Probability:
The Science of Uncertainty and Data

MITx 6.431x

2018/08/28 – 2018/12/23

Sumário

1 Welcome to 6.431x: Probability — The Science of Uncertainty and Data! (2018/08/28) 4

1.1 Welcome to 6.431x: Unit 0 released . . . 4
1.2 Quick info . . . 4
1.3 Updates . . . 5
1.3.1 2018/09/04: Unit 1 Released; Grading Policy; Discussion Forum Guidelines . . . 5
1.3.2 2018/09/07: Unit 1 due next Tuesday; No solutions on forum please . . . 6

2 Unit 0: Overview (2018/08/28) 7

2.1 Course overview . . . 7
2.1.1 Course character and objectives . . . 7
2.1.2 Why study probability? . . . 8
2.1.3 Course contents . . . 8
2.2 Course introduction, objectives and study guide . . . 8
2.2.1 Introduction . . . 8
2.2.2 Course objectives . . . 8
2.2.3 Study guide . . . 9
2.3 Syllabus, calendar, and grading policy . . . 10
2.3.1 Syllabus . . . 10
2.3.2 Calendar . . . 13
2.3.3 Gradingp policy . . . 14
2.4 edX Tutorial . . . 15
2.4.1 Basics . . . 15
2.4.2 Courseware navigation . . . 15
2.4.3 Top-level navigation . . . 15
2.4.4 Discussion forums . . . 15
2.4.5 Summary . . . 15
2.5 Discussion forum and collaboration guidelines . . . 15

2.5.1 Discussion forum guidelines . . . 15
2.5.2 Collaboration guidelines . . . 17
2.6 Homework mechanics and standard notation . . . 17
2.6.1 Checking and submitting an answer . . . 17
2.6.2 Standard notation . . . 18
2.7 Textbook information . . . 22
2.7.1 Textbook . . . 22
2.7.2 Ordering and other information . . . 22
2.8 Micromasters, Certification, and Honor Pledge . . . 22
2.8.1 Micromasters . . . 22
2.8.2 Certification . . . 24
2.8.3 EdX Honor Code Pledge . . . 24
2.9 Entrance survey . . . 25

3 Unit 1: Probability models and axioms (2018/09/03) 26
3.1 Lecture 1: probability models and axioms . . . 26
3.1.1 Motivation . . . 26
3.1.2 Overview and slides . . . 26
3.1.3 Sample space . . . 28
3.1.4 Exercise: Sample Space . . . 30
3.1.5 Sample space examples . . . 31
3.1.6 Exercise: Tree representations . . . 35
3.1.7 Probability axioms . . . 35
3.1.8 Exercises: axioms . . . 40
3.1.9 Simple properties of probabilities . . . 40
3.1.10 Exercise: Simple properties . . . 47
3.1.11 More properties of probabilities . . . 47
3.1.12 Exercise: More properties . . . 54
3.1.13 A discrete example . . . 54
3.1.14 Exercise: Discrete probability calculations . . . 58
3.1.15 A continuous example . . . 58
3.1.16 Exercise: Continuous probability calculations . . . 62
3.1.17 Countable additivity . . . 62
3.1.18 Exercise: Using countable additivity . . . 70
3.1.19 Exercise: Uniform probabilities on the integers . . . 70
3.1.20 Exercise: On countable additivity . . . 71
3.1.21 Interpretations and uses of probabilities . . . 71
3.2 Mathematical background: Sets; sequences, limits, and series; (un)countable sets . . . 74
3.2.1 Mathematical background overview . . . 74
3.2.2 Sets . . . 75
3.2.3 De Morgan's laws . . . 80
3.2.4 Sequences and their limits . . . 84
3.2.5 When does a sequence converge? . . . 88
3.2.6 Infinite series . . . 90
3.2.7 The geometric series . . . 91

3.2.8 About the order of summation in series with multiple indices . . . 94
3.2.9 Countable and uncountable sets . . . 103
3.2.10 Proof that the set of real numbers is uncountable . . . 106
3.3 Solved problems . . . 109
3.3.1 The probability of the difference of two events . . . 109
3.3.2 Geniuses and chocolates . . . 113
3.3.3 Uniform probabilities on a square . . . 116
3.3.4 Bonferroni's inequality . . . 120
3.4 Problem Set 1 . . . 126
3.4.1 Venn diagrams . . . 126
3.4.2 Set operations and probabilities . . . 130
3.4.3 Three tosses of a fair coin . . . 130
3.4.4 Parking lot problem . . . 130
3.4.5 Probabilities on a continuous sample space . . . 130
3.4.6 Upper and lower bounds on the probability of intersection 131

4 Unit 2: Conditioning and independence (2018/09/10) 132
4.1 Unit overview . . . 132
4.1.1 Motivation . . . 132
4.1.2 Unit 2 overview . . . 133
4.2 Lecture 2: Conditioning and Bayes' rule . . . 134
4.2.1 Overview . . . 134
4.2.2 Conditional probabilities . . . 135
4.2.3 Exercise: Conditional probabilities . . . 139
4.2.4 A die roll example . . . 140
4.2.5 Exercise: Conditional probabilities in a continuous model 142
4.2.6 Conditional probabilities obey the same axioms . . . 143

1 Welcome to 6.431x: Probability — The Science of Uncertainty and Data! (2018/08/28)

1.1 Welcome to 6.431x: Unit 0 released

The course site is now open. We have released Unit 0, which introduces the course and summarizes the objectives and what you can expect to learn. It also contains lots of important information that you should read over carefully. We have also released an Entrance Survey, and will appreciate your help in improving this course. Unit 1 will be released next Monday.

This is a graduate level version of 6.041x, which has been offered several times, and we are once more excited to offer this material. We hope that you will find this course an enriching educational experience, helping you to master the fundamental concepts and tools of probability theory and its applications.

This is a *challenging* class. It is exactly at the same level, breadth, and depth as the corresponding residential MIT offering. MIT students typically *spend about 12 hours a week* on this subject, and you can expect to need a similar time commitment, perhaps even a bit more, depending on your background. But even if you do not have the time to do everything, you may still gain a lot by following just parts of the course.

We look forward to seeing you in class! And tell your friends about it!

Best wishes, Prof. John Tsitsiklis, Eren Kizildag (TA), and your course team

1.2 Quick info

This is a full semester course on the basic tools of probabilistic modeling. This course covers:

- the general framework of probability models
- conditional probabilities, independence, and the Bayes rule
- multiple discrete or continuous random variables
- expectations, conditional expectations, variance
- various powerful tools of general applicability, including methods for calculating probabilities and expectations.
- laws of large numbers
- the main tools of Bayesian inference methods
- an introduction to random processes (Poisson processes and Markov chains)

The contents of this course are essentially the same as those of the corresponding MIT class, which has been offered and continuously refined over more than 50 years. It is a *challenging class*, but will enable you to apply the tools of probability theory to real-world applications or your research.

If you are new to the course, please review the following features, which are located in the menu that runs across the top of every page. If this is your

first edX course, you may also consider taking the edX Demo course first to explore the edX learning experience.

Course Info & Updates You are currently in this tab, where you will find links to useful handouts at the right side of the page, as well as any updates and announcements regarding the course.

Courseware This tab contains the main materials of the course, including lectures, solved problems, problem sets, and exams. When you return to this site, edX will remember where you left off, for your convenience.

Discussion All of the discussion forums throughout the course can be navigated through the Discussion tab, but we recommend that you use the in-page discussion boards that are specially created within each unit to help focus the discussions.

Progress This tab allows you to track your progress in the course by providing you with a summary of your score on each assignment as well as a useful plot of your overall score.

Resources In this tab you will find various useful resources, such as a standard normal table and the textbook excerpts.

1.3 Updates

1.3.1 2018/09/04: Unit 1 Released; Grading Policy; Discussion Forum Guidelines

Unit 1 Probabilistic Models and Axioms, has been released. We will dive into the heart of the subject right away and learn about the elements of a probabilistic model.

The lecture exercises and the first problem set are due in a week, on Tuesday Sept 11. While some parts may appear easy, make sure you understand every single detail well in order to do well as the course becomes more demanding. Please also note that because of the tight schedule, there will be no extensions to any of the deadlines in this course.

After careful considerations, we have also adjusted the grading policy to put more weight on the timed exams. Furthermore, the grade of one of the eleven problem sets will also be dropped. The updated grading policy is posted in Unit 0. If you have not already, please make sure you go over Unit 0 carefully, as it contains a lot of important information. We also appreciate your filling in the Entrance Survey.

Finally, we are very excited to see your engagement in the discussion forum already. Please continue to give your constructive comments and help

to each other and to us, but don't forget: DO NOT POST OR GIVE AWAY SOLUTIONS on the forum!

We are happy to have you with us.

Best wishes,

Prof John Tsitsiklis, Eren Kizildag (TA), Karene Chu, and the rest of your course team

1.3.2 2018/09/07: Unit 1 due next Tuesday; No solutions on forum please

We hope you are enjoying your journey in the world of uncertainty thus far!

The lecture exercises for Unit 1: Probability models and axioms and the associated Problem Set 1 are due on Tuesday, September 11. We want to emphasize once more that even though some parts of the problems might appear easy, we strongly recommend you have a look at the solutions and make sure you fully understood everything, as this unit serves as the fundamental background for the rest of the course. The solutions to the problem set will be available after the due date.

We are also delighted with the activity on the discussion forum, and we hope you'll continue the constructive discussion for the rest of the course! Please continue to give hints and ask questions to lead others towards more understanding, or just help in any creative way! On the other hand, please be careful not to post or give away answers. We know this is the beginning of the course and you may not be aware, and we will try our best to remove answers from posts, but we warn now that if a learner posts answers repeatedly, we might have to revoke their course access. Please be sure to review the discussion forum guidelines.

Finally, Unit 2: Conditioning and independence and the associated Problem Set 2 will be released next Monday, September 10; and the due date will be on Tuesday, September 18.

Best wishes,

Prof John Tsitsiklis, Eren Kizildag (TA), Dr. Karene Chu, and the rest of your course team

2 Unit 0: Overview (2018/08/28)

Unit 0 is the first unit available in the Courseware. It introduces the course and summarizes the objectives and what you can expect to learn. It also contains lots of important information that you should read over carefully.

Course overview; Course introduction, objectives, and study guide These sections introduce and overview the course and provide a guide for how to make the most of the wealth of materials that this course offers.

