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

Lecture Notes 4

September 27, 2015 **Scribe: Wei-Chang Lee, Chi-Ning Chou**

Today we're going to talk about method of counting to be able to study probability assignments on finite space, we will then give definition about conditional probability on discrete case and intuition about continuous case. After knowing how to calculate conditional probabilities, we can then use Bayes' theorem to connect experimental study to observation study which is called the odds ratio. At the end of the class, Prof.Chiang introduce the concept "independence" and how to interpret it beyond geometric views in a more statistical way.

1 Counting Methods

There are two kinds of interpretation of probability measure.

1 The first kind of view is based on the "frequency of occurrence". They do random experiments many times to study how many times event which is interested in would take place. Then, they assign the limit ratio as the probability of certain events. The intuition of this interpretation is that we believe the underlying parameter is invariant so after many repeated experiment the random effect can be cancelled out and the real instinct reveals.

2 The second kind of view is just based on the "subjective belief of interpreter" in the chance of an event occurring. We can just give probability under our faith before any experiment has been done and then do experiment to study whether our prior assumption fit the data or not and then modify it. This procedure can be done by Bayes' theorem.

The counting problem is sometimes sophisticated and along with many restrictions. The way to solve such problem is to break them into series of simple tasks and apply rules to combine it back.

Theorem 1 (Fundamental Theorem of Counting) *Suppose a job consists of k separate tasks, the ith of which can be done in n_i ways, then entire job can be done in $\prod_{i=1}^{k} n_i$ ways.*

Intuition (Classification of counting methods)

The proof of Fundamental Theorem of Counting is quite trivial. Sometimes it is better to think of task as partition criteria such as love or hate, different gender or income level then we just construct a sample space $\Omega = \Omega_1 \times \Omega_2 ... \Omega_k$ which is the cartesian product of k criteria we interested in then we can apply the theorem to calculate there are how many possible outcomes.
In reality, we may face situation such as replacement and unordered. Replacement means that $n_i = n_{i+1}$ and unordered means that we can perform k separate tasks arbitrary without certain order. In such case we have to carefully apply or modify the fundamental counting theorem.

Consider number of possible arrangements of size r drawing from n different subjects. We can divide the case into four categories.

	Without replacement	With replacement
Ordered	$n*(n-1)*(n-2)...(n-r+1)=\frac{n!}{(n-r)!}$	$n*n*n...n=\prod_{i=1}^{r} n=n^r$
Unordered	$\frac{\prod_{i=1}^{r}(n-i+1)}{r!}=\frac{n!}{(n-r)!(r!)}=\binom{n}{r}$	$\binom{n+r-1}{r}$

Proof: Case 1 and case 2 is quite simple just applying the Fundamental Theorem of Counting. For case 3, r different objects can be permuted in $r*(r-1)*(r-2)...1=r!$ ways but they represent the same arrangement in unordered situation. So we have case 3 equals case 1 divided by r!. Case 4 is the most difficult, we may simply view it as case 2 divided by r! but it will underestimate the possible arrangements since r objects with some of them are of same kind do not construct r! different permutations. A clever way to solve case 4 is to think of r as numbers of coins and place it arbitrary but all into n different box. A coin in ith box means in our ultimate arrangements we have one ith objects. Consider a small case as n=3 and r=3. Then the following figure is the realization of picking two 2th object and one 3th object. And all possible realization cab be expressed as all possible arrangements of 2 | and 3 O which is $\binom{3-1+3}{3}$.

$$\_\_^1|\underline{OO}^2|\underline{O}^3$$

So for case 4, the answer is $\binom{n-1+r}{r}$. ■

The counting techniques are useful when the sample space is finite and every possible outcomes in S are equally likely. The probability of certain event can be calculated by the number of outcomes in that event times the probability of each outcome from the countably addictive axiom.

Theorem 2 (Enumerating outcomes) *Let* $\Omega=\{w_1,w_2...w_n\}$ *with* $P(\{w_i\})=\frac{1}{n}\ \forall i$ *then* $P(A)=\sum_{\{w_i\}\in A}P(\{w_i\})=\sum_{\{w_i\}\in A}\frac{1}{n}=\frac{\sharp(A)}{\sharp(\Omega)}\ \forall A$.

