Statistical Inference

Aakash Ghosh

Contents

1 Introduction 3
1.1 Why statistical inference? 3
1.2 Problem of non-indetifiability 3
1.2.a Example 4
1.3 Parameter 4
1.3.a Examples 4
1.4 Regular families 5

2 Statistic 6
2.1 Introduction 6
2.2 Ancillary Statistics 6
2.2.a Example 1 6
2.2.b Example 2 7
2.2.c Example 3 7
2.3 Scale Family 8
2.4 Location Family 8
2.5 Location-Scale Family 8
2.6 Limiting conditions 9
2.7 Characterization in terms of level sets 9
2.8 Sufficient Statistic 10
2.8.a Example 10
2.9 Fisher-Neyman factorization 10
2.9.a Example 1 11
2.9.b Example 2 11
2.9.c Example 3 11
2.9.d Example 4 11
2.10 Minimal sufficient Statistic 12
2.11 Example 12
2.11.a Example 1 12
2.11.b Example 2 12
2.12 Complete statistics 13
2.13 Misc. Theorems 13
2.14 Exponential Families 13
2.15 Examples 13

2.15.a Example 1 13
2.15.b Example 2 14

3 Assignment-1 15

Chapter 1

Introduction

1.1 Why statistical inference?

There are three main questions we would like to answer:

1. Given a random vector X, determine which $\theta \in \Theta$ is most compatible with the data given [i.e. which $\theta \in \Theta$ generates X]. This is point estimation.
2. Given an X we wish to determine if a given value of θ or distribution f_θ is consistent. This is hypothesis testing.
3. After (1), we would like to find a suitable rang of values of θ which is compatible. This is interval estimation.

We would also like our answers to be optimal in each of the above cases.

1.2 Problem of non-indetifiability

This is a problem which arises when more than one value of $\theta \in \Theta$ can give rise to the same f_θ.

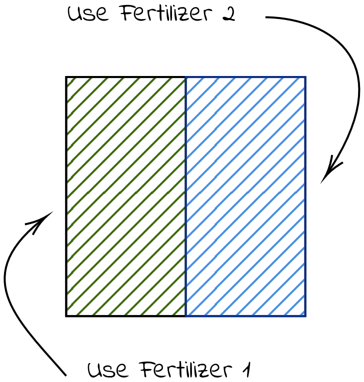
$$g : \Theta \to \mathcal{F}$$

g is onto, but not one-one

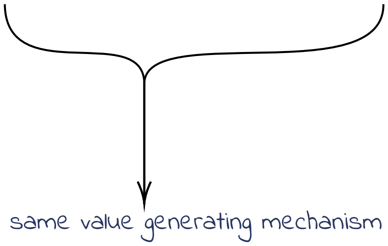
$\exists \theta_1, \theta_2$ such that $g(\theta_1) = g(\theta_2)$. The difficulty arising from this is very obvious: with non-identifiability, θ_1 can be confused with θ_2 as we are essentially claiming that both θ_1 and θ_2 results in the same data generating mechanism. We must understand that this is an inherent problem of our data generating mechanism/experiment and must be addressed before any further analysis is carried out.

1.2.a Example

Consider a scenario where we want to test the effectiveness of two fertilizers on a field. What we might do is divide the field in two divisions such, use fertilizer one on the first half and fertilizer two on the second half and then record our crop yield. Assume that the field has yield follows a normal distribution with mean α with standard deviation 1. Further, assume that each fertilizer i increases yield by α_i. We repeat this process several$(= n)$ times.
Assume on the j^{th} iteration with fertilizer i the yield is X_{ij}. Define $X_1 = [X_{11}, X_{12}, X_{13} \ldots X_{1n}]$ and $X_2 = [X_{21}, X_{22}, X_{23} \ldots X_{2n}]$. Then by using our setup, we can use X_1 and X_2 to estimate $\alpha + \alpha_1$ and $\alpha + \alpha_2$ but not $\alpha, \alpha_1, \alpha_2$ individually.



$$\theta = (\alpha, \alpha_1, \alpha_2) \qquad \theta_c = (\alpha - c, \alpha_1 + c, \alpha_2 + c)$$



Again note this inability is an inherent flaw in the experiment itself, and we can't do anything to circumnavigate it. However, we might address this problem by assuming that $\alpha_1 + \alpha_2 = 0$. Note, this is definitely not true. But we are claiming something along the lines that the net fertility remains constant, and this helps in computing all the values.