Syllabus, calendar, and grading policy Here you will find an outline of the units of this course, together with release and due dates. The same information is presented in a calendar format for your convenience. The grading policy is also explained in detail.

edX tutorial This sequence of videos gives a visual tutorial of how to use the basic elements of the edX platform.

Discussion forum and collaboration guidelines This section contains the course's guidelines for collaboration and using the dicussion forum. Please read them carefully and follow them throughout the course.

Homework mechanics and standard notation This section explains how to submit answers to problems and details the standard notation that should be used throughout the course when entering symbolic responses. Please read carefully and refer back to these documents when needed.

Textbook information This section describes how to access and navigate through the e-reader of excerpts from the course textbook. There is also information for purchasing a physical copy of the textbook as well as a link to textbook errata. While this textbook is recommended, the materials provided by this course are self-contained.

Micromasters, Certificates, and Honor Pledge This section provides information on how to earn a verified certificate for this course, as well as how to obtain the credential for the MITx Micromasters Program in Statistics and Data Science. You will also be asked to make a pledge to abide by the EdX Honor Code.

2.1 Course overview

2.1.1 Course character and objectives

Video: Course character and objectives (transcripts, slides)

2.1.2 Why study probability?

Video: Why study probability? (transcripts, slides)

2.1.3 Course contents

Video: Course contents (transcripts, slides)

2.2 Course introduction, objectives and study guide

2.2.1 Introduction

Welcome to 6.431x, an introduction to probabilistic models, including random processes and the basic elements of statistical inference.

The world is full of uncertainty: accidents, storms, unruly financial markets, noisy communications. The world is also full of data. Probabilistic modeling and the related field of statistical inference are the keys to analyzing data and making scientifically sound predictions.

The course covers all of the basic probability concepts, including:

- multiple discrete or continuous random variables, expectations, and conditional distributions
- laws of large numbers
- the main tools of Bayesian inference methods
- an introduction to random processes (Poisson processes and Markov chains)

2.2.2 Course objectives

Upon successful completion of this course, you will:

At a conceptual level:

- Master the basic concepts associated with *probability models*.
- Be able to translate models described in words to mathematical ones.
- Understand the main concepts and assumptions underlying *Bayesian and classical inference*.
- Obtain some familiarity with the range of *applications of inference methods*.

At a more technical level:

- Become familiar with basic and common *probability distributions*.
- Learn how to use *conditioning* to simplify the analysis of complicated models.
- Have facility manipulating *probability mass functions*, *densities*, and *expectations*.
- Develop a solid understanding of the concept of *conditional expectation* and its role in inference.

- Understand the power of *laws of large numbers* and be able to use them when appropriate.
- Become familiar with the basic inference methodologies (for both *estimation* and *hypothesis testing*) and be able to apply them.
- Acquire a good understanding of two *basic stochastic processes* (*Bernoulli* and *Poisson*) and their use in modeling.
- Learn how to formulate simple dynamical models as *Markov chains* and analyze them.

2.2.3 Study guide

This class provides you with a great wealth of material, perhaps more than you can fully digest. This "guide"offers some tips about how to use this material.

Start with the overview of a unit, when available. This will help you get an overview of what is to happen next. Similarly, at the end of a unit, watch the unit summary to consolidate your understanding of the "big picture"and of the relation between different concepts.

Watch the lecture videos. You may want to download the slides (clean or annotated) at the beginning of each lecture, especially if you cannot receive high-quality streaming video. Some of the lecture clips proceed at a moderate speed. Whenever you feel comfortable, you may want to speed up the video and run it faster, at 1.5x.

Do the exercises! The exercises that follow most of the lecture clips are a most critical part of this class. Some of the exercises are simple adaptations of you may have just heard. Other exercises will require more thought. Do your best to solve them right after each clip — do not defer this for later — so that you can consolidate your understanding. After your attempt, whether successful or not, do look at the solutions, which you will be able to see as soon as you submit your own answers.

Solved problems and additional materials. In most of the units, we are providing you with many problems that are solved by members of our staff. We provide both video clips and written solutions. Depending on your learning style, you may pick and choose which format to focus on. But in either case, it is important that you get exposed to a large number of problems.

The textbook. If you have access to the textbook, you can find more precise statements of what was discussed in lecture, additional facts, as well as several examples. While the textbook is recommended, the materials provided by this course are self-contained. See the "Textbook information"tab in Unit 0 for more details.

Problem sets. One can really master the subject only by solving problems — a large number of them. Some of the problems will be straightforward applications of what you have learned. A few of them will be more challenging. Do not despair if you cannot solve a problem – no one is expected to do everything perfectly. However, once the problem set solutions are released (which will happen on the due date of the problem set), make sure to go over the solutions to those problems that you could not solve correctly.

Exams. The midterm exams are designed so that in an on-campus version, learners would be given two hours. The final exam is designed so that in an on-campus version, learners would be given three hours. You should not expect to spend much more than this amount of time on them. In this respect, those weeks that have exams (and no problem sets!) will not have higher demands on your time. The level of difficulty of exam questions will be somewhere between the lecture exercises and homework problems.

Time management. The corresponding on-campus class is designed so that students with appropriate prerequisites spend about 12 hours each week on lectures, recitations, readings, and homework. You should expect a comparable effort, or more if you need to catch up on background material. In a typical week, there will be 2 hours of lecture clips, but it might take you 4–5 hours when you add the time spent on exercises. Plan to spend another 3–4 hours watching solved problems and additional materials, and on textbook readings. Finally, expect about 4 hours spent on the weekly problem sets.

Additional practice problems. For those of you who wish to dive even deeper into the subject, you can find a good collection of problems at the end of each chapter of the print edition of the book, whose solutions are available online.

2.3 Syllabus, calendar, and grading policy

2.3.1 Syllabus

6.431x Fall 2018 Syllabus

- Unit 0: Overview (released Tue. August 28)
- Unit 1: Probability models and axioms (released Mon. Sep 3; Sections 1.1–1.2)
  - L1: Probability models and axioms
  - Problem Set 1 due on Tue Sept 11
- Unit 2: Conditioning and independence (released Mon. Sept 10; Sections 1.3–1.5)
  - L2: Conditioning and Bayes' rule

  - L3: Independence
  - Problem Set 2 due on Tue Sept 18
- Unit 3: Counting (released Mon. Sept 17; Section 1.6)
  - L4: Counting
  - Problem Set 3 due on Tue Sept 25
- Unit 4: Discrete random variables (released Wed. Sept 19; Sections 2.1–2.7)
  - L5: Probability mass functions and expectations
  - L6: Variance; Conditioning on an event; Multiple r.v.'s
  - L7: Conditioning on a random variable; Independence of r.v.'s
  - Problem Set 4 due on Tue Oct 2
- Exam 1 (Timed) : Covers material from L1 to L7 (released Wed. Oct 3; due on Tue. Oct 9)
- Unit 5: Continuous random variables (released Mon. Oct 1; Sections 3.1–3.5)
  - L8: Probability density functions
  - L9: Conditioning on an event; Multiple r.v.'s
  - L10: Conditioning on a random variable; Independence; Bayes' rule
  - Problem Set 5 due on Tue. Oct 16
- Unit 6: Further topics on random variables (released Mon. Oct 15; Sections 4.1–4.3, 4.5)
  - L11: Derived distributions
  - L12: Sums of r.v.'s; Covariance and correlation
  - L13: Conditional expectation and variance revisited; Sum of a random number of r.v.'s
  - Problem Set 6 due on Tue. Oct 23
- Unit 7: Bayesian inference (released Mon. Oct 22 Sections 3.6, 8.1–8.4)
  - L14: Introduction to Bayesian inference
  - L15: Linear models with normal noise
  - L16: Least mean squares (LMS) estimation
  - L17: Linear least mean squares (LLMS) estimation
  - Problem Set 7a due on Tue. Oct 30
  - Problem Set 7b due on Tue. Nov 6
- Exam 2 (Timed): Covers material from L8 to L17 (released Wed. Nov 1; due on Nov 13)
- Unit 8: Limit theorems and classical statistics (released Mon. Nov 5; Sections 5.1–5.4, pp. 466–475)
  - L18: Inequalities, convergence, and the Weak Law of Large Numbers
  - L19: The Central Limit Theorem (CLT)
  - L20: An introduction to classical statistics
  - Problem Set 8 due on Tue. Nov 27

- Unit 9: Bernoulli and Poisson processes (released Tue. Nov 14; Sections 6.1-6-2)
  - L21: The Bernoulli process
  - L22: The Poisson process
  - L23: More on the Poisson process
  - Problem Set 9 due on Tue. Dec 4
- Unit 10: Markov chains (released Tue. Nov 26; Sections 7.1–7.4)
  - L24: Finite-state Markov chains
  - L25: Steady-state behavior of Markov chains
  - L26: Absorption probabilities and expected time to absorption
  - Problem Set 10 due on Tue. Dec 11
- Final Exam (Timed) (released Wed. Dec 12; due on Sun. Dec 23)

Note: Problem set and exam due dates are at the end of the specified date, at 23:59 UTC.

2.3.2 Calendar

6.431x Fall 2018 Calendar

MONDAY	TUESDAY	WEDNESDAY
9/3 Unit 1 released: Probability models and axioms (Secs. 1.1-1.2)	**9/4**	**9/5**
9/10 Unit 2 released: Conditioning and independence (Secs. 1.3-1.5)	**9/11** Problem Set 1 due	**9/12**
9/17 Unit 3 released: Counting (Sec. 1.6)	**9/18** Problem Set 2 due	**9/19** Unit 4 released: Discrete r.v.'s (Ch. 2)
9/24	**9/25** Problem Set 3 due	**9/26**
10/1 Unit 5 released: Continuous r.v.'s (Secs. 3.1-3.5)	**10/2** Problem Set 4 due	**10/3** Exam 1 (Timed) released
10/8	**10/9** Exam 1 (Timed) due	**10/10**
10/15 Unit 6 released: Further topics on r.v.'s (Secs. 4.1-4.3, 4.5)	**10/16** Problem Set 5 due	**10/17**
10/22 Unit 7 released: Bayesian inference (Secs. 3.6, 8.1-8.4)	**10/23** Problem Set 6 due	**10/24**
10/29	**10/30** Problem Set 7a due	**10/31**
11/5 Unit 8 released: Limit theorems and classical statistics (Secs. 5.1-5.4, pp. 466-475)	**11/6** Problem Set 7b due	**11/7** Exam 2 (Timed) released
11/12	**11/13** Exam 2 (Timed) due	**11/14** Unit 9 released: Bernoulli and Poisson processes (Secs. 6.1-6.-2)
11/19	**11/20**	**11/21**
11/26 Unit 10 released: Markov chains (Secs. 7.1-7.4)	**11/27** Problem Set 8 due	**11/28**
12/3	**12/4** Problem Set 9 due	**12/5**
12/10	**12/11** Problem Set 10 due Final Exam (Timed) released	**12/12**
12/17	**12/18**	**12/19** 12/20 Final Exam (Timed) due

Notes:

- The due dates for the weekly problem sets and the exams are fixed and cannot be changed or modified for any individuals. Please plan accordingly.
- Problem set and exam due dates are at the end of the specified date, at 23:59 UTC.
- The calendar above shows only Tuesdays, Wednesdays, and Thursdays, since these are the only days of the week when materials will be released or due, except the final exam.

