STAT 611-600

Theory of Statistics - Inference
Introduction, Exponential Family & Sufficient Stats

Tiandong Wang

Department of Statistics
Texas A&M University

Register for the online materials

- Logon to http://ecampus.tamu.edu/
- Choose the TAMU (Net ID) logon option
- Logon with your Net ID and password
- Choose the STAT 611 course list link
- Piazza sign-up: `piazza.com/tamu/spring2021/stat611600`.
  - **Note:** All course-related questions will be addressed via Piazza and students are strongly encouraged to answer each other's questions.

The goal of this couse

- STAT 610, introduction of probabilistic models and properties of random variables, common probability distributions
- STAT 611, continuation of mathematical theory of statistics (principles for statistical inference, formulation of statistical models, point estimation, confidence intervals, hypothesis testing, and Bayesian inference.)

Statistical Inference

- **Population**: The entire group of interest (unknown).
- **Sample**: A part of the population selected to draw conclusions about the entire population (observed).
- **Probability model**: Probability distribution for the **population** (usually specified by pdf or cdf.)
- **Parameter**: numerical characteristics associated with probability distribution.

Statistical Inference

Goal is to infer aspects of population from information in sample.

- Probability theory
  - Reasoning is from population to sample.
  - Population is known and we determine probability of obtaining a given sample from the population.
- Statistical theory
  - Reasoning is from sample to population.
  - Sample is known and we try to guess form of population from info in sample.

Statistical Inference

Data $\mathbf{X} = (X_1, \cdots, X_n)$: from a probability distribution $f(\mathbf{x}; \theta)$, with θ unknown.

- Our task is to estimate θ based on data
  Examples:
  - to estimate the success probability p in a Bernoulli trial
  - to estimate the supporting rate p of a president candidate
  - to estimate the average SAT score of the freshmen at a national level
- Three types of methods to estimate θ
  - point estimation (Chapter 7)
  - hypothesis testing (Chapter 8)
  - interval estimation (Chapter 9)

Review on common distributions: Discrete

The range (sample space) of X is countable.

- Discrete Uniform
- Hypergeometric
- Binomial
- Poisson
- Geometric
- Negative binomial

Continuous distributions

- Continuous Uniform
- Exponential distribution
- Gamma Family
- Normal Distribution
- Beta

More on distribution theories

- Sampling distributions related to normal: T, Chi-square, and F distributions.
- Exponential Family: Density functions for distributions belong to the exponential family satisfy

$$f(x;\theta) = c(\theta)h(x)\exp[\sum_{j=1}^{k} w_j(\theta)t_j(x)], \theta = (\theta_1, \cdots, \theta_d).$$

  The exponential family is **full-rank** if $d = k$; curved if $d < k$.
- Natural parametrization of the exponential family:

$$f(x;\eta) = h(x)c^*(\eta)\exp\left\{\sum_{j=1}^{k} \eta_j t_j(x)\right\}.$$

Examples

- Binomial; Poisson.
- Normal(μ, σ^2).
- Normal(μ, μ^2).
- Zero-Inflated Poisson.

Review Thm 3.4.2 in C&B for properties of the exponential family.

Statistics

(1) Summarize information about θ in data with one or a few statistics $T = T(\mathbf{X})$.
(2) Use T to construct point estimators, test statistics, upper/lower confidence limit.

Definition

A statistic is a real- or vector-valued function of the random sample:

$$T = T(X_1, \ldots, X_n)$$

- sample mean: $\bar{X} = \dfrac{1}{n} \sum X_i$
- sample variance: $S^2 = \dfrac{1}{n-1} \sum (X_i - \bar{X})^2$
- sample standard deviation: $S = \sqrt{S^2}$
- minimum of sample: $T = \min(X_1, \ldots, X_n)$
- part of sample: $T = (\sum X_i, \sum X_i^2)$

REMARK: A statistic can not be a function of parameter.

Which statistic?

- Data $X_1, \cdots, X_n$ contains much information, some are relevant for θ and some are not.
- Dropping irrelevant information is desirable, but dropping relevant information is undesirable.
- the dimension of T is generally smaller than the sample size n.