1.3 Parameter

A parameter is a function

$$\nu : \mathcal{F}_\theta \to \mathcal{N}$$

where $\mathcal{N}$ is an arbitrary set. When F_θ is an identifiable parameterization, then there exists function q such that:

$$\nu(F_\theta) = q(\theta)$$

Identifiability ensures that we can recover ν from q. Going from ν to q requires no conditions on indetifiability.

1.3.a Examples

1. $\mu(\theta) = \int x dF_\theta(x)$
2. $\sigma^2(\theta) = \int x^2 dF_\theta(x) - (\mu(\theta)^2)$
3. $m(\theta) =$ median of F_θ

1.4 Regular families

For our work we will assume that:

1. All $F_\theta \in \mathcal{F}$ have continuous CDF with density f_θ or are discrete with pmf f_θ
2. There is identifiable parameterization.

Chapter 2

Statistic

2.1 Introduction

Any measurable function $T : X \to \mathbb{R}^k$ is called a statistic.
The idea is to compress the given data without losing valuable information. We look at two extreme cases:

- $T(X)$ =constant: This tells us nothing and all information is lost,
- $T(X) = X$: while this retains all information, there is no compression that takes place.

2.2 Ancillary Statistics

We wish to find good statistics. And it turns out that to achieve this it is better to sieve out the bad ones and then look further in whatever is left*.

An ancillary statistic $T = T(X)$ is such that the distribution of $T(X)$ is independent of θ.

2.2.a Example 1

Assume $X_1, X_2 \dots X_n \sim (\mu, 1), \mu \in \mathbb{R}$. Let

$$T(X) = \sum_{i=0}^{n} \sum_{j=0}^{n} (X_i - X_j)^2$$

Two approaches for solving this are:

*Honestly, this feels like something you find on cleanup: You do everything and then understand that doing it this way is smoother.

1. Define $\tilde{X}_i = X_i - \bar{X}$. Note that:

$$\begin{aligned} T(X) &= \sum_{i \neq j} (X_i - X_j) \\ &= \sum_{i \neq j} (\tilde{X}_i - \tilde{X}_j) \\ &= (n-1) \sum_i {\tilde{X}_i}^2 - 2 \sum_{i \neq j} \tilde{X}_i \tilde{X}_j \\ &= (n+1) \sum_i {\tilde{X}_i}^2 - 2 \sum_{i,j} \tilde{X}_i \tilde{X}_j \\ &= (n+1) \sum_i {\tilde{X}_i}^2 - 2 \left(\sum_i \tilde{X}_i \right) \left(\sum_j \tilde{X}_j \right) \\ &= (n+1) \sum_i \left(X_i - \bar{X} \right)^2 \end{aligned}$$

   Now distribution of $X_i - \bar{X} \sim \mathcal{N}(0,1)$ which is independent of μ. Therefore, T is ancillary.

2. The above argument can be considerable shortened by arguing that the distribution of $X_i - X_j$ is independent of μ, and therefore the whole sum is independent of μ.

2.2.b Example 2

Assume $X_1, X_2 \dots X_n \sim Exp(\theta)$. The density function is given by: $f_\theta(x) = \theta e^{-\theta x}$. Such a distribution often occurs when modelling things like lifetime of a particular make of a bulb, etc. Define:

$$T(X) = \frac{\max_i X_i}{\min_i X_i}$$

Finding the distribution of T explicitly is a nightmare. Instead, what we do is define $Y_i = \theta X_i$. Note that:

$$P(Y_i < y) = P(\theta X_i < y) = \int_0^{y/\theta} \theta e^{-\theta x} dx = \int_0^y e^{-u} du$$

Therefore, $Y_i \sim Exp(1)$. we rewrite T as:

$$T(X) = \frac{\max_i X_i}{\min_i X_i} = \frac{\max_i Y_i}{\min_i Y_i} = T(Y)$$

But, it is trivial to note that any parameter on Y is independent of θ. Therefore, $T(X)$ is ancillary.

2.2.c Example 3

Assume $X_1, X_2 \dots X_n \sim Unif(0, \theta)$. Define:

$$T(X) = \frac{\max_i X_i}{\min_i X_i}$$

Define $Y_i = X_i/\theta$. Note that:

$$P(Y_i < y) = P(X_i/\theta < y) = \int_0^{y\theta} \frac{1}{\theta} dx = \int_0^y du$$

Therefore, $y \sim Unif(0,1)$. we rewrite T as:

$$T(X) = \frac{\max_i X_i}{\min_i X_i} = \frac{\max_i Y_i}{\min_i Y_i} = T(Y)$$

Again as any parameter on Y is independent of θ. Therefore, $T(X)$ is ancillary.
The idea in the examples above is that we are "scaling" the function to fit our needs.