2.3.3 Gradingp policy

Grading policy Your overall score in this class will be a weighted average of your scores for the different components, with the following weights:

- 20% for the lecture exercises (divided equally among the 26 lectures)
- 20% for the problem sets (divided equally among 11 problem sets)
- 18% for the first midterm exam (timed)
- 18% for the second midterm exam (timed)
- 24% for the final exam (timed)

To earn a verified certificate for this course, you will need to obtain an *overall score* of 60% or more of the maximum possible overall score.

Note that not every problem set or set of lecture exercises will have the same number of raw points. For example, Problem Set 1 may have 30 points and Problem Set 2 may have 35 points. However, each one receives the same weight for the purposes of calculating your overall score.

As an illustrative example, if you receive 20 points out of 30 on Problem Set 1, this will contribute $\frac{20}{30} \times \frac{20\%}{11} = 1.21\%$ to your overall score. Similarly, if you receive 30 points out of 35 on Problem Set 2, this will contribute $\frac{30}{35} \times \frac{20\%}{11} = 1.56\%$ to your overall score.

Under the “Progress"tab at the top, you can see your score broken down for each assignment, as well as a summary plot.

Timed Exams The 2 midterm exams and one final exam are *timed exams*. This means that each exam is available for approximately a week, but once you open the exam, there is a limited amount of time (48 hours), counting from when you start, within which you must complete the exam. Please plan in advance for the exams. If you do not complete the whole exam during the allowed time, you will miss the points associated with the questions that have not been answered. The exams are designed to assess your knowledge. There are no extensions granted to these deadlines. You can find the exam dates on the calendar on the previous page. Note that the timed exams cannot be completed using the edX mobile app.

MITx Committment to Accessibility If you have a disability-related request regarding accessing an MITx course, including exams, please contact the course team as early in the course as possible (at least 2 weeks in advance of exams opening) to allow us time to respond in advance of course deadlines. Requests are reviewed via an interactive process to meet accessibility requirements for learners with disabilities and uphold the academic integrity for MITx.

2.4 edX Tutorial

2.4.1 Basics

Video: edX Basics

2.4.2 Courseware navigation

Video: Courseware navigation

2.4.3 Top-level navigation

Video: Top-level navigation

2.4.4 Discussion forums

Video: Discussion forums

2.4.5 Summary

Video: Summary

2.5 Discussion forum and collaboration guidelines

2.5.1 Discussion forum guidelines

Discussion forum overview The course provides an online discussion forum for you to communicate with the course team and other learners. You may access the forum through the “Discussion"tab at the top of the page, as well as through many embedded discussions within each unit. We recommend using the embedded discussions within each unit to discuss topics related to a specific unit’s materials, whether it’s lectures, solved problems, or problem set problems. Please see the guidelines below for more information on how to use these embedded discussions.

For other more general discussions, you may use the “Discussion"tab at the top of the page. When creating a new post, *please choose one of the following categories that best describes your post*:

- *Introductions*: Introduce yourself to your fellow learners and find out more about them!

- *Micromasters*: Ask questions related to the MITx Micromaster Program in Statistics and Data Science and meet other Micromasters fellows!
- *Course Feedback*: Let the course team know how you are finding the course, what you think works well, and what you would like to see improved.
- *Technical Problems*: Let the course team know about any technical issues you are dealing with (e.g., playing videos, entering answers, etc).
- *General*: Other general discussions.

Discussion forum guidelines The discussion forum is the main way for you to communicate with the course team and other learners. We hope it contributes to a sense of community and serves as a useful resource for your learning. Here are some guidelines to help you successfully navigate and interact on the forum:

- *Use discussion while working through the material*. Beginning with Unit 1, each lecture will contain an embedded discussion located at the bottom of the lecture overview clip, which is the first or second clip of that lecture sequence. You should discuss anything related to that lecture's video clips or exercises there. Click "Show Discussion"to see all discussions associated with the lecture, and click "Add a Post"to post a new topic. In addition, every solved problem and problem set problem will have its own embedded discussion located at the bottom of their respective pages. As with the lecture discussions, click "Show Discussion"and "Add a Post"to see and create discussion topics related to that specific problem. We recommend that you use these in-page discussion boards to help focus discussions on specific topics.
- *Use informative topic titles and tags*. To make it easier to identify relevant discussion topics, please use informative titles and tags when creating a new discussion topic. We suggest using titles or tags that are as informative as possible, e.g., "lecture X, exercise Y on topic W, clarify part Z"
- *Be very specific*. Provide as much information as possible about what you need help for: Which part of what problem or video? Why do you not understand the question? Do you need help understanding a particular concept? What have you tried doing so far? Use a descriptive title to your post. This will attract the attention of other learners having the same issue.
- *Observe the honor code*. We encourage collaboration and help, but please do not ask for nor post problem solutions.
- *Upvote good posts*. This applies to questions and answers. Click on the green plus button so that good posts can be found more easily.
- *Search before asking*. The forum can become hard to use if there are too many threads, and good discussions happen when people participate in the same thread. Before asking a question, use the search feature by clicking on the magnifying glass on the left-hand side.

- *Write clearly*. We know that English is a second language for many of you but correct grammar will help others to respond. Avoid ALL CAPS, abbrv of wrds (abbreviating words), and excessive punctuation!!!!

Please Introduce Yourself! Let's get started by introducing yourselves on the discussion forum. A lot of the learning in this class will happen in your interactions with each other. Click on the post titled "Introduce yourself!"below, and respond to it by telling everyone your name, where you are from, why you are taking this course, and whatever else you would like to share! Your post will be indexed in the "Introductions"category in the forum.

2.5.2 Collaboration guidelines

We encourage you to interact with your fellow learners and engage in active discussion about the course. Please use the guidelines below for acceptable collaboration. The staff will be proactive in removing posts and replies in the discussion forum that have stepped over the line.

- Given a problem, it is ok to discuss the general approach to solving the problem.
- You can work jointly to come up with the general steps for the solution.
- It is ok to get a hint, or several hints for that matter, if you get stuck while solving a problem.
- You should work out the details of the solution yourself.
- It is not ok to take someone else's solution and simply copy the answers from their solution into your checkboxes.
- It is not ok to take someone else's formula and plug in your own numbers to get the final answer.
- It is not ok to post answers to homework and lab problems before the submission deadline.
- It is not ok to look at a full step-by-step solution to a problem before the submission deadline.
- It is ok to have someone show you a few steps of a solution where you have been stuck for a while, provided of course, you have attempted to solve it yourself without success.
- After you have collaborated with others in generating a correct solution, a good test to see if you were engaged in acceptable collaboration is to make sure that you are able to do the problem on your own.

2.6 Homework mechanics and standard notation

2.6.1 Checking and submitting an answer

Checking and submitting an answer For each problem, you will have between 2 to 5 attempts to submit an answer, with the exception of problems where an

attempt essentially reveals the answer (e.g., True/False questions), for which you will be limited to a single attempt.

To submit your answer, click the “Submit "button. This will automatically submit the problem for grading purposes, and the edX platform is able to verify your answer and give you immediate feedback as to whether or not your answer is correct. To save your answer without submitting it for grading purposes, click the “Save"button. Your answer will be restored when you return to the problem.

The number of attempts allowed as well as the number of attempts you’ve already made will always be visible on a problem’s page at the bottom, next to the “Check"button. Please note that for problems consisting of multiple parts, hitting the button will count as an attempt for all parts of the problem. Unfortunately, it is not possible to submit answers for one part at a time.

For lecture exercises, a “Show Answer(s)"button will appear immediately after you submit the correct answer or use all of your attempts. Clicking this button will reveal the correct answers and solutions.

For homework problems, the “Show Answer(s)"button will appear after the due date of the homework.

You are strongly encouraged to look at the solutions even if your answer is correct.

Answer formats This course will use several answer formats:

- Multiple choice: Select the correct option from the dropdown menu or radio buttons.
- Numerical answers: Enter a number, either in decimal (e.g., ’3.14159’) or fractional form (e.g., ’22/7’). Do not enter any non-numerical letters or symbols. To account for rounding, the system will accept a range of answers as correct. Unless otherwise specified in the problem, the default tolerance range will be +/-3% of the correct answer.
- Symbolic answers: Some problems will ask for a symbolic answer (e.g., ’n*(n+1)/2’). See the next section on “Standard notation"for details on how to submit such answers.

Below are some example problems for you to familiarize yourselves with how these problem types work with different number of attempts. These problems are not graded and have no impact on your grade.

2.6.2 Standard notation

Many exercises and problems throughout the course will ask you to provide an algebraic answer in terms of symbols. Please follow the guidelines below when entering your responses. Below your answer textbox, the system will also display, in a "pretty"format, what it has interpreted your input to be. However, this display is not perfect (for example, it does not catch all cases of missing close parentheses) so please also check your text input carefully.

