STA257 Notes

Max Xu

'25 Fall

Contents

§1 Day 1: Sample Spaces and Probability (Sept 3, 2025)

Life is very random and uncertain, with interesting problems to solve (e.g. whats the probability that I win the lottery). This course uses certain mathematics to study the uncertainty of probabilities. You will be able to solve problems like this:

Problem 1.1. Which is more likely: getting at least one six when rolling a fair 6 sided die 4 times, or getting one pair of sixes when rolling two six sided dice 24 times?

5% of your grade is poll-based, more info here.

Definition 1.2 (Sample Space)**.** A non-empty set containing all possible outcomes, written S.

e.g. coin-flipping: $S = \{\text{Heads}, \text{Tails}\}$, two die: $S = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$

Definition 1.3 (Event)**.** Any subset $A \subseteq S$ is an event.

Prof says in some continuous sample spaces, there may exist some non-measurable subsets to which the probability measure defined later won't work on, but don't worry about it in this course. (yay!!)

Definition 1.4 (Probability)**.** For any event A, define probability $\mathrm{P}(A)$ that satisfies:

- For all $A \subseteq S$, $0 \leq \mathrm{P}(A) \leq 1$
- $A = S$, $\mathrm{P}(A) = \mathrm{P}(S) = 1$
- $A = \emptyset$, corresponding to no outcome, then $\mathrm{P}(A) = \mathrm{P}(\emptyset) = 0$
- **Additivity**: if $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$

If $A_1, \ldots, A_n$ are disjoint[1] events, we have

- **Finite Additivity**: For some $n \in \mathbb{N}$,

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathrm{P}(A_i)$$

- **Countable Additivity**:

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

[2]

When looking at the probability of getting heads from a coinflip, $\mathrm{P}(H)$ with $S = \{T, H\}$ is really shorthand for $\mathrm{P}(\{H\})$, since H may not be a subset of S. For uniformly picking any number between 0 and 1, denoted Uniform$[0,1]$, we can define a probability $\mathrm{P}([a,b]) = b - a$ whenever $0 \leq a \leq b \leq 1$. (don't know what uniform means yet)

Note that by definition of probability, $\mathrm{P}(A_i)$ is positive, so the right hand side is an absolutely convergent series (prof didn't mention this).

[1] prof said this but i think he meant pairwise disjoint?

[2] we could've derived all the same results with just 3, see probability axioms

§2 Day 2: Properties of Probability (Sept 8, 2025)

§2.1 Additional Properties of Probability

Today we will be deriving properties of probability from the 'axioms' we stated last class. Note that most of these follow from the additivity property 1.4. Have $A, B \subseteq S$ be events.

Theorem 2.1. If A^C is the complement of A, then $P(A^C) = 1 - P(A)$.

Proof. A and A^C are by definition disjoint, and their union is S. By additivity 1.4 $P(A) + P(A^C) = P(S) = 1$. □

Theorem 2.2. $P(A) = P(A \cap B) + P(A \cap B^C)$

The set of $\{x \in A : x \in B\}$ and $\{x \in A : x \notin B\}$ are by definition disjoint, and the union of the two is A. This then follows by additivity 1.4.

Theorem 2.3. If A contains B, $P(A) = P(B) + P(A \cap B^C)$

Proof. Have 2.2, except $P(A \cup B) = P(B)$ where $A \supseteq B$. □

Theorem 2.4 (Monotonicity)**.** If $A \supseteq B$, then $P(A) \geq P(B)$

Immediately follows from 2.3, since $P(A \cap B^C)$ must be non-negative, giving the inequality.

