Statistical Inference I **Prof. Chin-Tsang Chiang**

Lecture Notes 3

September 21, 2015 **Scribe: Chi-Ning Chou**

Today we're going to talk about some properties of probability function and some useful theorems such as Boole's inequality, Bonferroni's inequality, the first Borel-Cantelli lemma. In the next lecture, we will talk about Counting Theory, and today Prof.Chiang also demonstrate some intuitions about counting theory.

1 Basic Properties of Probability Function and Their Intuitions

Theorem 1 (properties of probability function) *Suppose $(\Omega, \mathcal{A}, P)$ is a probability space where P satisfies the axioms of probability function. Then,*

1. *(minimizer) $P(\emptyset) = 0$*
2. *(finite partition) $P(A) = 1 - P(A^C), \forall A \in \mathcal{A}$*
3. *(upper bound) $P(A) \leq 1, \forall A \in \mathcal{A}$*
4. *(conditioned on finite partition) $P(B \backslash A) = P(B \cap A^C) = P(B) - P(B \cap A)$*
5. *(union) $P(A \cup B) = P(A) + P(B) - P(A \cup B)$*
6. *(sub-event) $P(A) \leq P(B), \forall A, B \in \mathcal{A}, A \subseteq B$*
7. *(conditioned on countably partition) $P(B) = \sum_{i=1}^{\infty} P(B \cap A_i)$, where $\{A_i\}$ is a partition of Ω and $A_i \in \mathcal{A}, \forall i$.*
8. *(Boole's inequality) $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$. Note that here $\{A_i\}$ is arbitrary sequence of events.*

The following gives some intuitions behind each result:

1. To prove empty set is a minimizer, we can simply choose a sequence of event as $A_1 = \Omega, A_i = \emptyset, \forall i = 2, 3, ...$ Then the results is obvious as we apply countably mutual exclusive additive axiom.
2. By choosing $A_1 = A, A_2 = A^C, A_i = \emptyset, \forall i = 3, 4, ...$, countably mutual exclusive additive axiom shows the results.
3. With 2. and the lower bound axiom.
4. Consider $B = B \cap \Omega = B \cap (A \cup A^C)$, by distributive law, $B = (B \cap A) \cup (B \cap A^C)$ where the two terms are mutually exclusive. Thus, we an apply the mutually exclusive additive axiom.

5. By think of $A \cup B$ as $A \cup (B \cap A^C)$ such that the two terms are mutually exclusive. Then consider $B = (B \cap A) \cup (B \cap A^C)$ where the two terms are also mutually exclusive. Apply the mutually exclusive additive axiom, we have $P(A \cup B) = P(A) + P(B \cap A^C) = P(A) + (P(B) - P(B \cap A)) = P(A) + P(B) - P(A \cap B)$.

6. Simply consider $B = A \cup (B \cap A^C)$.

7. The same as 4. by consider $B = \bigcup_{i=1}^{\infty}(B \cap A_i)$ where all the terms are mutually exclusive.

8. Set $A_1^* = A_1, A_2^* = A_2 \backslash A_1, A_3^* = A_3 \backslash (A_1 \cup A_2), ..., A_j^* = A_j \backslash (\bigcup_{i=1}^{j-1} A_i)$ such that $\{A_j^*\}$ is mutually exclusive and $\bigcup_{i=1}^{\infty} A_i = \bigcup_{j=1}^{\infty} A_j^*$, which means that $P(\bigcup_{i=1}^{\infty} A_i) = P(\bigcup_{j=1}^{\infty} A_j^*)$. With an observation on $P(A_j^*) = P(A_j \backslash (\bigcup_{i=1}^{j-1} A_i)) \leq P(A_j)$, we have $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$.

Intuition (properties of probability function)

The key point here is to make the sets become **mutually exclusive** so that we can apply the mutually exclusive additive axiom. On the contrary, it also shows that the mutually exclusive additive axiom is so strong that what we think is reasonable in intuition can be described by it!

2 Useful Inequalities, and the First Borel-Cantelli Lemma

In the previous section, Boole's inequality provides an loose upper bound for the probability of the **union** of a sequence of events. Now, we can draw a similar lower bound for the probability of the **intersection** of a sequence of events.

2.1 Bonferroni's inequality

Theorem 2 (Bonferroni's inequality)
Let $A_1, A_2, ...$ be a sequence of events, then

$$P(\bigcap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$$

Proof:

$$P(\bigcap_{i=1}^{n} A_i) = 1 - P(\bigcup_{i=1}^{n} A_i^C) \geq 1 - \sum_{i=1}^{n} P(A_i^C) = 1 - \sum_{i=1}^{n}(1 - P(A_i)) = \sum_{i=1}^{n} P(A_i) - (n-1)$$

■

Just as Boole's inequality, Bonferroni's inequality provides a loose lower bound. What does Bonferroni's inequality brings to us? There's an application in **pairwise comparison**: Suppose we are going to estimate three parameters $\theta_1, \theta_2, \theta_3$ which are highly correlated. How can we lower bound the probability of the event: $(\theta_1 - \theta_2) \in A_{12}, (\theta_1 - \theta_3) \in A_{13}, (\theta_2 - \theta_3) \in A_{23}$? Note that this probability is used for constructing confidence interval.