- Symbols are case-sensitive: a and A are different — make sure to use the correct case as specified in the problem
- Parentheses: make sure that your parentheses are properly balanced — each open parenthesis should have a matching close parenthesis!
- Elementary arithmetic operations: use the symbols + , - , * , / for addition, subtraction, multiplication, and division, respectively
- $1 + bc - d/e$ should be entered as 1+b*c-d/e
- For multiplication, use * explicitly:
  - in the example above, enter b*c; do NOT enter bc
  - for $2n(n+1)$, enter 2*n*(n+1); do NOT enter 2n(n+1)
  - although the "pretty"display underneath your answer looks correct if you do not include * s, your answer will be marked incorrect!
- Exponents: use the symbol ^ to denote exponentiation
  - 2^n should be entered as 2^n
  - x^{n+1} should be entered as x^(n+1)
- Square root: use the string of letters sqrt , followed by enclosing what is under the square root in parentheses
  - $\sqrt{-1}$ should be entered as sqrt(-1)
- Mathematical constants: use the symbol e for the base of the natural logarithm, e; use the string of letters pi for π
  - $e^{i\pi} + 1$ should be entered as e^(i*(pi))+1
- Order of operations: 1) parentheses, 2) exponents and roots, 3) multiplication and division, 4) addition and subtraction
  - $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ should be entered as (1/sqrt(2*(pi)))*e^(-(x^2)/2)
  - a/b*c is interpreted as $\frac{a}{b}c$
  - enter a/(b*c) for $\frac{a}{bc}$
  - When in doubt, use additional parentheses to remove possible ambiguitites
- Natural logarithm: although in lectures and solved problems we will sometimes use the notation "log"(instead of "ln"), you should use the string of letters ln , followed by the argument enclosed in parentheses
  - $\ln(2x)$ should be entered as ln(2*x)
- Trigonometric functions: use the usual 3-letter symbols to denote the standard trigonometric functions
  - $\sin(x)$ should be entered as sin(x)
- Greek letters: use the Latin-character name to denote each Greek letter
  - $\lambda e^{-\lambda t}$ should be entered as lambda*e^(-lambda*t)
  - $\mu\alpha\theta$ should be entered as mu*alpha*theta

- Factorials, permutations, combinations: you will not need enter these for any symbolic answers; do NOT use ! in your answers as it will not be evaluated correctly!

Figura 1: Standard Notation Summary: 1

Symbols	These are case sensitive. Use the correct case as specified in the problem.	**n** and **N** are different.	Do NOT enter x for X
Parentheses	Match each open parenthesis with a close parenthesis.		
Elementary Arithmetic Operations	Use the symbols **+**,**-**,*****,**/** for addition, subtraction, multiplication, and division, respectively.	Enter **1+b*c-d/e** for $1 + bc - d/e$	
	For multiplication, use ***** explicitly. Although the "pretty" display underneath your answer looks correct if you do not include ***** s, your answer will be marked incorrect!	Enter **b*c** for bc in the example above Enter **2*n*(n+1)** for $2n(n+1)$	Do NOT enter bc for bc Do NOT enter 2n(n+1) for $2n(n+1)$
Exponents	Use the symbol **^** to denote exponentiation.	Enter **2^n** for 2^n Enter **x^(n+1)** for x^{n+1}	
Square Root	use the string of letters **sqrt**, followed by enclosing what is under the square root in parentheses.	Enter **sqrt(-1)** for $\sqrt{-1}$	

Figura 2: Standard Notation Summary: 2

Mathematical Constants	Use the symbol **e** for the base of the natural logarithm, e. Use the string of letters **pi** for π.	Enter **e^(i*(pi))+1** for $e^{i\pi}+1$	
Order Of Operations	1) parentheses 2) exponents and roots 3) multiplication and division 4) addition and subtraction When in doubt, use additional parentheses to remove possible ambiguitites.	Enter **a/b*c** for $\frac{a}{b}\cdot c$ Enter **a/(b*c)** for $\frac{a}{bc}$ Enter **(1/sqrt(2*(pi)))*e^(-(x^2)/2)** for $\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$	Do NOT enter a/b*c for $\frac{a}{bc}$
Natural Logarithm	Although in lectures and solved problems we will sometimes use the notation "log" (instead of "ln"), you should use the string of letters **ln**, followed by the argument enclosed in parentheses.	Enter **ln(2*x)** for $\ln(2x)$	Do NOT enter log(2*x) for $\ln(2x)$
Trigonometric Functions	Use the usual 3-letter symbols to denote the standard trigonometric functions	Enter **sin(x)** for $\sin(x)$	Do NOT enter sin x for $\sin(x)$
Greek Letters	Use the Latin-character name to denote each Greek letter	Enter **lambda*e^(-lambda*t)** for $\lambda e^{-\lambda t}$ Enter **mu*alpha*theta** for $mu\alpha\theta$	
Factorials, Permutations, Combinations	You will not need enter these for any symbolic answers.		Do NOT use ! in your answers as it will not be evaluated correctly!

2.7 Textbook information

2.7.1 Textbook

The class follows closely the text Introduction to Probability, 2nd edition, by Bertsekas and Tsitsiklis, Athena Scientific, 2008; see the publisher's website or Amazon.com for more information.

While this textbook is recommended, the materials provided by this course are self-contained. Furthermore, the publisher has made available, for the purposes of this class, the summary tables that are included in the text. These can be found under the "Resources"tab, or directly by following this link. In various places within the courseware, there will also be links to specific sections and pages to the excerpts from the textbook relevant to the material at hand. These links will also take you to the e-reader, jumping directly to the specific sections and pages.

To adjust the zoom level in the e-reader, click the '+' and '-' buttons at the top-right to zoom in and out, respectively. Or, choose a specific zoom level using the drop-down menu. Depending on your operating system and web browser, you may encounter occasional artifacts or imperfect rendering of some formulas. Please try adjusting the zoom level to find the one that gives the best readability. We recommend using Firefox as it renders the text most accurately.

2.7.2 Ordering and other information

The class follows closely the text Introduction to Probability, 2nd Ed., by Bertsekas and Tsitsiklis. If interested in purchasing a copy of the textbook, it is available through Amazon.

Textbook errata can be found here. Ignore "Corrections to the 1st and 2nd printing."These do not apply to the currently available printed version.

2.8 Micromasters, Certification, and Honor Pledge

2.8.1 Micromasters

Video: MITs Micromasters in Statistics and Data Science

This course is part of the MITx Micromaster Program in Statistics and Data Science. Welcome to the program!

About the Program The MITx Micromasters program in Statistics and Data Science is comprised of four EdX courses and a virtually proctored exam that will provide you with the foundational knowledge essential to understanding the methods and tools used in data science, and hands-on training in data analysis and machine learning. You will dive into the fundamentals of probability and statistics, as well as learn, implement, and experiment with data analysis techniques and machine learning algorithms. This program will prepare you to become an informed and effective practitioner of data science who

adds value to an organization and will also accelerate your path towards an MIT PhD or a Master's at other universities.

Anyone can enroll in the Micromasters program just as in any EdX courses. It is designed for learners who want to acquire sophisticated and rigorous training in data science without leaving their day job but also without compromising quality. There is no application process. To excel in the entire program, make sure you learn the foundational material covered in this course. You will also need some knowledge of matrices and proficiency in Python programming.

What You'll Learn You'll lean about:

- Master the foundations of data science, statistics, and machine learning
- Analyze big data and make data-driven predictions through probabilistic modeling and statistical inference; identify and deploy appropriate modeling and methodologies in order to extract meaningful information for decision making
- Develop and build machine learning algorithms to extract meaningful information from seemingly unstructured data; learn popular unsupervised learning methods, including clustering methodologies and supervised methods such as deep neural networks
- Master techniques in modern data analysis to leverage big datasets; use python and R skillfully to analyze data

How to earn the Micromasters credential To earn the MITx Micromasters credential in statistics and data science, you must successfully pass and receive a Verified Certificate in each of the 4 courses listed below and pass the final Capstone Exam:

- 6.431x Probability–the Science of Uncertainty and Data
- 14.310Fx Data Analysis in Social Sciences
- 18.6501x Fundamentals of Statistics
- 6.86x Machine Learning with Python–From Linear Models to Deep Learning
- DS-CFx Captone Exam in Statistics and Data Science

All the courses are taught by MIT faculty at a similar pace and level of rigor as an on-campus course at MIT.

More information If you are interested in the Micromasters program, visit https://www.edx.org/micromasters/mitx-statistics-and-data-science. For more detail on this program and credit pathways, please visit the MITx Micromasters Portal, which includes a "Contact us"link at the very bottom left. You may also find the FAQ helpful. Finally, you can start connecting with fellow Micromasters learners on the discussion forum!

2.8.2 Certification

To earn a Verified Certificate in this course, you need to:

- Upgrade your status to be a verified learner - The fee is $300.
- Verify your identity - ID Verification.
- Pass the course - at least 60% on your final grade.

You have limited time to switch to a Verified Certificate learner – you should get ID Verified as soon as you register as a Verified learner. See the EdX FAQ for more details on certificates.

Note: It is your responsibility to make sure that your ID verification is valid during the whole course.

A verified certificate indicates that you have successfully completed the course, but will not include a specific grade. Certificates are issued by edX under the name of MITx and are delivered online through your dashboard on edx.org.

2.8.3 EdX Honor Code Pledge

By enrolling in an EdX course, you have already agreed with the EdX Honor Code, which means that you will do the following:

- Complete all graded material (graded assignments and exams) with your own work and only your own work. You will not submit the work of any other person or have anyone else submit work under your name.
- Maintain only one user account and not let anyone else use your username and/or password. Having two user accounts registered in this course will constitute cheating. Not engage in any activity that would dishonestly improve your results, or improve or hurt the results of others."."Not collaborate with anyone other than staff on the exam questions. This means comparing answers, working as teams, or sharing answers in any way.
- Not post answers to any problems that are used to assess learner performance.
- Always be polite and respectful when communicating across the platform (with other learners and the staff).

We will strictly enforce this honor code pledge. Learners found violating this pledge will be dealt with directly. If we become aware of any suspicious activity we reserve the right to remove credit, not award a certificate, revoke a certificate, ban from this and other courses in the MITs Micromasters Program in Statistics and Data Science as well as notify edX for other actions. We take academic honesty very, very seriously at MIT. With the introduction of the Micromasters Credential, the importance of honesty in work has been elevated to a much higher level than before. We will diligently monitor this and be very proactive.

2.9 Entrance survey

For us to offer the best course experience possible, we'd like to ask you to answer a few questions about yourself: Entrance Survey.

3 Unit 1: Probability models and axioms (2018/09/03)

3.1 Lecture 1: probability models and axioms

Atention: Exercises due Sep 11, 2018 20:59:59 -03.