2.3 Scale Family

A scale family is a family $\mathcal{F}$ of distribution given by:

$$\left\{ f_\theta | f_\theta(x) = \frac{1}{\theta} g\left(\frac{x}{\theta}\right) \right\}$$

where g is a known distribution. Set $Y_i = X_i/\theta$. Any $T(X)$ which can be written as $S(Y)$ is ancillary.

2.4 Location Family

A scale family is a family $\mathcal{F}$ of distribution given by:

$$\{ f_\theta | f_\theta(x) = g(x - \theta) \}$$

where g is a known distribution. Set $Y = X_i - \theta$. Any $T(X)$ which can be written as $S(y)$ is ancillary.

2.5 Location-Scale Family

A scale family is a family $\mathcal{F}$ of distribution given by:

$$\left\{ f_\theta | f_\theta(x) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right) \right\}$$

where g is a known distribution. Set $Y_i = (X_i - \mu)/\sigma$. Any $T(X)$ which can be written as $S(Y)$ is ancillary.

2.6 Limiting conditions

other ways of identifying a "bad-statisttic" is to look at limiting conditions. For example, let $X_i \sim Unif(0, \theta)$. Consider two candidates:

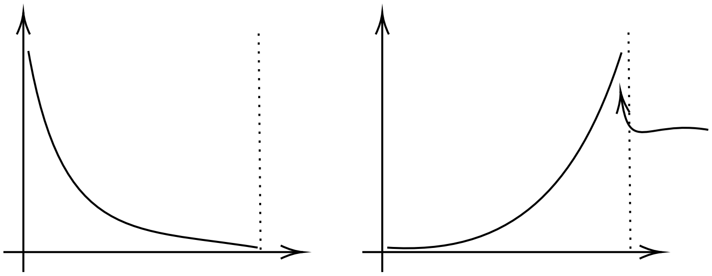
$$T_1(X) = \max_i X_i$$

$$T_2(X) = \min_i X_i$$

It is easy to compute that:

$$f_{T_2}(t) = \frac{n}{\theta}\left(1 - \frac{t}{\theta}\right)^{n-1}, 0 < t < \theta$$

$$f_{T_1}(t) = \frac{n}{\theta}\left(\frac{t}{\theta}\right)^{n-1}, 0 < t < \theta$$



f_{T_2} f_{T_1}

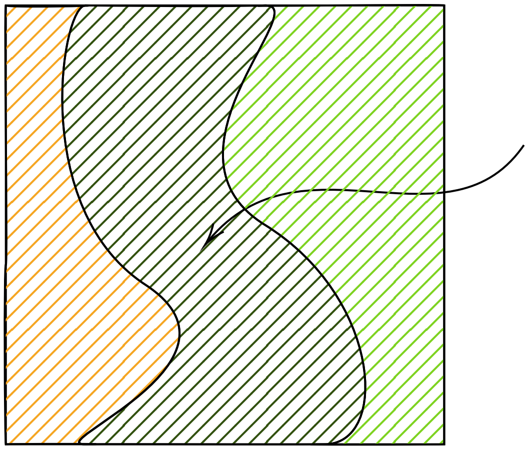
The density is concerntrated here, which gives more information about θ

2.7 Characterization in terms of level sets

Consider a level set of T

$$A_t = \{x \in \mathbb{R}^n | T(x) = t\}$$

Ideally what we want is T to do is compress all information about θ and leave out all the rest. This essentially translates to arguing that any variability of x in A is independent of $theta$, or more formally, conditional distribution of X given $T(X) = t$ is independent of θ. But that implies that for any t, T restricted to $T^{-1}(t)$ is ancillary in nature.



T is ancillary in a level set (each shaded region).
Any vaiability
is independent of θ

Such a T is called a sufficient statistic.

2.8 Sufficient Statistic

$T = T(x)$ is said to be sufficient for θ if $P(X = B|T = t)$ is independent of θ for all $B \in \mathcal{B}_{\mathbb{R}^n}, t \in \mathbb{R}^k$. Note that any one-one map of a sufficient statistic is also sufficient.