Theorem 2.5 (Law of Total Probability)

Suppose $A_1, A_2, \ldots$ are a sequence of events which form a *partition* of S (pairwise disjoint) , with their union being the entire sample space ($\bigcup_i A_i = S$). Let B be any event. Then we have

$$P(B) = \sum_i P(A_i \cap B)$$

Theorem 2.6 (Principle of Inclusion-Exclusion)**.** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof. The events $A \cap B^C$, $B \cap A^C$, $A \cap B$ are disjoint events.

$$\begin{aligned} P(A \cup B) &= P(A \cap B) + P(A \cap B^C) + P(B \cap A^C) \\ &= P(A \cap B) + [P(A) - P(A \cap B)] + [P(B) - P(A \cap B)] \qquad \text{from 2.2} \end{aligned}$$

□

For a more generalized version of inclusion-exclusion formula, look at Challenge 1.3.10 in textbook.

Theorem 2.7 (Subadditivity)**.** For any sequence of events $A_1, A_2, \ldots$ not necessarily pairwise disjoint, have

$$P(A_1 \cup A_2 \ldots) \leq P(A_1) + P(A_2) + \ldots$$

This is Theorem 1.3.4 in the textbook, with its proof found in section 1.7.

Remark 2.8. This is more of a worry in grad-level courses, where you study more pathological probability spaces, but 'uncountable' subadditivity does not exist. Consider $S = \text{Uniform}([0,1])$. Have $A_x = \{x\}$ for $x \in S$. $P(\bigcup_{x\in S})A_x = P(S) = P([0,1]) = 1$. Yet for any 'singleton' x, $P(A_x) = P(\{x\}) = 0$, meaning $\sum_{x\in S} P(A_x) = 0$.

§2.2 Uniform Probabilities on Finite Spaces

Have $S = \{s_1, \ldots, s_n\}$. For all $\{s_i\}$ to have the same probability, $P(\{s_i\}) = \frac{1}{n}$, called a *discrete uniform distribution*.

Any $A \subseteq S$ with k elements, would have $P(A) = \frac{k}{n}$, meaning

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

A problem solving technique to find P of a rather complicated event is to see if the probability of its complement can be easily found, then use 2.1.

§3 Day 3: 'Combinatorics' (Sept 10, 2025)

For the first few weeks, we will mostly be dealing with uniform spaces. Be careful, if the probability is non-uniform, meaning not all outcomes are equally likely, the counting technique from last class would not apply.

The sample space can also be a discrete infinite set, e.g. $S = \mathbb{N} = \{1, 2, \dots\}$, with $P(\{i\}) = 2^{-i}$ for $i \in \mathbb{N}$. We can check that this is valid by checking that each $0 \leq P(\{i\})$

$$\sum_{i=1}^{\infty} 2^{-i} = 1$$

To get the probability of the even numbers, we can compute the sum

$$\sum_{i=2,4,6,\dots}^{\infty} 2^{-i} = \frac{1}{3}$$

which is quite surprising.

On a discrete infinite space, we cannot have a uniform distribution.

§3.1 More Finite Uniform Probabilities

The number of ways to pick k distinct items *in order* out of n items total, is

$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

which is also called a 'permutation', written $P(n, k)$.

There are $k!$ ways to order k distinct objects. For this reason, the number of ways to pick k distinct *unordered* objects,

$$n(n-1)\cdots(n-k+1)/k! = \frac{n!}{(n-k)!k!}$$

This formula is called 'combinations', 'choose formula', or 'binomial coefficient', written $C(n, k)$, n choose k, and $\binom{n}{k}$ respectively.

$$C(n, k) = \frac{P(n, k)}{k!}$$

> **Remark 3.1.** Regarding the lottery, my advice is to not buy a lottery ticket. But if you really wanted to, you should avoid common patterns, valid birthdays etc... so you can avoid having to share the winnings with another person. - Prof Rosenthal

In a standard deck of playing cards, there are 4 suits, with each suit having 13 ranks, making $4 \cdot 13 = 52$ cards total.