3.1.1 Motivation

Video: Motivation

Let's face it. Life is uncertain. But one thing is certain. We need a way to make predictions and make decisions under uncertainty. Probabilistic models can help you answer questions, such as:

- What are the odds that there will be a long line at the supermarket checkout counter?
- How likely is it that my GPS device is off by more than 10 meters?
- What are the odds that I will have a car accident next year?
- How likely is it that the air traffic control radar will miss the approaching plane?
- Should I invest in the stock market now or wait?
- Can I use a probabilistic model of social networking data to create a marketing campaign?
- How do we use a statistical model to decide if a medical treatment is effective?
- How do I model the huge amounts of data that are now becoming available in so many different fields, big data, as they call it, and extract useful information?

I am *John Tsitsiklis*. And I'm *Patrick Jaillet*. Our mission in this class is to give you the tools to model and analyze uncertain situations no matter what your discipline.

To do that, we will use the language and precision of mathematics, but we will also build your intuition. This is an ambitious class. The online version is at *the same level as the one offered to MIT students*.

It covers a lot of material. Beyond the basics, you will learn about random processes and about extracting information from data. In the end, you will be able to make much better sense of the uncertainty around you. The rewards are certain to come.

Let's face it. Life is uncertain.

3.1.2 Overview and slides

This lecture sequence introduces the basic structure of probability models, including the sample space and the axioms that any probabilistic model should obey, together with some consequences of the axioms and some simple examples.

Video: Lecture 01: Probability Models and Acioms (transcripts, slides, annotated slides)[1] [2] [3]

Welcome to the first lecture of this class. You may be used to having a first lecture devoted to general comments and motivating examples. This one will be different.

We will dive into the heart of the subject right away. In fact, today we will accomplish a lot. By the end of this lecture, you will *know about all of the elements of a probabilistic model*.

A probabilistic model is a quantitative description of a situation, a phenomenon, or an experiment whose outcome is uncertain. Putting together such a model involves two key steps:

- First, we need to describe the possible outcomes of the experiment. This is done by specifying a so-called *sample space*.

- Second, we specify a *probability law*, which assigns probabilities to outcomes or to collections of outcomes.

The probability law tells us, for example, whether one outcome is much more likely than some other outcome.

Probabilities have to satisfy certain basic properties in order to be meaningful. These are the *axioms of probability theory*. For example probabilities cannot be negative. Interestingly, there will be very few axioms, but they are powerful, and we will see that they have lots of consequences. We will see that they imply many other properties that were not part of the axioms.

We will then go through a couple of very simple examples involving models with either *discrete* or *continuous* outcomes. As you will be seeing many times in this class, discrete models are conceptually much easier. Continuous models involve some more sophisticated concepts, and we will point out some of the subtle issues that arise. And finally, we will talk a little bit about the big picture, about the role of probability theory, and its relation with the real world.

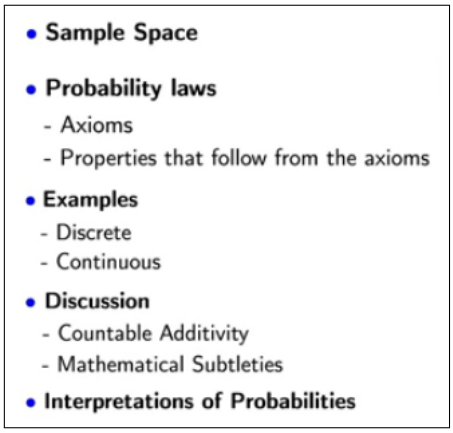
[1] The same material, in live lecture hall format, can be found here and here.

[2] You can also take this occasion to review some concepts related to sets (especially De Morgan's laws), sequences, and infinite series, by watching the "Mathematical background"sequence of clips.

[3] More information is given in the text:

- Sets: Section 1.1
- Probabilistic models: Section 1.2

Figura 3: Objectives of Lecture 1



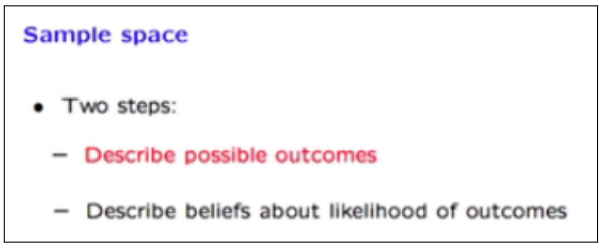
3.1.3 Sample space

Video: Sample space (transcript)

Putting together a probabilistic model — that is, a model of a random phenomenon or a random experiment — involves two steps.

- First step, we describe the *possible outcomes* of the phenomenon or experiment of interest.
- Second step, we describe our beliefs about the *likelihood of the different possible outcomes* by specifying a *probability law*.

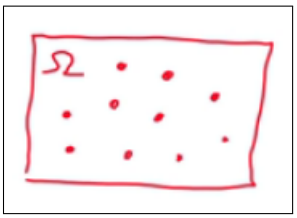
Here, we start by just talking about the first step, namely, the description of the possible outcomes of the experiment.



So we carry out an experiment. For example, we flip a coin. Or maybe we flip five coins simultaneously. Or maybe we roll a die.

Whatever that experiment is, it has a number of possible outcomes, and we start by *making a list of the possible outcomes* — or, a better word, instead of the word "list", is to use the word "set", which has a more formal mathematical meaning.

So we create a set that we usually denote by capital omega, Ω. That set is called the *sample space* and is the set of **all possible outcomes of our experiment**:



The elements of that set should have certain properties. Namely, the elements should be **mutually exclusive** and **collectively exhaustive**.

- List (set) of possible outcomes, Ω
- List must be:
  - Mutually exclusive
  - Collectively exhaustive
  - At the "right" granularity

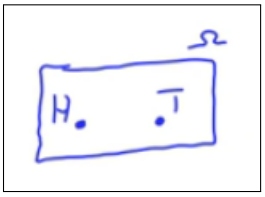
What does that mean? Mutually exclusive means that, if at the end of the experiment, I tell you that this outcome happened, then it should not be possible that this outcome also happened. At the end of the experiment, there can *only be one of the outcomes that has happened*.

Being collectively exhaustive means something else — that, together, all of these *elements of the set exhaust all the possibilities*. So no matter what, at the end, you will be able to point to one of the outcomes and say, that's the one that occurred.

To summarize: this set should be such that, at the end of the experiment, you should be always able to *point to one, and exactly one, of the possible outcomes* and say that this is the outcome that occurred.

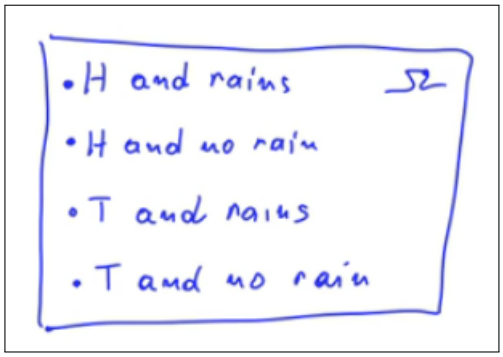
Physically *different outcomes should be distinguished in the sample space* and correspond to distinct points. But when we say physically different outcomes, what do we mean? We really mean *different in all relevant aspects* but perhaps not different in irrelevant aspects. Let's make more precise what I mean by that by looking at a very simple, and maybe silly, example, which is the following.

Suppose that you flip a coin and you see whether it resulted in heads or tails. So you have a perfectly legitimate sample space for this experiment which consists of just two points — heads and tails:



Together these two outcomes *exhaust all possibilities*. And the two outcomes are *mutually exclusive*. So this is a very legitimate sample space for this experiment.

Now suppose that while you were flipping the coin, you also looked outside the window to check the weather. And then you could say that my sample space is really, heads, and it's raining. Another possible outcome is heads and no rain. Another possible outcome is tails, and it's raining, and, finally, another possible outcome is tails and no rain.



This set, consisting of four elements, is also a perfectly legitimate sample space for the experiment of flipping a coin. The elements of this sample space are mutually exclusive and collectively exhaustive. Exactly one of these outcomes is going to be true, or will have materialized, at the end of the experiment.

So which sample space is the correct one? This sample space, the second one, involves some *irrelevant details*. So the preferred sample space for describing the flipping of a coin, the preferred sample space is the simpler one, the first one, which is sort of at the *right granularity*, **given what we're interested in**.

But ultimately, the question of which one is *the right sample space depends on what kind of questions you want to answer*. For example, if you have a theory that the weather affects the behavior of coins, then, in order to play with that theory, or maybe check it out, and so on, then, in such a case, you might want to work with the second sample space.

This is a common feature in all of science. Whenever you put together a model, you need to decide how detailed you want your model to be. And *the right level of detail is the one that captures those aspects that are relevant and of interest to you*.

3.1.4 Exercise: Sample Space

Exercise 3.1.4-1: Sample space
For the experiment of flipping a coin, and for each one of the following choices, determine whether we have a legitimate sample space:

$\Omega = \{\text{Heads and it is raining, Heads and it is not raining, Tails}\}$

◯ Yes
◯ No

$\Omega = \{\text{Heads and it is raining, Tails and it is not raining, Tails}\}$

◯ Yes
◯ No

3.1.5 Sample space examples

Video: Sample space examples (transcripts)

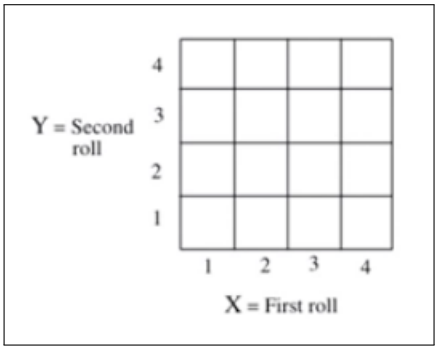
Let us now look at some examples of sample spaces. *Sample spaces are sets*. And a set can be:

- Discrete or continuous
- Finite or infinite

Let us start with a simpler case in which we have a sample space that is discrete and finite. The particular experiment we will be looking at is the following. We take a very special die, a tetrahedral die. So it's a die that has four faces numbered from 1 up 4. We roll it once. And then we roll it twice [again].