Partition of Sample Space by $T(\mathbf{X})$
Consider the discrete case. For any possible value t of T, there is a corresponding set

$$A_t = \{x : T(x) = t\}$$

The set collection $\{A_t, \text{all} \quad t\}$ makes a partition on the sample space of X.
Note

$$P(T(\mathbf{X}) = t) = \sum_{\mathbf{x} \in A_t} P(\mathbf{X} = x).$$

Statistics and Partition: Example

Example: Toss a coin $n = 3$ times, and let X_1, X_2, X_3 be respectively the outcome of each toss. Let T denote the total number of heads obtained, i.e., $T = \sum_{i=1}^{3} X_i$. Write down the partition of the sample space given by T.

Remark: Often T has a simpler data structure and distribution than the original sample $\mathbf{X} = (X_1, \cdots, X_n)$, so it would be nice if we can use $T(\mathbf{X})$ to summarize and then replace the entire data.

Important Issues in Data Reduction:

We should think about the following questions carefully before the simplification process

- Is there any loss of information due to summarization?
- How to compare the amount of information about θ in the original data $\mathbf{X}$ and in $T(\mathbf{X})$?
- Is it sufficient to consider only the reduced data T

Sufficient Statistics

Definition

A statistic T is called sufficient if the conditional distribution of $\mathbf{X}$ given T is free of θ (that is, the conditional is a completely known distribution).

- Remark 1:
  - A distribution *free* of θ means that the distribution is completely known, hence the corresponding random quantity can be generated with a random number generator.
  - Example. Toss a coin n times, and the probability of head is an unknown parameter p. Let $T =$ the total number of heads. Is T sufficient for p?

Example: Bernoulli Trials

Suppose that $X_1, \ldots, X_n$ are iid Bern(p) and $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$. Is $T(X_1, \ldots, X_n)$ sufficient?

Solution: Compute

$$\mathbb{P}\left(\mathbf{X} = \mathbf{x} \middle| \sum_{i=1}^n X_i = k\right) = \frac{1}{\binom{n}{k}}.$$

Note: Given $T(\mathbf{X})$, $\mathbf{X}$ is uniformly distributed on the set of all n-vectors of 0's and 1's with exactly $T(\mathbf{X})$ 1's.

Sufficient Statistics: Continue

Remark 2:

- Consider the discrete random variable case. Assume T is sufficient for θ. Given any value of T, we can define a conditional distribution of $\mathbf{X}$ given on $T(\mathbf{X}) = t$ (with the restricted sample space $\mathcal{A}_t$), and generate a pseudo data $\mathbf{X}'$ from $P(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x})). \quad (\star)$
- Note $(\star)$ is actually a probability distribution defined on the set $A_{T(\mathbf{x})}$. Both $\{\mathbf{X} = \mathbf{x}\}$ and $\{\mathbf{X}' = \mathbf{x}\}$ are subsets of $\{T(\mathbf{X}) = T(\mathbf{x})\}$. We can show that $\mathbf{X}$ and $\mathbf{X}'$ have the identical distribution, i.e.

$$P_\theta(\mathbf{X} = \mathbf{x}) = P_\theta(\mathbf{X}' = \mathbf{x}), \forall \mathbf{x}, \theta.$$

Sufficient Statistics: Remark 2 Continue

- Since $X_1', \cdots, X_n'$ can be regarded as another random sample from the same population as the original data $X_1, \cdots, X_n$, they contain an equal amount of probabilistic information about θ. Therefore, we can recover the data if we retain T and discard $X_1, \cdots, X_n$. That's why T is sufficient.
- For the continuous case, we have the same conclusion (proof omitted for technical reasons).

Sufficiency Principle

- If T is sufficient, the extra information carried by $\mathbf{X}$ is worthless as long as θ is concerned. It is then natural to only consider inference procedures which do not use extra irrelevant information.
- This leads to the Sufficiency Principle :
  Any inference procedure should depend on the data only through sufficient statistics.
- Question: How to check sufficiency?

Good News: Exponential family

Exponential Family: Recall the density function of an exponential family

$$f(x;\theta) = c(\theta)h(x)\exp[\sum_{j=1}^{k} w_j(\theta)t_j(x)], \theta = (\theta_1, \cdots, \theta_d).$$

Theorem

Let $X_1, \cdots, X_n$ be a random sample from the exponential family. Then

$$T(\boldsymbol{X}) = (\sum_{i=1}^{n} t_1(\boldsymbol{X}_i), \cdots, \sum_{i=1}^{n} t_k(\boldsymbol{X}_i))$$

is sufficient for $\theta = (\theta_1, \cdots, \theta_d)$.

- Applies to many standard families discussed above such as binomial, Poisson, normal, exponential, gamma.

Example

Show that the gamma distribution belongs to the exponential family and find the sufficient stats.