$$P(\text{Clubs or 7}) = P(\text{Clubs}) + P(7) - P(\text{Clubs and 7})$$

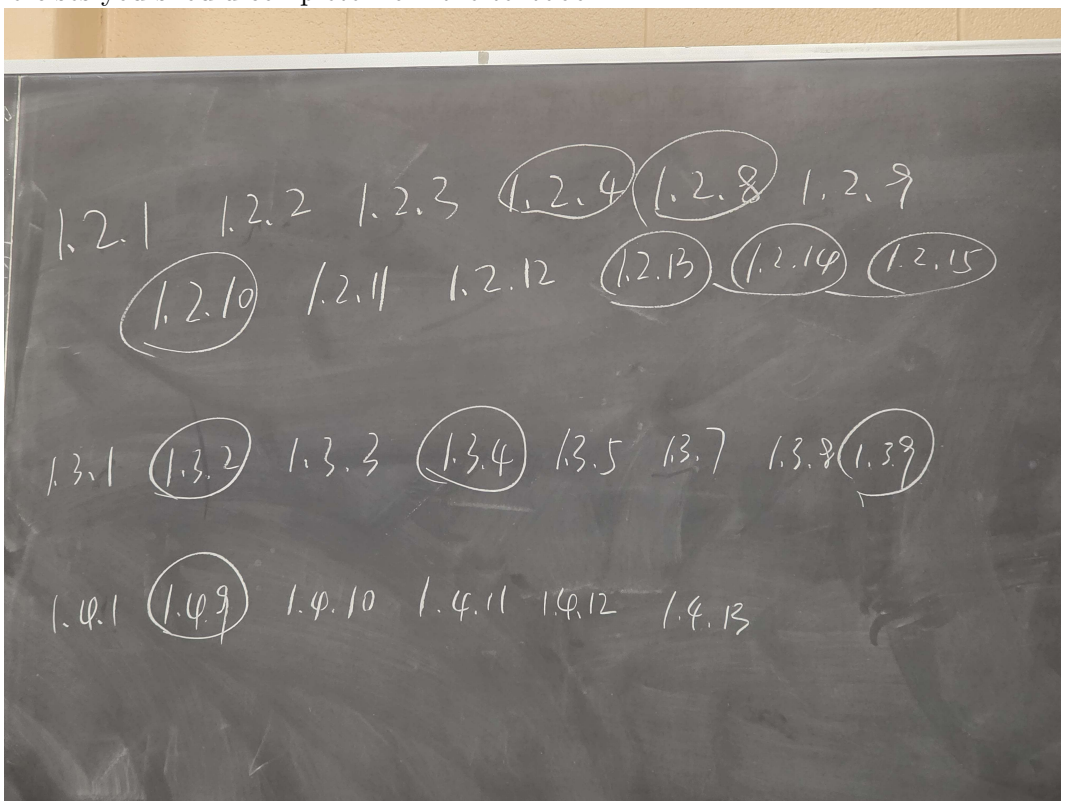
which is by inclusion-exclusion 2.6.

§4 Tutorial 1 (Sept 10, 2025)

DeMorgan's laws state that

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

Exercises you should complete from the textbook:



§5 Day 4: Conditional Probability (Sept 15, 2025)

From last class: choosing a subset in order is called a permutation, choosing a subset irrespective of the order is called a combination. You don't have to use R in this course, but if you wanted to there is some info here.

Problem 5.1. Suppose we flip 4 fair coins, what is $P(\text{exactly 2 heads})$?

You could solve this by writing out the entire sample space. Or by computing $\frac{\binom{4}{2}}{2^4} = \frac{3}{8}$. In general, for flipping n coins, the probability of getting exactly k heads is

$$\frac{\binom{n}{k}}{2^n} \text{for } 0 \leq k \leq n$$

§5.1 Conditional Probability

We now receive some information that restricts the sample space of interest to some subset of the original sample space S. If P was a discrete uniform distribution, P on said subset is remains a discrete uniform distribution.

Definition 5.2 (Conditional Probability)**.** If A and B are two events, where $P(B) > 0$, then the *conditional probability* of A given B is written $P(A \mid B)$ represents the fraction of the times when B occurs, in which A also occurs.

