Statistical Inference Notes

Chapter 1: Probability Theory

Contents

Probability Theory

"The subject of **probability theory** is the foundation upon which all of statistics is built, providing a means for modeling populations, experiments, or almost anything else that can be considered a random phenomenon. Through these models, statisticians are able to draw inferences about populations, inferences based on examination of only a part of the whole."

Set Theory

Definition 1.1. The set S of all possible outcomes of a particular experiment is called the **sample space** of the experiment.

Example 1.2. H, T is the sample space for flipping a coin. $2, 3, ..., 11, 12$ is the sample space for the rolling of two dice.

A sample space may be countable or uncountable. The above are two examples of countable sample spaces, while examples of an *un*countable sample space might include $S = [0, 1]$ for the generation of a random number between 0 and 1, or $[0, \inf)$ representing the distance traveled by some particle in an experiment.

Definition 1.3. An **event** A is any collection of possible outcomes of an experiment, i.e., any subset of S. A is said to occur if the outcome of the experiment is in the set A.

Properties 1.4. Let A and B be any two sets (events). The following are some useful characteristics of sets.

- **Containment**: $A \subset B \iff x \in A \implies x \in B$
- **Equality**: $A = B \iff A \subset B$ and $B \subset A$
- **Union**: $A \cup B = \{x : x \in A \text{ or } x \in B\}$
- **Intersection**: $A \cap B = \{x : x \in A \text{ and } x \in B\}$
- **Complement**: $A^c = \{x : x \notin A\}$

Theorem 1.5. *For any events A, B, C defined on a sample space S, the following properties hold:*

Commutativity:

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

Associativity:

$$A \cup (B \cup C) = (A \cup B) \cup C$$
$$A \cap (B \cap C) = (A \cap B) \cap C$$

Distributive Laws:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

De Morgan's Laws:

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

We can also extend the notions of union and intersection to infinite collections of sets $A_1, A_2, \ldots$:

$$\bigcup_{i=1}^{\infty} A_i = \{x \in S : x \in A_i \text{ for some } i\}$$
$$\bigcap_{i=1}^{\infty} A_i = \{x \in S : x \in A_i \text{ for all } i\}$$

Note: We denote with $\emptyset$ the **empty set**, i.e., the set having no elements in it.

Definition 1.6. Two events A and B are **disjoint**, or **mutually exclusive**, if $A \cap B = \emptyset$. The events $A_1, A_2, \ldots$ are **pairwise disjoint** if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Definition 1.7. If $A_1, A_2, \ldots$ are pairwise disjoint and $\cup_{i=1}^{\infty} A_i = S$, then we say that $A_1, A_2, \ldots$ form a **partition** of S.

Example 1.8. The most obvious partition of S would be the sets $\{A, A^c\}$ for any event A. (This may seem rather pointless but we will revisit its use later when discussing Bayes' Theorem.)

Basics of Probability Theory

To begin this section we outline the *very basic* facets of measure theory, which provide the under-the-hood mechanisms with which we can formally define and talk about probability.

Definition 1.9. A collection of subsets of S is a **sigma algebra**, or **Borel field**, denoted by $\mathcal{B}$, if it satisfies the following properties:

- $\emptyset \in \mathcal{B}$
- If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$ (closed under complementation)
- If $A_1, A_2, \cdots \in \mathcal{B}$, then $\cup A_i \in \mathcal{B}$ (closed under countable unions)

Example 1.10. The simplest example of a sigma algebra is the **trivial sigma algebra**, $\{\emptyset, S\}$.

Example 1.11. If S is finite or countable, we can take $\mathcal{B} = \{\text{all subsets of } S\}$.

Example 1.12. If S is uncountable, we may take $\mathcal{B}$ = all sets of the form $[a,b], [a,b), (a,b], (a,b)$

Usually we'll just take $\mathcal{B}$ to be the simplest sigma algebra possible: the smallest one containing all open sets in a given sample space S. Now we may take a look at defining probability...

Definition 1.13. Let S be a sample space with an associated sigma algebra $\mathcal{B}$. A **probability function** is a function P with domain $\mathcal{B}$ satisfying

- $P(A) \geq 0$ for all $A \in \mathcal{B}$
- P(S) = 1
- If $A_1, A_2, \dots \in \mathcal{B}$ are pairwise disjoint, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

These conditions are known as the Axioms of Probability, or the Kolmogorov Axioms. In particular, the last axiom is known as the Axiom of Countable Additivity.

In practice, we don't want to always check these axioms to verify whether a function is a valid probability function. The below result helps in this regard by giving a systematic way to confirm a valid probability function.

Theorem 1.14. *Let $S = \{s_1, \dots, s_n\}$ be a finite set and let $\mathcal{B}$ be any sigma algebra of sets of S. Also let $p_1, \dots, p_n$ be nonnegative integers that sun to 1. For any $A \in \mathcal{B}$, let $P(A)$ be defined by*

$$P(A) = \sum_{i : s_i \in A} p_i$$

Then, P is a probability function on $\mathcal{B}$.