Remark 1 This is also the classical definition of probability from Pierre-Simon Laplace.

2 Conditional probability

In reality, we may need to study something like if she is a girl, what is the probability that she wants to get married. Studying these kinds of probabilities under certain situation or restrictions of sample space needs the definition of conditional probability.

Theorem 3 *Let P(.) be a probability measure on* σ*-algebra* F*, A and B be events in* F *and P(B)>0 then,* $P(.|B)$*, the conditional probability of A given B is denoted by*

$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

First we need to verify $P(.|B)$ is a probability measure.

1 $P(A|B)\geq 0\ \forall A\in F$: Since $P(B)>0$ and $P(A\cap B)\geq 0$ because P(.) is a probability measure, then the ratio $\frac{P(A\cap B)}{P(B)}\geq 0\ \forall A\in F$.

2 $P(\Omega|B) = 1 : \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$

3 $A_1, A_2 ... A_n ... \in F,\ A_i \cap A_j = \phi$ then $P(\cup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} P(A_i|B)$: First notice that $A_1 \cap B, A_2 \cap B ... A_n \cap B ...$ are all disjoint then $P(\cup_{i=1}^{\infty} A_i|B) = \frac{P(\cup_{i=1}^{\infty} A_i \cap B)}{P(B)} = \frac{P(\cup_{i=1}^{\infty}(A_i \cap B))}{P(B)} = \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)} = \sum_{i=1}^{\infty} P(A_i|B).$

$P(.|B)$ is indeed a probability measure, but what is its statistical meaning?

Example 1 Consider a case that we need to study how accurate our new AIDS test is, then the sample space is partitioned by two criteria. The test result is positive/negative and the subject has AIDS or not. We have $\Omega : \{(Test, Disease) : T \in \{+, -\}, D \in \{+, -\}\} = \Omega_T \times \Omega_D$, then

$$P(D_+|T_+) = \frac{P(D_+ \cap T_+)}{P(T_+)} = \frac{P(\{(D_+, T_-), (D_+, T_+)\} \cap \{(D_+, T_+), (D_-, T_+)\})}{P(\{(D_+, T_+), (D_-, T_+)\})}$$

So the conditional probability is actually the study of how sub-sample space Ω_T, Ω_D will affect each other on Ω i.e. what is the relation between each partition criteria, we may not simply think conditional probability in geometric view.

Intuition (Conditional Probability)

Generally speaking, $P(B) > 0$ cannot always be satisfied since we may be interested in continuous data such as heights and weights. We face division by zero when study problem like given father is 1.75 tall what is the probability the son is 1.80 tall. We have $P(\text{father is } 1.75) = 0$ and above definition does not work any more. A natural way to save it is to think of B as a small neighbourhood $(B - \frac{\triangle}{2}, B + \frac{\triangle}{2})$ so we have

$$P(A|B) = \frac{\lim_{\triangle \to 0} P(A \cap (B - \frac{\triangle}{2}, B + \frac{\triangle}{2}))/\triangle}{\lim_{\triangle \to 0} P(B - \frac{\triangle}{2}, B + \frac{\triangle}{2})/\triangle}$$

Noting that both the numerator and denominator are of change rate form so $P(A|B)$ can be thought of as the intensity of the ratio of change rate.
In the previous test-disease example, we can not do medical behavior relying on uncertified test for moral issue. We may only do the reverse direction test. Do the test to patients who have the disease and what is the test result respond to it. So we have only $P(T_+|D_+)$ and also the prevalence rate of the disease $P(D_+)$. How can we know $P(D_+|T_+)$? This is where we need the Bayes' theorem.

Example 2 [Casella 1.3.4] Three prisoners A,B and C are on death row. The governor decides to pardon one of the three prisoners randomly. He informs the warden of his choice but ask warden to keep it secret. A tries to get the warden to tell him who had been pardoned. The warden refuses so A asks another way. A asks which of B and C will be executed. And the following table is warden's possible react.

The warden tells A that B would be executed, does it reveal any message to A?