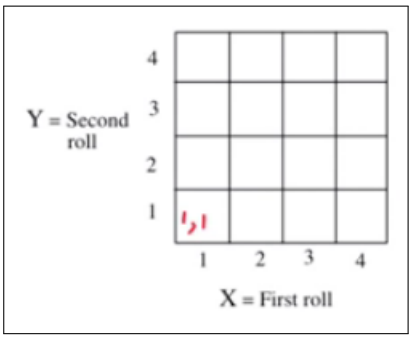
Were not dealing here with two probabilistic experiments. We're dealing with a single probabilistic experiment that involves two rolls of the die within that experiment. What is the sample space of that experiment?

Well, one possible representation is the following:

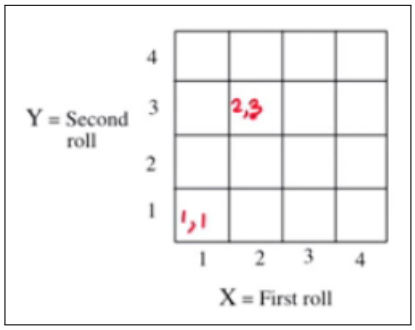


We take note of the result of the first roll. And then we take note of the result of the second roll. And this gives us a pair of numbers. Each one of the possible pairs of numbers corresponds to one of the little squares in this diagram.

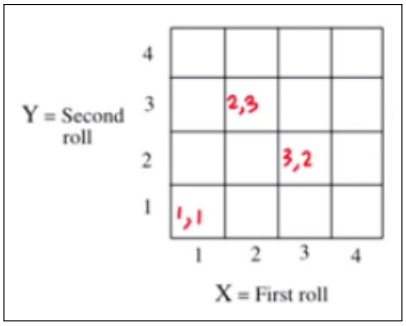
For example, if the first roll is 1 and the second is also 1, then this particular outcome has occurred:



If the first roll is it 2 and the second is a 3, then this particular outcome occurs:



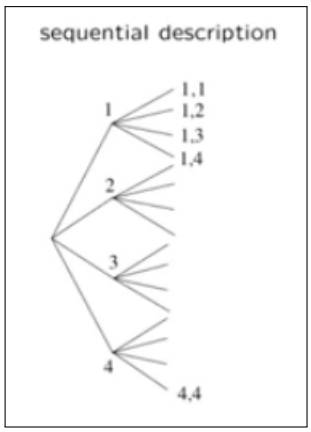
If the first roll is a 3 and then the next one is a 2, then this particular outcome occurs:



Notice that these two outcomes are pretty closely related. In both cases, we observe a 2 and we observe a 3. But we distinguish those two outcomes because in those two outcomes, the 2 and the 3 happen in different order. And the order in which they appear may be a detail which is of interest to us. And so we make this distinction in the sample space. So we keep the (3, 2) and the (2, 3) as separate outcomes.

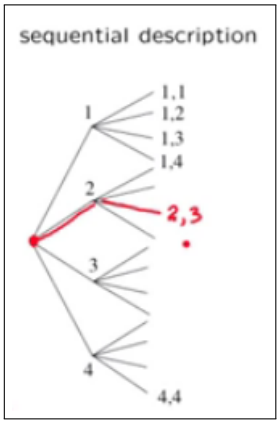
Now this is a case of a model in which *the probabilistic experiment can be described in phases or stages*. We could think about rolling the die once and then going ahead with the second roll. So we have two stages.

A very useful way of describing the sample space of experiments — whenever we have an experiment with several stages, either real stages or imagined stages — it is by providing a *sequential description in terms of a tree*.

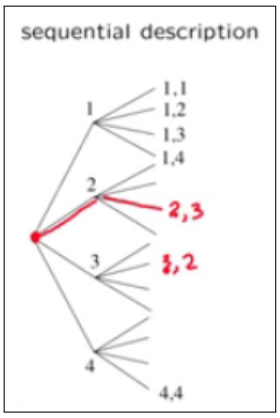


So a diagram of this kind, we call it a tree. You can think of this as the *root* of the tree from which you start. And the endpoints of the tree, we usually call them the *leaves*.

So the experiment starts. We carry out the first phase, which in this case is the first roll. And we see what happens. So maybe we get a 2 in the first roll. And then we take note of what happened in the second roll. And maybe the result was a 3. So we follow this branch here. And we end up at this particular leaf, which is the leaf associated with the outcome 2, 3:



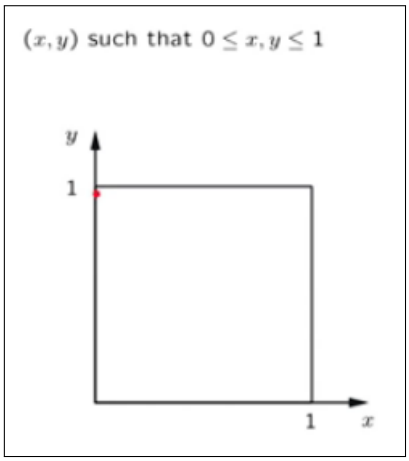
Notice that in this tree we once more have a distinction. The outcome 2 followed by a 3 is different from the outcome 3 followed by a 2, which would correspond to this particular place in the diagram:



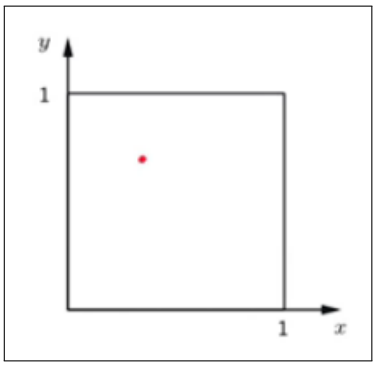
In both cases, we have 16 possible outcomes. 4 times 4 makes 16. And similarly, if you count here, the number of leaves is equal to 16.

The previous example involves a sample space that was discrete and finite. There were only 16 possible outcomes. But sample spaces can also be infinite. And they could also be continuous sets. Here's an example of an experiment that involves a continuous sample space.

So we have a rectangular target which is the unit square:



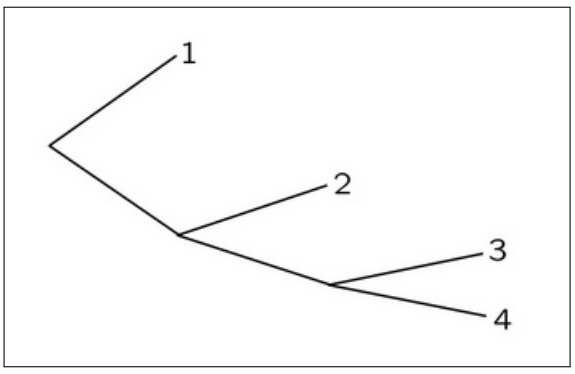
And you throw a dart on that target. And suppose that you are so skilled that no matter what, when you throw the dart, it always falls inside the target:



Once the dart hits the target, you record the coordinates x and y of the particular point that resulted from your dart throw. And we record x and y with *infinite precision*. So x and y are real numbers. So in this experiment, the sample space is just the set of x, y pairs that lie between 0 and 1 (inclusive): $\Omega = \{(x, y) \quad | \quad 0 \leq x, y \leq 1\}$

3.1.6 Exercise: Tree representations

Exercise 3.1.6-1: Tree representations
Paul checks the weather forecast. If the forecast is good, Paul will go out for a walk. If the forecast is bad, then Paul will either stay home or go out. If he goes out, he might either remember or forget his umbrella. In the tree diagram below, identify the leaf that corresponds to the event that the forecast is bad and Paul stays home.



- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4

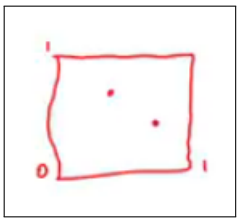
3.1.7 Probability axioms

Video: Probability axioms (transcripts)

We have so far discussed the first step involved in the *construction of a probabilistic model*, namely, the **construction of a sample space**, which is a description of the possible outcomes of a probabilistic experiment.

We now come to the second and much more interesting part. We need to *specify which outcomes are more likely to occur* and which ones are *less likely to occur* and so on. And we will do that by **assigning probabilities to the different outcomes**. However, as we try to do this assignment, we run into some kind of difficulty, which is the following.

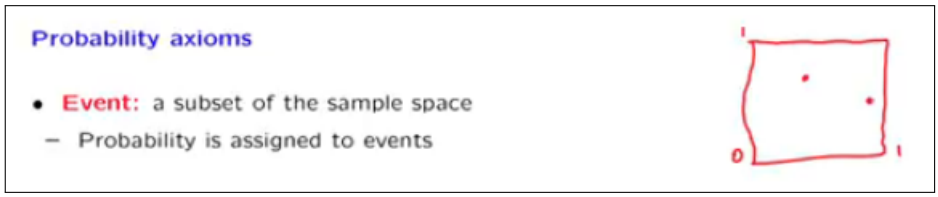
Remember the previous experiment involving a continuous sample space, which was the unit square and in which we throw a dart at random and record the point that occurred:



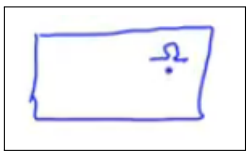
In this experiment, what do you think is the probability of a particular point?

Let's say what is the probability that my dart hits exactly the center of this square. Well, this probability would be essentially 0. Hitting the center *exactly with infinite precision* should be 0.

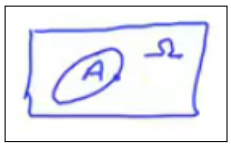
And so it's natural that *in such a continuous model any individual point should have a 0 probability*. For this reason instead of assigning probabilities to individual points, we will instead *assign probabilities to whole sets*, that is, to subsets of the sample space.



So here we have our sample space, which is some abstract set omega:

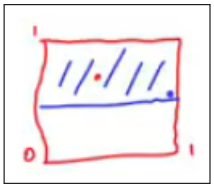


Here is a subset of the sample space, call it capital A:



We're going to assign a probability to that subset A, which we're going to denote with this notation, $P(A)$, which we read as the probability of set A:

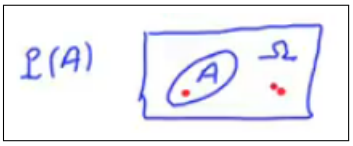
So *probabilities will be assigned to subsets*. And these will not cause us difficulties in the continuous case because even though individual points would have 0 probability, if you ask me what are the odds that my dart falls in the upper half, let's say, of this diagram, then that should be a reasonable positive number:



So even though individual outcomes may have 0 probabilities, sets of outcomes in general would be expected to have positive probabilities.