There are many important properties of probability functions as well as relations we can make between multiple events that will allow us to find probability values for events. We outline these now.

Theorem 1.15. *Let P be a probability function and let $A \in \mathcal{B}$ be an event. Then:*

- $P(\emptyset) = 0$
- $P(A) \leq 1$

- $P(A^c) = 1 - P(A)$

Theorem 1.16. *Let P be a probability function and let $A, B \in \mathcal{B}$ be any events. Then:*

- $P(A \cap B^c) = P(A) - P(A \cap B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $If A \subset B$*, then* $P(A) \leq P(B)$

Corollary 1.17. *For any events $A, B \in \mathcal{B}$,*

$$P(A \cap B) \geq P(A) + P(B) - 1$$

This result is known as Bonferroni's inequality and is useful for bounding the probability of the intersection of two events, using the probabilities of the individual events.

Theorem 1.18. *Let P be a probability function. Then:*

- $P(A) = \sum_{i=1}^{\infty} A \cap C_i$ *for any partition* $C_1, C_2, \ldots$
- $P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$ *for any sets* $A_1, A_2, \ldots$ *(Boole's inequality)*

We can apply Boole's inequality above to the event A^c and obtain an expanded version of Bonferroni's inequality, in more general terms, in the following way:

$$P\left(\bigcup_{i=1}^{\infty} A_i^c\right) \leq \sum_{i=1}^{\infty} P(A_i^c)$$

Now, note that $P(A_i^c) = 1 - P(A_i)$ and $\cup A_i^c = (\cap A_i)^c$. Then,

$$1 - P\left(\bigcap_{i=1}^{\infty} A_i\right) \leq n - \sum_{i=1}^{\infty} P(A_i) \tag{1}$$

$$P\left(\bigcap_{i=1}^{\infty} A_i\right) \geq \sum_{i=1}^{\infty} P(A_i) - (n-1) \tag{2}$$

Now we move on to the topic of counting, which is used for constructing probability assignments on finite sample spaces as well as other tasks.

Theorem 1.19 (**Fundamental Theorem of Counting**)**.** *Suppose that a "job" consists of k separate tasks where the ith task can be done in n_i different ways, for $i = 1, \ldots, k$. Then, the entire job can be done in $n_1 \times \cdots \times n_k$ ways.*

While this theorem can be used for basic examples, usually we must make a couple distinctions depending on the problem at hand. Namely, we must differentiate between counting with versus without replacement, and between the cases where ordering does and does not matter to the outcome. We can arrange the following table of the number of possible arrangements of size r from n objects.

	without replacement	with replacement
Ordered	$\frac{n!}{r!(n-r)!}$	n^r
Unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

If we have equally likely outcomes in S, then we can use the above to count the number of outcomes in a given event A to find $P(A)$:

$$P(A) = \sum_{s_i \in A} P(\{s_i\}) = \sum_{s_i \in A} \frac{1}{N} = \frac{\text{number of elements in } A}{\text{number of elements in } S}$$

Example 1.20. Find the probability of various poker hands.

$$P(4 \text{ of a kind}) = \frac{13 \times 48}{\binom{52}{5}} =$$

Conditional Probability and Independence

The notion of conditional probability will allow us to update our sample space at hand based on new information/data. For example, to find the probability of drawing 4 aces from a deck of cards, we can say $P(4 \text{ aces}) = \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49}$, where with each term we are updating the sample space of possible cards we may draw from, using the knowledge of the card(s) just drawn.

Definition 1.21. If A and B are events in S, where $P(B) > 0$, then the conditional probability of A given B is given by:

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

Applying the above definition, we can also conveniently write probabilities of intersections, with

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Theorem 1.22 (Bayes' Theorem). *Let A, B be sets in S. Then,*

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

In general, let $A_1, A_2, \ldots$ be a partition of S and B be any event set. Then,

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{\infty} P(A_j)P(B|A_j)}$$

Definition 1.23. Events A and B are statistically independent if

$$P(A \cap B) = P(A)P(B) \iff P(A|B) = P(A) \iff P(B|A) = P(B)$$

Theorem 1.24. *Let A and B be independent events. Then, the following pairs of events are also independent:*

- *A and B^c*
- *A^c and B*
- *A^c and B^c*

Definition 1.25. A collection of events $A_1, \ldots, A_n$ are mutually independent if for any subcollection of events $A_{i_1}, \ldots, A_{i_k}$, we have:

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j})$$

Random Variables

Definition 1.26. A random variable X is a function from a sample space S into $\mathbb{R}$

Suppose that we have a sample space $S = \{s_1, \ldots, s_n\}$ with a probability function P and we define a new random variable X with range $\mathcal{X} = \{x_1, \ldots, x_m\}$. Then, we can define a probability function P_X on $\mathcal{X}$ with:

$$P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\})$$

We say that P_X is an *induced* probability function on $\mathcal{X}$.