Pardoned	Warden tells	Probability
A	B dies	r
	C dies	1-r
B	C dies	1
C	B dies	1

Solve: Let W denote the event that warden says B will die then A can update his probability given W.

$$\begin{aligned} P(\text{A pardoned } |W) &= \frac{P(A \cap W)}{P(W)} \\ &= \frac{P(\text{A pardoned B dies})}{P(\text{A pardoned B dies}) + P(\text{C pardoned B dies}) + P(\text{B pardoned B dies})} \\ &= \frac{\frac{1}{3} \cdot r}{\frac{1}{3} \cdot r + \frac{1}{3} + 0} = \frac{r}{r+1} \end{aligned}$$

The answer would rely on warden's behavioral pattern.

	$P(A\|W)$	Message
$r = 1$	$\frac{1}{2}$	good news
$r = \frac{1}{2}$	$\frac{1}{3}$	no news
$r = 0$	0	bad news

Intuition (Three Prisoners)

When r tends to 1, the probability A would be pardoned reaches $\frac{1}{2}$. This is because when A had been pardoned, the warden would always say B and event W becomes a more good predictor of A pandered and vice versa. When $r = \frac{1}{2}$, the warden has the same chance to tell B or C dies when A had been pardoned so W can not be used to trace back A pardoned. Sometimes people may think $P(A|W) = \frac{1}{2}$ because one of A and C would be saved. But it is $P(A|B^C)$ actually.
Consider $P(C|W)$ it is actually $1 - P(A|W)$. So when $r = \frac{1}{2}$, $P(C|W) = \frac{2}{3}$. Though warden reveals no message to A but it is actually a good news for C. It is like the situation we switch door in the Monty Hall problem. Conditional probability can be really tricky.

■

3 Bayes Rule

3.1 Bayes rule

Imagine a situation that there are two sample spaces $\Omega_1 = \{D_1, ..., D_p\}, \Omega_2 = \{O_1, ..., O_q\}$ that augments a larger sample space $\Omega = \Omega_1 \times \Omega_2$. What kind of inferences can we draw from the knowledge of outcome O? If I know the conditional probability $P(O|D)$ what can do to inference $P(D|O)$?

Intuitively, there are four categories of probabilities here:

- $P(D)$: probability of getting disease.
- $P(O)$: probability of yielding some symptoms (outcomes).
- $P(D|O)$: probability of getting disease conditioned on having some symptoms.
- $P(O|D)$: probability of having some symptoms conditioned on getting a disease.

The four categories are connected together with the following Bayes rule:

$$P(D = d|O = o) = \frac{P(O = o|D = d)}{\sum_{d \in D} P(D = d)P(O = o|D = d)}$$

The important philosophy here is that, in real life, we won't have all the four categories of probability in hand. That is, we have to draw inference on the probability we care from the probability we known. For example, a doctor want to know the probability of a patient getting disease d conditioned on he has symptom o. If the doctor knows the probability of getting a disease ($P(D)$), and the probability of a outcome to happen conditioned on having disease or not ($P(O|D)$). Then, by applying Bayes rule, he can calculate the conditioned probability $P(D|O)$, which is what he concerns.

Here I a little abuse the notation of D, actually, the event in D must be a **partition**. For example $D = \{\text{disease}, \text{no disease}\}$.

Now, we formally state the Bayes rule as follow:

Theorem 4 (Bayes rule) *Let $A_1, ..., A_n$ be a partition over $\mathcal{A}$ and $B \in \mathcal{A}$, then*

$$P(A_i|B) = \frac{P(B|A_i)}{\sum_j P(A_j)P(B|A_j)}$$

Intuition (Bayes rule)

Bayes rule help us update the probability over a **partition** on event space. It utilize what we have:

- The probability of population on the partition.
- The conditioned probability of the observed event conditioned on each event in the partition.

Then calculate the conditioned probability we want to inference: The probability of each event in the partition to happen conditioned on an observation.

3.2 Relative risk and Odd ratio

Both relative risk and odd ratio are important statistical concept in epidemiology/experimental study. In this context, we care the following scenario: There's a disease and a treatment, we want to know how well the treatment is but in the meantime hoping the treatment has less side-effect. We can summarize the above scenario in Table 1. Here, D refers to the event that the patient

has the disease and E is the event that the patient is under treatment (or formally, being exposed to the treatment). In the context of experimental study, see Table 2, E can be regarded as the experimental group and E^C can be seen as the control group. Here D and D^C can be simply considered as the presence and absence of an event.