Now, consider that we have already lower bound the significant level of each parameter, say

$$P(\theta_1 - \theta_2 \in A_{12}) \geq 1 - \alpha_1$$
$$P(\theta_1 - \theta_3 \in A_{13}) \geq 1 - \alpha_2$$
$$P(\theta_2 - \theta_3 \in A_{23}) \geq 1 - \alpha_3$$

By Bonferroni's inequality, we have

$$\begin{aligned} & P((\theta_1 - \theta_2) \in A_{12}, (\theta_1 - \theta_3) \in A_{13}, (\theta_2 - \theta_3) \in A_{23}) \\ \geq\ & P(\theta_1 - \theta_2 \in A_{12}) + P(\theta_1 - \theta_3 \in A_{13}) + P(\theta_2 - \theta_3 \in A_{23}) - 2 \\ \geq\ & (1 - \alpha_1) + (1 - \alpha_2) + (1 - \alpha_3) - 2 \\ =\ & 1 - (\alpha_1 + \alpha_2 + \alpha_3) \end{aligned}$$

Intuition (Boole's and Bonferroni's inequalities)

The important concept here is that both Boole's and Bonferroni's inequalities are relatively **loose**. As a result, they can only provide good estimate as the probability of the individual events are **small**. Take a look at the above example, we can see that normally the significant level α_i is relatively small. And that's why here Bonferroni's inequality has some usage.

2.2 The first Borel-Cantelli lemma

Before we give the statement of the first Borel-Cantelli lemma, let's consider the following motivation: Imagine after defining the axiom of probability and deduce some nice properties and now you're going to define the concept of **conditional probability** $P(A|B)$. However, an issue coming out:

What if P(B)=0?

This makes sense as now we consider the sample space with countably many outcomes. The probability of a single outcome is measure 0 ($P(\text{outcome}) = 0$). As a result, the conditional probability taught in elementary probability $P(A|B) = \frac{P(A \cap B)}{P(B)}$ is not well-defined!

A simple idea is to relax B or construct a sequence of event $B_1, B_2 ...$ such that $\lim_{n\to\infty} B_n = B$ and define the conditional probability as $P(A|B) := \lim_{n\to\infty} \frac{P(A \cap B_n)}{P(B_n)}$, which is much more making sense. But here comes a question:

Is the probability of a convergent sequence of events also converge?

Theorem 3 (The first Borel-Cantelli lemma)

Let $\{A_i\}$ be a sequence of events such that $\sum_{i=1}^{\infty} P(A_i) < \infty$, then

$$P(A_{n.i.o}) = 0$$

, where $A_{n.i.o} = \limsup_{n\to\infty} A_n$

Proof: By the definition of limsup, we have $\bigcup_{i=n}^{\infty} A_i \to \limsup_{n\to\infty} A_n$ monotonously. Now, we proof the lemma with two steps:

1. (Continuity axiom) $P(A_{n.i.o}) = \lim_{n\to\infty} P(\bigcup_{i=n}^{\infty} A_i)$
2. (Boole's inequality) $P(\bigcup_{i=n}^{\infty} A_i) \leq \sum_{i=n}^{\infty} P(A_i)$

Since $\sum_{i=1}^{\infty} P(A_i) < \infty$, we now that $\lim_{n\to\infty} \sum_{i=n}^{\infty} P(A_i) = 0$. Thus, $P(A_{n.i.o}) = 0$. ■

Note that there's no independence involves here, which is different from the second Borel-Cantelli's lemma that we'll introduce later.

3 Intuitions About Counting Theory

Why do we need counting techniques?

3.1 Sampling

What kind of sampling mechanisms do we have? That is, how do we design the mechanism for us to draw samples from a large population? Do we sample periodically? Or, sample from the subset that we believe is more important? Or, we cluster the sample space into a smaller subspace and sample in them respectively?

The chance mechanism we choose will affect the underlying possibility for us to get a certain outcome. That is, the way we analyze or the model we choose will differ as the sampling techniques are distinct. And how do we react to different mechanism is to utilize the tools in counting theory.

3.2 Bootstrap

In high-dimensional statistics, the number of parameters are so large that sometimes even the number of samples is less than that of parameters! However, we believe the underlying structure might not be so much. Namely, there might exist some sparse or low rank structure behind the scheme.

But, how can we find out such implicit properties with only a few samples and a great amount of candidate parameters? There's a method called **bootstrap**, which reuses the same samples and generate e a subset of new samples from it. With some counting argument, we can see that this will not affect the performance and can increase the number of samples to search for special structure in a high-dimensional setting.