So coming back, we're going to *assign probabilities to the various subsets* of the sample space. And here comes a piece of terminology, that a subset of the sample space is called an **event**.

Why is it called an event? Because once we carry out the experiment and we observe the outcome of the experiment, either this outcome is inside the set A and in that case we say that *event A has occurred*, or the outcome falls outside the set A in which case we say that *event A did not occur*:



Now we want to move on and describe certain rules. The rules of the game in probabilistic models, which are basically the rules that these probabilities should satisfy. They shouldn't be completely arbitrary.

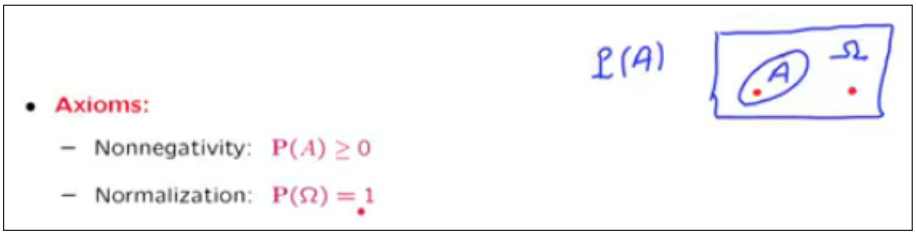
First, by convention, probabilities are always given in the range between 0 and 1. Intuitively:

- probability 0 means that we believe that something practically cannot happen
- probability 1 means that we're practically certain that an event of interest is going to happen.

So we want to specify rules of these kind for probabilities. These *rules that any probabilistic model should satisfy* are called the **axioms of probability theory**.

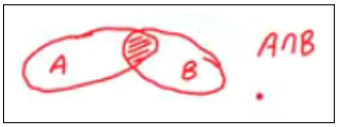
Our first axiom is a **nonnegativity axiom**, $P(A) \geq 0$. Namely, probabilities will always be non-negative numbers. It's a reasonable rule.

The second rule is that if the subset that we're looking at is actually not a subset but is the *entire sample space* Ω, the probability of it should always be equal to 1. What does that mean? We know that the outcome is going to be an element of the sample space. This is the definition of the sample space. So we have absolute certainty that our outcome is going to be in Ω. Or in different language we have absolute certainty that event Ω is going to occur. And we capture this certainty by saying that the probability of event omega is equal to 1: $P(\Omega) = 1$.



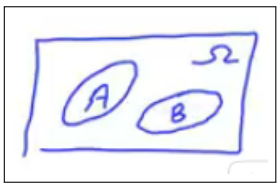
These two axioms are pretty simple and very intuitive. The more interesting axiom is the next one that says something a little more complicated.

Before we discuss that particular axiom, a quick reminder about set theoretic notation. If we have two sets, let's say a set A, and another set, another set B, we use this particular notation, $A \cap B$, which we read as "A intersection B"to refer to the collection of elements that belong to both A and B. So in this picture, the intersection of A and B is this shaded set:



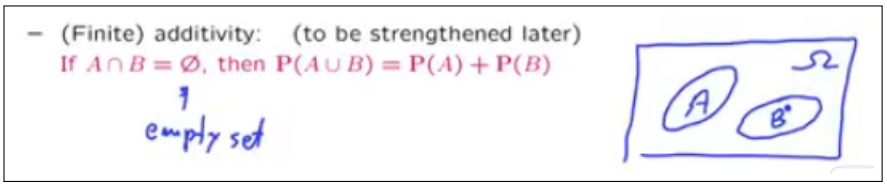
We use this notation, $A \cup B$, which we read as "A union B", to refer to the set of elements that belong to A or to B or to both. So in terms of this picture, the union of the two sets would be this blue set:

After this reminder about set theoretic notation, now let us look at the form of the third axiom. What does it say? If we have two sets, *two events*, *two subsets* of the sample space, which are *disjoint*. So here's our sample space. And here are the two sets that are disjoint:



In mathematical terms, two sets being disjoint means that *their intersection has no elements*, $A \cap B = \varnothing$. So their intersection is the empty set. And we use this symbol, $\varnothing$, here to denote the empty set.

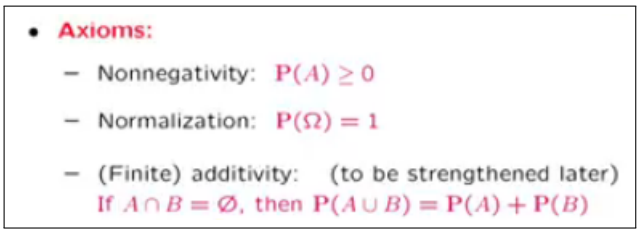
So *if the intersection of two sets is empty*, then the probability that the outcome of the experiments falls in the union of A and B, that is, the probability that the outcome is here or there, is equal to the *sum of the probabilities of these two sets*:



This is called the **additivity** axiom. So it says that *we can add probabilities of different sets when those two sets are disjoint*.

In some sense we can think of probability as being one pound of some substance which is spread over our sample space and the probability of A is how much of that substance is sitting on top of a set A. So what this axiom is saying is that the total amount of that substance sitting on top of A and B is how much is sitting on top of A plus how much is sitting on top of B. And that is the case whenever the sets A and B are disjoint from each other.

The additivity axiom needs to be refined a bit. We will talk about that a little later. Other than this refinement, *these three axioms are the only requirements in order to have a legitimate probability model*:



At this point you may ask, shouldn't there be more requirements? Shouldn't we, for example, say that probabilities cannot be greater than 1? Yes and no. We do not want probabilities to be larger than 1, but we do not need to say it. As we will see in the next segment, such a requirement follows from what we have already said. And the same is true for several other natural properties of probabilities.

3.1.8 Exercises: axioms

Exercise 3.1.8-1: Axioms
Let A and B be events on the same sample space, with $P(A) = 0.6$ and $P(B) = 0.7$. Can these two events be disjoint?

◯ Yes
◯ No

3.1.9 Simple properties of probabilities

Video: Simple properties of probabilities (transcripts)

The *probability axioms are the basic rules of probability theory*. And they are surprisingly few. But they imply many interesting properties that we will now explore.

First we will see that what you might think of as missing axioms are actually implied by the axioms already in place. For example, we have an axiom that probabilities are non-negative. We will show that probabilities are also less than or equal to 1. We have another axiom that says that the probability of the entire sample space is 1. We will show a counterpart that the probability of the empty set is equal to 0.

This makes perfect sense. The empty set has no elements, so it is impossible. There is 0 probability that the outcome of the experiment would lie in the empty set.

We also have another intuitive property. The probability that an event happens ($P(A)$) plus the probability that the vendor does not happen ($P(A^c)$) exhaust all possibilities. And these two probabilities together should add to 1 ($P(A) + P(A^c) = 1$):

$$\mathrm{P}(A) + \mathrm{P}(A^c) = 1$$

For instance, if the probability of heads is 0.6, then the probability of tails should be 0.4.

Finally, we can generalize the additivity axiom, which was originally given for the case of two disjoint events to the case where we're dealing with the union of several disjoint events. By disjoint here we mean that the intersection of any two of these events is the empty set ($A \cap B = \varnothing$, $A \cap C = \varnothing$, $B \cap C = \varnothing$):

For disjoint events:

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$$

$$\mathrm{P}(A) + \mathrm{P}(A^c) = 1$$

$$\mathrm{P}(A \cup B \cup C) = \mathrm{P}(A) + \mathrm{P}(B) + \mathrm{P}(C)$$

and similarly for k disjoint events

We will prove this for the case of three events and then the argument generalizes for the case where we're taking the union of k disjoint events, where k is any finite number. So the intuition of this result is the same as for the case of two events. But we will derive it formally and we will also use it to come up with a way of calculating the probability of a finite set by simply adding the probabilities of its individual elements:

$$\begin{aligned}\mathrm{P}(\{s_1, s_2, \ldots, s_k\}) &= \mathrm{P}(\{s_1\}) + \cdots + \mathrm{P}(\{s_k\}) \\ &= \mathrm{P}(s_1) + \cdots + \mathrm{P}(s_k)\end{aligned}$$

All of these statements that we just presented are intuitive. And you do not to really need to be convinced about their validity. Nevertheless, it is instructive to see how these statements follow from the axioms that we have put in place:

Some simple consequences of the axioms

Axioms	Consequences
$\mathrm{P}(A) \geq 0$	$\mathrm{P}(A) \leq 1$
$\mathrm{P}(\Omega) = 1$	$\mathrm{P}(\varnothing) = 0$
For disjoint events:	$\mathrm{P}(A) + \mathrm{P}(A^c) = 1$
$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$	$\mathrm{P}(A \cup B \cup C) = \mathrm{P}(A) + \mathrm{P}(B) + \mathrm{P}(C)$
	and similarly for k disjoint events
	$\mathrm{P}(\{s_1, s_2, \ldots, s_k\}) = \mathrm{P}(\{s_1\}) + \cdots + \mathrm{P}(\{s_k\})$
	$= \mathrm{P}(s_1) + \cdots + \mathrm{P}(s_k)$

So we will now present the arguments based only on the three axioms that we have available. And in order to be able to refer to these axioms, let us give them some names, call them axioms A, B, and C:

Some simple consequences of the axioms

Axioms

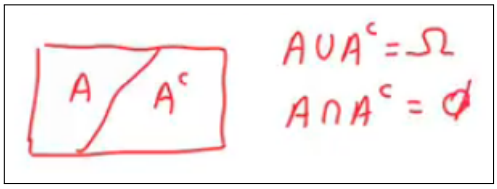
(a) $\mathrm{P}(A) \geq 0$

(b) $\mathrm{P}(\Omega) = 1$

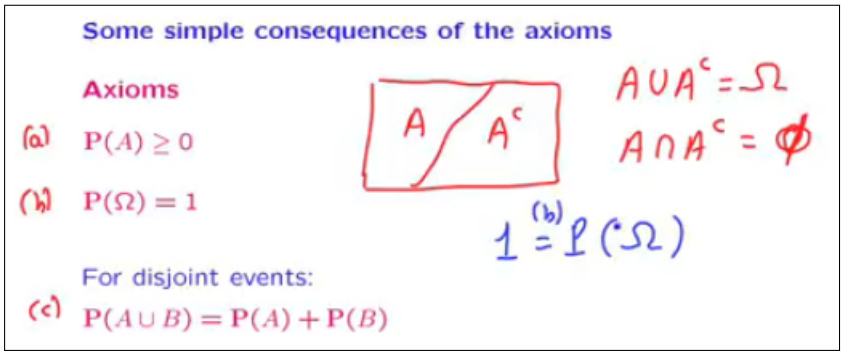
For disjoint events:

(c) $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$

We start as follows. Let us look at the sample space and a subset of that sample space. Call it A. And consider the complement of that subset, A^c. The complement is the set of all elements that do not belong to the set A. So a set together with its complement make up everything, which is the entire sample space: $A \cup A^c = \Omega$. On the other hand, if an element belongs to a set A, it does not belong to its complement. So the intersection of a set with its complement is the empty set, $A \cap A^c = \varnothing$:



Now we argue as follows. We have that the probability of the entire sample space is equal to 1. This is true by our second axiom, $1 = P(\Omega)$.