Example 1.27. Consider an experiment involving flipping a coin three times and define the random variable X to be the number of heads in those 3 rolls. Then, we have the following:

s	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(s)$	3	2	2	2	1	1	1	0

The induced probability function P_X will be given by:

x	0	1	2	3
$P_X(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

The above were for the case where S and $\mathcal{X}$ are finite or countable. We can define P_X for an uncountable $\mathcal{X}$ similarly. For any set $A \subset \mathcal{X}$,

$$P_X(X \in A) = P(\{s \in S : X(s) \in A\})$$

Distribution Functions

Definition 1.28. The cumulative distribution function (cdf) of a random variable X is defined by:

$$F_X(x) = P_X(X \leq x)$$

for all x.

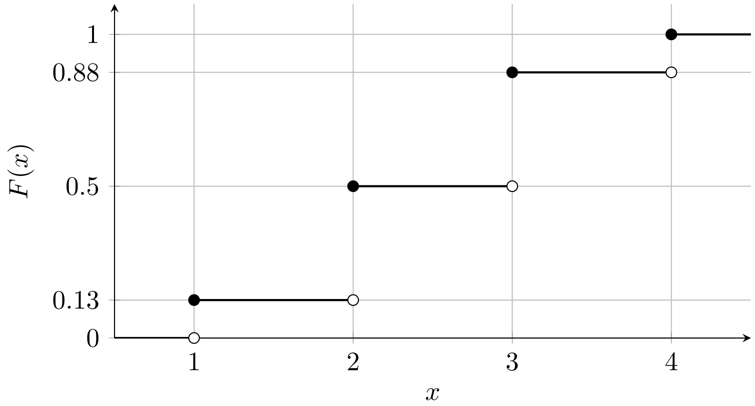

Figure 1: Cumulative distribution function (cdf) of Example **??**

Theorem 1.29. *The function $F_X(x)$ is a cdf if and only if the following hole:*

- $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$
- *F(x) is a nondecreasing function*
- *F(x) is right-continuous, i.e.,* $\lim_{x\to x_0^+} F(x) = F(x_0)$ *for any* x_0

Example 1.30. Consider an experiment where we flip a coin until we get a heads. (TO DO: Example 1.5.4)

Definition 1.31. A random variable X is continuous if $F_X(x)$ is a continuous function and discrete if $F_X(x)$ is a step function.

Definition 1.32. Let $\mathcal{B}$ be the smallest sigma algebra containing all the intervals of real numbers of the form $[a, b], [a, b), (a, b], (a, b)$. The random variables X and Y are identically distributed if for all $A \in \mathcal{B}$, $P(X \in A) = P(Y \in A)$

Example 1.33. If we revisit the experiment of flipping a coin 3 times and define the random variables X = number of heads and Y = number of tails, then X and Y are identically distributed (even though we don't have $X(s) = Y(s)$ for any sample $s \in S$).

Theorem 1.34. *The random variables X and Y are identically distributed if and only if $F_X(x) = F_Y(x)$ for all x.*

Density and Mass Functions

Definition 1.35. The probability mass function (pmf) of a discrete random variable X is given by

$$f_X(x) = P(X = x) \quad \text{for all } x$$

Definition 1.36. The probability density function (pdf) of a continuous random variable X is the function $f_X(x)$ that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \quad \forall x$$

Naturally, we also have the relationship $\frac{d}{dx}F_X(x) = f_X(x)$

Notation: If X follows a distribution given by the cdf F_X, we write $X \sim F_X(x)$. Likewise, we may also write $X \sim f_X(x)$. Lastly, if X follows the same distribution as another random variable Y, then we could write $X \sim Y$.

We can use the relationship between the cdf and pdf to find probabilities.

$$P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx$$

For a continuous random variable X, we will have $P(X = x) = 0$ for all x. So, for some points $a, b \in S$, we have that

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b)$$

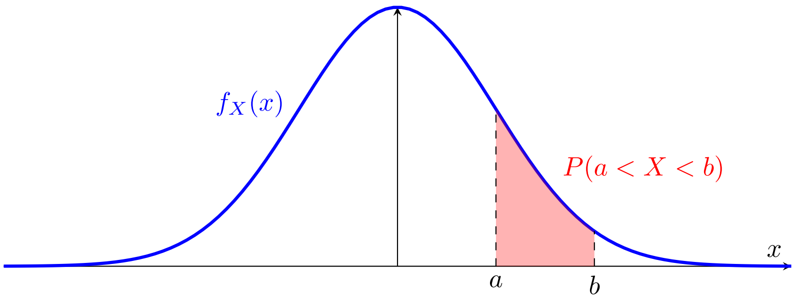

Figure 2: Probability using area under the density curve $f_X(x)$.

Theorem 1.37. *A function $f_X(x)$ is a pdf/pmf of a random variable X if and only if:*

- $f_X(x) \geq 0 \quad \forall x$
- $\sum_x f_X(x) = 1$ *if f_X is a pmf;* $\int_{-\infty}^{\infty} f_X(x)dx = 1$ *if f_X is a pdf.*