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

If $P(B) = 0$, then $P(A|B)$ is undefined. Assuming that an event with probability 0 occurring would lead to all sort of contradictions that I don't want to explore.

Theorem 5.3 (Conditional Multiplication Formula)**.**

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

Combining with law of total probability 2.5 we can get a more useful version where we replace $P(A_i \cap B)$ according to 5.3, giving

$$P(B) = \sum_i P(A_i)P(B \mid A_i)$$

which is a lot more useful. Remember that $\bigcup_i A_i = S$.

Problem 5.4 (Challenge)**.** Roll n fair six sided dice. What are the odds we get more than $0 \leq k \leq n$ 5s?

§6 Day 5: Independence (Sep 17, 2025)

When solving problems, solve them systematically, it is very easy to get tricked. English is already a complex language, and reading comprehension is even more difficult. Don't rely on analogies to familiar phenomena, when asked to obtain results do everything from the definition.

We start this class with some review of conditional probability. The main challenge in conditional probability questions is finding the correct partition for S.

Recall the conditional multiplication formula 5.3. We are given $P(A)$, $P(B)$, and one of $P(A \mid B)$ or $P(B \mid A)$. To solve for the other, we can use the following (simply divide by $P(B)$)

Theorem 6.1 (Bayes Theorem)**.**

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

§6.1 Independence

If A and B are any to events, to say that they are independent is that knowing that one happens does not effect the probability of the other. We could define independence as $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$, but they are undefined for $P(A), P(B) = 0$. To try to see how we can deal with this drawback, we look into the definition of conditional probability, where we find that we are dividing by $P(B)$.

$$\begin{aligned} P(A \mid B) = \frac{P(A \cup B)}{P(B)} &= P(A) \\ P(A \cup B) &= P(A)P(B) \end{aligned}$$

which is a much better definition, as it's more symmetric, easier to work with, and defined for more events.

Definition 6.2 (Independence)**.** We say 2 events $A, B \subseteq S$ are independent if

$$P(A \cup B) = P(A)P(B)$$

So far, independence is defined pairwise. What about triplets, or more? We cover that next class.

§7 Tutorial 2

The TA performed computations slightly differently, he defined the permutation number A_m^n as

$$A_m^n = \frac{m!}{(m-n)!} = P(m,n)$$

the combination number C_m^n is defined as

$$C_m^n = \frac{m!}{n!(m-n)!} = \frac{A_m^n}{n!} = C(m,n)$$

§8 Day 6: Continuity of Probabilities (Sep 22, 2025)

Midterm #1 is just 2 weeks away, Monday Oct 6 in this very room. Usually this is not the case. If you have a calculator, you would need to get it pre-approved 1 week before the midterm. It cannot be programmable, do algebra, solve equations, etc.

Last class, we looked at conditional probability, now we study independence. Pairs of events can be independent, but the intersection of them may not, as illustrated by the following example:

Example 8.1

Have $A = \{\text{first coin heads}\}$, $B = \{\text{second coin heads}\}$, and $C = \{\text{both coins flip the same}\}$.

Realize that pairwise, A, B, C are independent, but as a whole, knowing A, B lets you 'deduce' C, and having C and exactly one of A or B lets you deduce the other. For events A, B, C to be truly independent, we need $P(A \cap B \cap C) = P(A)P(B)P(C)$, motivating the following definition.

Definition 8.2 (Independence)**.** For a collection $\{A_i\}$ of events to be independent, for any finite subcollection of the events $\{i_k\}$,

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$$

Theorem 8.3. If $P(A), P(B) > 0$, then A and B cannot be both independent and disjoint.

Proof. TODO: prove this □

§8.1 Continuity of Probabilities

Recall for functions $f : \mathbb{R} \to \mathbb{R}$, continuity means if $\lim_{n\to\infty} x_n = x$, then $\lim_{n\to\infty} f(x_n) = f(x)$. We want an analogue for probabilities P, which motivates:

Definition 8.4 (Nested Increasing)**.** We say that $\{A_n\} \nearrow A$ if $\bigcup_n A_n = A$, with $A_n \subseteq A_{n+1}$ for all n.