Risk		Treatment	
		Exposed	Not exposed
Disease	Diseased	$P(D\|E)$	$P(D\|E^C)$
	Healthy	$P(D^C\|E)$	$P(D^C\|E^C)$

Table 1: Epodemiology scenario.

Risk		Experiment	
		Experiment	Control
Outcome	Presence	$P(D\|E)$	$P(D\|E^C)$
	Absence	$P(D^C\|E)$	$P(D^C\|E^C)$

Table 2: Experimental study.

With this scenario, we an immediately some intuitive term to help us making inference. The first one is **relative risk (RR)**. Relative risk is the probability ratio of an event to happen under certain exposure or not. In this context, it's simply $\frac{P(D|E)}{P(D|E^C)}$.

Definition 5 (relative risk) *Relative risk (RR) is the probability ratio of an event to happen under an exposure or not. That is*

$$RR := \frac{P_{event\ when\ exposed}}{P_{event\ when\ not\ exposed}}$$

Note that RR can help us inference on the effectiveness of the treatment on the disease. See Table 3 for more details on RR.

	Epidemiology	Experimental study
RR>1	Thetreatment is worse than having no treatment.	The event is more likely to happen in the **experimental** group.
RR<1	The treatment is effective.	The event is more likely to happen in the **control** group

Table 3: Relative risk.

Now, we can use relative risk to describe the effectiveness of a treatment/experiment on a certain disease/event, we might wander: how about the relation between two diseases/events under the same treatment/experiment? And this is actually the definition of odd ratio.

Definition 6 (odd ratio) *The odd ratio (OD) is the ratio of the relative risk of two events. That is, the odd ratio of event A and B w.r.t a treatment E is*

$$\begin{aligned} OD :&= \frac{RR_A}{RR_B} \\ &= \frac{P(A|E)/P(A|E^C)}{P(B|E)/P(B|E^C)} \end{aligned}$$

The simplest odd ratio is to consider a event and its complement, say D and D^C, then the odd ratio will become

$$\begin{aligned} \frac{P(D|E)/P(D|E^C)}{P(D^C|E)/P(D^C|E^C)} &= \frac{P(D\cap E)/P(D\cap E^C)}{P(D^C\cap E)/P(D^C\cap E^C)} \\ &= \frac{P(D\cap E)/P(D^C\cap E^C)}{P(D^C\cap E)/P(D\cap E^C)} \end{aligned}$$

With this transformation, we can use the four intersection probability to calculate the odd ratio instead of the four conditional probability. In some circumstances, this will be more convenient and intuitive.

4 Independence

The initial idea (and the most intuitive concept) of independence is that two event A and B is said to be independent if the probability of A to happen will not change after we have the knowledge of B, and vice versa. Thus, formally we can write

$$P(A) = P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B) = P(B|A) = \frac{P(A \cap B)}{P(A)}$$

And we can see that actually as A and B are independent, $P(A \cap B) = P(A)P(B)$, while the converse is also correct. As a result, this has become the definition of independence.

Definition 7 (independence) *We say two event A and B are independent if $P(A \cap B) = P(A)P(B)$.*

However, note that there are more than one concept about independence in mathematics. For instance, in linear algebra, there's so called linear independence. And even in probability theory, as we consider more than one event, say a group of events, the idea of independence varies. Actually, there are two kinds of independence for more than two events: **pairwise independence** and **mutually independence**.

Definition 8 (pairwise independence) *We say a finite set of events $\{A_i\}$ is pairwise independent if $\forall A_i \neq A_j$, $P(A_i \cap A_j) = P(A_i)P(A_j)$.*

Definition 9 (mutually independence) *We say a finite set of events $\{A_i\}$ is mutually independent if for any subset of $\{A_i\}$, say $\{A'_j\}$, $P(\bigcap_j A'_j) = \prod_j P(A'_j)$.*

Note that mutually independence is **strictly stronger** than pairwise independence. That is, the former implies the latter while the converse is not necessarily true.