2.8.a Example

Let $X_1, X_2 \ldots X_i$ be Bernoulli iid with parameter θ, i.e.

$$P(X_i = 1) = \theta \qquad P(X_i = 0) = 1 - \theta$$

Take $T = \sum X_i$. we calculate $P(X = x_i|T = t)$.

$$P(X = x_i|T = t) = \frac{P(X = x, T(X) = t)}{P(T = t)}$$

the denominator easily comes out to be $\binom{n}{t}\theta^t(1-\theta)^{n-t}$. For the numerator, consider two cases:

1. $T(x) = t$: The event $[X = x]$ is a subset of $[T(x) = t]$. Therefore, the numerator simplifies to $P(X = x) = \theta^t(1-\theta)^{n-t}$, the whole expression simplifies to $\binom{n}{t}^{-1}$.
2. $T(x) \neq t$: The events $[X = x]$ and $[T = t]$ are disjoint. Therefore, the numerator simplifies to 0.

In either case, we see that the value of the expression is independent of θ.

2.9 Fisher-Neyman factorization

Calculating conditional probabilities is hard for continuous distributions. Finding candidate T for checking is even harder. A nice characterization for overcoming these difficulties is Fisher-Neyman factorization. $T(X)$ s sufficient for X if and only if $f_\theta(x)$ can be written as

$$f_\theta(x) = g(T(x), \theta)h(x)$$

2.9.a Example 1

Let X_is be iid Bernoulli with parameter θ.

$$\begin{aligned} f_\theta(x) &= P_\theta(X = x) && \text{[Good practice to include } \theta \text{ to show dependence]} \\ &= \prod_{i=1}^{n} \theta^{x_1}(1-\theta)^{1-x_i} 1_{x_i \in \{0,1\}} && \text{[Make sure to characterize the domain]} \\ &= \underbrace{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}_{\substack{\text{Set } T(x) = \sum x_i \\ \text{This part becomes } g(T(x),\theta)}} \quad \underbrace{1_{x_i \in \{0,1\}}}_{\text{This becomes } h(x)} \end{aligned}$$

which is exactly what we had concluded before.

2.9.b Example 2

Let X_is be iid exponentials with parameter θ.

$$f_\theta(x) = \prod \theta e^{-\theta x_i} 1_{x_i > 0} = \theta^n e^{-\theta \sum x_i} \prod 1_{x_i > 0}$$

we have a similar decomposition as before with $T(x) = \sum x_i$

2.9.c Example 3

Let X_is be iid uniform with parameter θ.

$$\begin{aligned} f_\theta(x) &= \prod_i \theta^{-1} 1_{x_i \in (0,\theta)} && \text{[Note how the domain is parameter dependent]} \\ &= \theta^{-n} 1_{0 < \min x_i \leq \max x_i \leq 1} \\ &= \theta^{-n} 1_{0 < \min x_i, \max x_i \leq 1} \\ &= \underbrace{\theta^{-n} 1_{\max x_i < 1}}_{\substack{\text{Set } T(x) = \max x_i \\ \text{and get } g(T(x),\theta)}} \underbrace{1_{0 \leq \min x_i}}_{\substack{\text{This becomes} \\ h(x)}} \end{aligned}$$

2.9.d Example 4

Let X_is be iid Cauchy with parameter θ (pmf $\frac{1}{\pi}\frac{1}{(x-\theta)^2}$).

$$f_\theta(x) = \frac{1}{\pi^n} \prod \frac{1}{1 + (x_i - \theta)^2}$$

Note there is not much to exploit over here. The best(to be proven later) that we can do is to order them in increasing order. This does not provide a reduction in dimensionality but compresses some information as there are multiple combinations of x which results in the same order.

2.10 Minimal sufficient Statistic

T is said to be minimally sufficient for θ if

1. T is sufficient for θ.
2. If S is any other sufficient statistic for θ, then there exists h such that

$$T = h(S)$$

It trivially follows that two minimally sufficient statistics are related by a one-one function.

Theorem 1. Let X have the joint pmf/pdf $f_\theta(X)$ and let T be a statistic. Then T is minimally sufficient for θ if the following holds:

$$\frac{f_\theta(x)}{f_\theta(y)} \text{ is independent of } \theta \text{ iff } T(x) = T(y)$$

2.11 Example

2.11.a Example 1

If $X_1, X_2 \dots X_n \sim \mathcal{N}(0, \sigma^2), \sigma > 0$ then we have:

$$\frac{f_\theta(X)}{f_\theta(Y)} = e^{-\left(-\frac{1}{2\sigma^2}\left(\sum x_i^2 - \sum y_i^2\right)\right)}$$

It is easy to see that $T(X) = \sum x_i^2$ is minimally sufficient.