As in $A_1 \subseteq A_2 \subseteq \cdots$. For example if $A_n = \{1, 2, \cdots, n\}$ then $\{A_n\} \nearrow \mathbb{N}$.

Theorem 8.5 (Continuity of Probabilities Theorem)

If $\{A_n\} \nearrow A$, then $\lim_{n\to\infty} P(A_n) = P(A)$.

The proof is similar to that of the subadditivity property of probability 2.7. However, since we know that $\{A_n\} \nearrow A$, meaning $A_1, \cdots, A_{n-2} \subseteq A_{n-1}$, we can take $B_n = A_n \cap (A_1 \cup \cdots \cup A_{n-1})^c = A_n \cap (A_{n-1})^c$ which is much less complicated.

Definition 8.6 (Nested Decreasing)**.** Write $\{A_n\} \searrow A$ if $\bigcap_n A_n = A$ and $A_n \supseteq A_{n+1}$ for all n.

If we only have probabilities defined on closed intervals, and wish to compute the probabilities of open intervals, we can either take the complement, or express an open interval as an arbitrary union of closed intervals. For example, to express (a, b) as a union of closed intervals, we can compute

$$\bigcup_{i=1}^{\infty} \left[a + \frac{1}{n}, b - \frac{1}{n}\right] = \lim_{n\to\infty} \bigcup_{i=1}^{n} \left[a + \frac{1}{n}, b - \frac{1}{n}\right] = (a, b)$$

This marks the end of chapter 1 in the textbook. Chapter 2 is much longer than chapter 1 though.

§8.2 Random Variables

Definition 8.7 (Random Variable)**.** A random variable is any function from S to $\mathbb{R}$.

A constant random variable, X, is a function such that for all $s \in S$, $X(s) = c$. Random variables can also be unbounded, where for every $M \in \mathbb{R}$, there exist $s \in S$ such that $X(s) > M$ or $X(s) < M$ or $|X(s)| > M$.

Definition 8.8 (Indicator Function)**.** $I_A(s) = 1$ if $s \in A$. Otherwise $I_A(s) = 0$.

§9 Day 7: Distributions (Sep 24, 2025)

§9.1 Distributions

Definition 9.1 (Distribution)**.** The distribution of a random variable is the collection of all of the probabilities of the variable being in every possible subset of $\mathbb{R}$.

We write $P(X \in B)$ to mean $P(X^{-1}(B)) := P\{s \in S : X(s) \in B\}$. We call $X^{-1}(B)$ the inverse image of B.

Definition 9.2 (Discrete Random Variables)**.** A random variable is called discrete if

$$\sum_{x \in \mathbb{R}} P(X = x) = 1$$

Also define the **probability function** as $p_X(x) := P(X = x)$

Definition 9.3 (Bernoulli (θ) Distribution)**.** $P(X = 1) = \theta, P(X = 0) = 1 - \theta$.

If X is a distribution with this property we write $X \sim$ Bernoulli(θ). Though for completeness on homework problems, you should say that otherwise for all other x, $p_X(x) = P(X = x) = 0$.

Definition 9.4 (Binomial(n, θ) Distribution)**.** $p_X(k) := P(X = k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$.

To prove that X sums to 1, recall the binomial theorem, which states that

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

which can be proved inductively. Substitute $a = \theta, b = (1-\theta)$, and $(\theta + (1-\theta))^n = 1$.

There exists a special case $\theta = \frac{1}{2}$, making $P(X = k) = \frac{\binom{n}{k}}{2^n}$. Bernoulli($\theta$) is the same as Binomial(1, θ). Suppose $X_1, \cdots, X_n \sim$ Bernoulli(θ) are independent random variables. We claim that the distribution of $Y = \sum_i X_i \sim$ Bernoulli(n, θ).

Picking independent at random with replacement, with binary outcomes is usually binomial?