reporter claimed that the poll's "**margin of error**" was 3%. Therefore, the confidence interval for the (unknown) population proportion *p* is 69% ± 3%. That is, we can be really confident that between 66% and 72% of all U.S. adults think using a hand-held cell phone while driving a car should be illegal.

(1-α) *t*-interval for the population mean μ

confidence interval = Sample mean ± (t-multiplier × standard error) $= \bar{x} \pm t_{\alpha/2, n-1}\left(\frac{s}{\sqrt{n}}\right)$

- the "**t-multiplier**," which we denote as $t_{\alpha/2, n-1}$, depends on the sample size through *n* - 1 (called the "**degrees of freedom**") and the confidence level $(1 - \alpha) \times 100$ through $\alpha/2$.
- the "**standard error**," quantifies how much the sample means $\bar{x}$ vary from sample to sample. That is, the standard error is just another name for the estimated standard deviation of all the possible sample means.
- the quantity to the right of the ± sign, *i.e.*, "**t-multiplier × standard error**," is just a more specific form of the margin of error. That is, the margin of error in estimating a population mean μ is calculated by multiplying the *t*-multiplier by the standard error of the sample mean.

- **Hypotheses Testing**

Hypotheses Testing is to test whehter unknown parameter lies in a specified subset of the parameter space.

> **Example** There is enough statistical evidence to conclude that the mean normal body temperature of adults is lower than 98.6 degrees F."

The general idea of hypothesis testing involves:

1. Making an initial assumption. $H_0 : Null\ Hypothesis\ \ H_A : alternative\ Hypothesis$

   In statistical, we usually **assume the null hypothesis is true**
2. Collecting evidence (data).
3. Based on the available evidence (data), deciding whether to reject or not reject the initial assumption.

We merely state that there is enough evidence to behave one way or the other. Because of this, whatever the decision, **there is always a chance that we made an error**.

Type I error: The null hypothesis is rejected when it is true.

Type II error: The null hypothesis is not rejected when it is false.

But, a good scientific study will minimize the chance of doing so! When **making a desicion,** It is either *likely* or *unlikely* that we would observe the evidence we did given our initial assumption. If it is *likely*, we do not reject the null hypothesis. If it is *unlikely*, then we reject the null hypothesis in favor of the alternative hypothesis.

Bootstrap Method

The bootstrap is a method for estimating standard errors and computing confidence intervals.

- **Bootstrap Varience**

  Let $T_n = g(X1, \ldots, Xn)$ be a statistic, that is, $T_n$is any function of the data. Suppose we want to know $V_F(T_n)$, the variance of T_n. The idea is to simlulate $V_{\hat{F}}(T_n)$ from sample data as real $V_F(T_n)$. Fisrt is to simulation distribution of T_n using data drawing from distribution $\hat{F}$ to compute each T_n. And then using distribution of T_n simulate $V_{\hat{F}}(T_n)$. The bootstrap methods steps is following:

  Step1: Draw $X_1^*, \ldots, X_n^* \sim Fn$.

  Step2: Compute $T_n^* = g(X_1^*, \ldots, X_n^*)$, in parametric model, using point estimation compute T_n

  Step3: Repeat step1 and 2 , B times to get $T_1^*, \ldots, T_n^*$

  Step4: Let

  $v_{boot} = \frac{1}{B}\sum_{b=1}^{B}(T_{n,b}^* - \sum_{r=1}^{B} T_{n,r}^*)$

- **Boostrap Cofidence Interval**

  We have already know the estimated variance of the statistic parameter θ. Using variance of this parameter, we can easily build the $1-\alpha$ confidence interval of this parameter, using following formula:

  $$Confidence\ Interval = [T_n - z_{\alpha/2}\hat{se}_{boot},\ T_n + z_{\alpha/2}\hat{se}_{boot}]$$

  se is boostrap variance

References

[1] Wikipedia

[2] ***All of statistics: A Consice Course in Statistical Inference***, *Larry Wasserman*

[3] ***An Introduction to Probability and Statistical Inference***, *George Roussas*

[4] Online Course of Statistics, penn state university