2.11.b Example 2

If $X_1, X_2 \dots X_n \sim Unif(\theta, \theta+1), \sigma > 0$ then we have:

$$\frac{f_\theta(x)}{f_\theta(y)} = \frac{1_{x_{(n)}-1<\theta<x_{(1)}}}{1_{y_{(n)}-1<\theta<y_{(1)}}} \qquad [X_{(i)} \text{ is the } i^{th} \text{ order statistic}]$$

Define $c_x = (x_{(n)} - 1, x_{(1)}), c_y = (y_{(n)} - 1, y_{(1)})$. we get different values of the ratios for each of $\theta \in c_x \cap c_y$,$\theta \in c_x^c \cap c_y$, $\theta \in c_x \cap c_y^c$ and $\theta \in c_x^c \cap c_y^c$. Set $T(X) = \{X_{(n)}, X_{(1)}\}$. It is trivially obvious that if $T(X) = T(Y)$ then $c_x = c_y$ and the ratio becomes 1[θ is kinda restricted in c_x or c_y],

2.12 Complete statistics

we say a statistic T is complete if for any h

$$E_\theta\left[h(T)\right] = 0 \Rightarrow P_\theta(h(T) = 0) = 1 \forall \theta \in \Theta$$

2.13 Misc. Theorems

Theorem 2. Debobroto Basu's theorem- Any boundedly complete minimal sufficient statistic is independent of any ancillary statistic.

Theorem 3. A complete sufficient statistic is a minimal statistic.

The converse is not true.

2.14 Exponential Families

A family of distribution $\mathcal{F}\theta$ is called an exponential family with k parameters if the joint distribution can be written as:

$$f_\theta = \exp\left\{C(\theta) \cdot T(X) - d(\theta)\right\} S(x)$$

and

$$\text{Support}(x) = \{x \in \mathbb{R}^n | f_\theta(x) \neq 0\} \text{is independent of } \theta$$

where:

- The vector $C(\theta)$ is independent of X
- The vector $T(X)$ is independent of θ
- C, T are vectors in $\mathbb{R}^k$
- Parameterization by $C(\theta)$ is identifiable. This parameterization is called the natural/canonical parameterization. A bit of calculation implies that this requires $T(X)$ to be full rank[not restricted to some hyperplane]. Similarly, we want $C(\theta)$ to be in the lowest form: i.e. no linear dependence among terms.

2.15 Examples

2.15.a Example 1

Let $X_1, X_2 \ldots X_n \sim Unif(0, \theta)$. X is not an exponential family as the support is dependent of θ.

2.15.b Example 2

Let $X_1, X_2 \dots X_n \in \mathcal{N}(\theta, \theta^2), \theta > 0$* Then $C(\theta) = (\theta^2, \theta)$. Now note, while the parameterization takes two argument, those two arguments are dependent on each other: knowledge of θ is enough. Therefore, while the parameters are 2 dimensional, the parameter space is 1 dimensional. Cases where the parameter space is of the form $(\theta, \gamma(\theta))$ where γ is a curve is known as curved exponential family.

* I know it and present me knows that future me knows it, it is not worthwhile to look through tedious calculations. So will just put the result here.

Chapter 3

Assignment-1

Problem 1

Let $X_1, \ldots, X_n$ be a sample from the uniform distribution on $(0, \theta)(\theta > 0)$. Show that $S = X_{(n)}$ is a complete sufficient statistic for θ. Show that $T = X_{(1)}/X_{(n)}$ is ancillary for θ. Use Basu's theorem to show that S and T are independent.

Solution:

1. we start with finding the max order statistics. Note
$$P\left(X_{(n)} < k\right) = \prod P(x_i < k)$$
we set $Y = X_{(n)}$ and get:
$$f_\theta(y) = \frac{ny^{n-1}}{\theta^n} 1_{0<y<\theta}$$
Now for any $h(Y)$, if $E[h(Y)] = 0$ then $h(Y) = 0$ a.e in $(0, \theta)$. Therefore, Y is sufficent.
2. T is ancillary as $unif(0, \theta)$ is scale family. Set $y_i = x_i/\theta$. Then $y_i \sim Unif(0, 1)$ and $Y_{(i)} = X_{(i)}$. So $\frac{X_{(1)}}{X_(n)} = \frac{Y_{(1)}}{Y_(n)}$ which is independent of θ.
3. -

Problem 2

Let $X_1, \ldots, X_n$ be a sample from the uniform distribution on $(-\theta, \theta)(\theta > 0)$. Show that $S = \left(X_{(1)}, X_{(n)}\right)$ is a sufficient statistic for θ and it is not complete. Show that $\max\left(-X_{(1)}, X_{(n)}\right)$ is minimal sufficient. Is it complete?

1. Ez to see that S is sufficient.