Now the sample space, as we just discussed, can be written as the union of an event and the complement of that event. This is just a set theoretic relation:

$$1 \overset{(b)}{=} P(\Omega) = P(A \cup A^c)$$

And next since a set and its complement are disjoint, this means that we can apply the additivity axiom and write this probability as the sum of the probability of event A with the probability of the complement of A. This is one of the relations that we had claimed and which we have now established:

$$1 \overset{(b)}{=} P(\Omega) = P(A \cup A^c) \overset{(c)}{=} P(A) + P(A^c)$$

Based on this relation, we can also write that the probability of an event A is equal to 1 minus the probability of the complement of that event:

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

And because, by the non-negativity axiom this quantity here, $P(A^c)$, is non-negative, 1 minus something non-negative is less than or equal to 1. We're using here the non-negativity axiom. And we have established another property, namely that probabilities are always less than or equal to 1:

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

Finally, let us note that 1 is the probability, always, of a set plus the probability of a complement of that set. And let us use this property for the case where the set of interest is the entire sample space:

$$1 = P(\Omega) + P(\Omega^c)$$

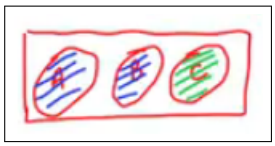
Now, the probability of the entire sample space is itself equal to 1, $P(\Omega) = 1$. And what is the complement of the entire sample space, $P(\Omega^c)$? The complement of the entire sample space consists of all elements that do not belong to the sample space. But since the sample space is supposed to contain all possible elements, its complement is just the empty set. And from this relation we get the implication that the probability of the empty set is equal to 0:

$$1 = P(\Omega) + P(\Omega^c)$$
$$1 = 1 + P(\emptyset) \Rightarrow P(\emptyset) = 0$$

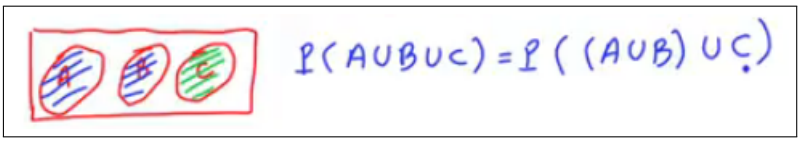
This establishes yet one more of the properties that we had just claimed a little earlier.

We finally come to the proof of the generalization of our additivity axiom from the case of two disjoint events to the case of three disjoint events. So we have our sample space. And within that sample space we have three events, three subsets. And these subsets are disjoint in the sense that any two of those subsets have no elements in common:

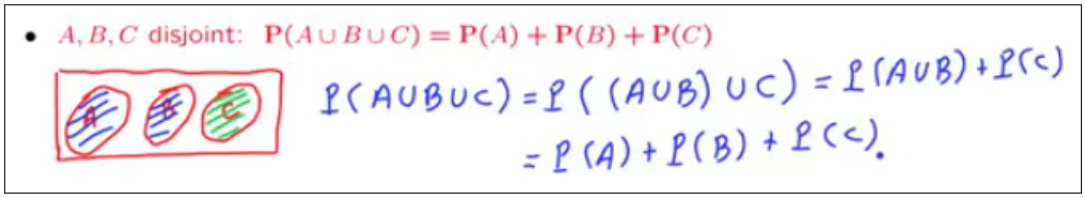
And we're interested in the probability of the union of A, B, and C, $P(A \cup B \cup C)$. How do we make progress? We have an additivity axiom in our hands, which applies to the case of the union of two disjoint sets. Here we have three of them. But we can do the following trick. We can think of the union of A, B, and C as consisting of the union of this blue set with that green set:



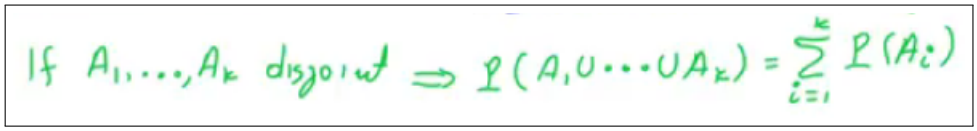
Formally, what we're doing is that we're expressing the union of these three sets as follows. We form one set by taking the union of A with B. And we have the other set C. And the overall union can be thought of as the union of these two sets:



Now since the three sets are disjoint, this implies that the blue set is disjoint from the green set and so we can use the additivity axiom here to write this probability as the probability of A union B plus the probability of C. And now we can use the additivity axiom once more since the sets A and B are disjoint to write the first term as probability of A plus probability of B. We carry over the last term and we have the relation that we wanted to prove:



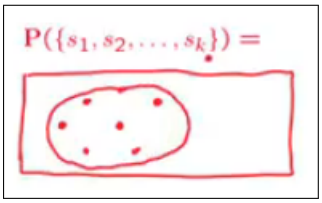
This is the proof for the case of three events. You should be able to follow this line of proof to write an argument for the case of four events and so on. And you might want to continue by induction. And eventually you should be able to prove that if the sets $A_1, \cdots A_k$ are disjoint, then the probability of the union of those sets is going to be equal to the sum of their individual probabilities:



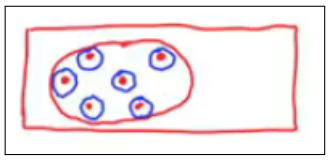
So this is the generalization to the case where we're dealing with the union of finitely many disjoint events.

A very useful application of this comes in the case where we want to calculate the *probability of a finite set*. So here we have a sample space, and within

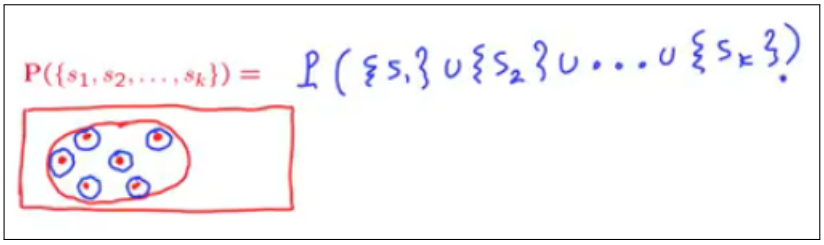
that sample space we have some particular elements S_1, S_2, up to S_k, k of them. And these elements together form a finite set:



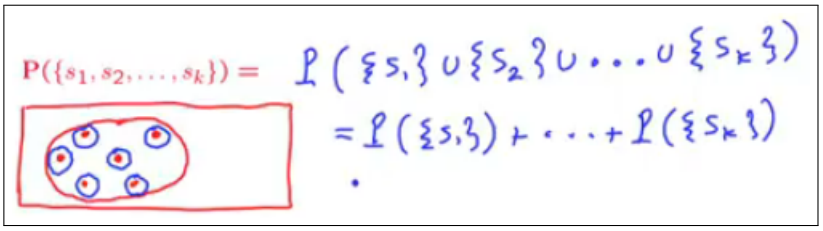
What can we say about the probability of this finite set? The idea is to take this finite set that consists of k elements and think of it as the union of several little sets that contain one element each:



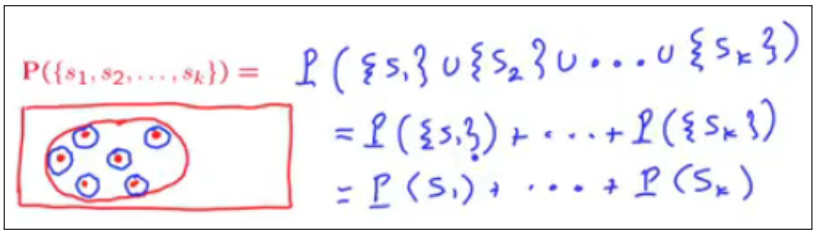
So set theoretically what we're doing is that we're taking this set with k elements and we write it as the union of a set that contains just S_1, a set that contains just the second element S_2, and so on, up to the k-th element:



We're assuming, of course, that these elements are all different from each other. So in that case, these sets, these single element sets, are all disjoint. So using the additivity property for a union of k disjoint sets, we can write this as the sum of the probabilities of the different single element sets:



At this point, it is usual to start abusing, or rather, simplifying notation a little bit. *Probabilities are assigned to sets*. So here we're talking about the probability of a set that contains a single element. But intuitively, we can also talk as just the probability of that particular element and use this simpler notation. So when using the simpler notation, we will be talking about the probabilities of individual elements. Although in terms of formal mathematics, what we really mean is the **probability of this event** that's *comprised only of a particular element* S_1 and so on:



3.1.10 Exercise: Simple properties

Exercise 3.1.10-1: Simple properties
Let A, B, and C be disjoint subsets of the sample space. For each one of the following statements, determine whether it is true or false. Note: "False"means "not guaranteed to be true."

$P(A) + P(A^c) + P(B) = P(A \cup A^c \cup B)$

- ◯ True
- ◯ False

$P(A) + P(B) \leq 1$

- ◯ True
- ◯ False

$P(A^c) + P(B) \leq 1$

- ◯ True
- ◯ False

$P(A \cup B \cup C) \geq P(A \cup B)$

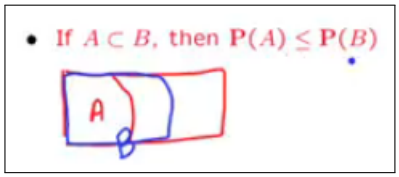
- ◯ True
- ◯ False

3.1.11 More properties of probabilities

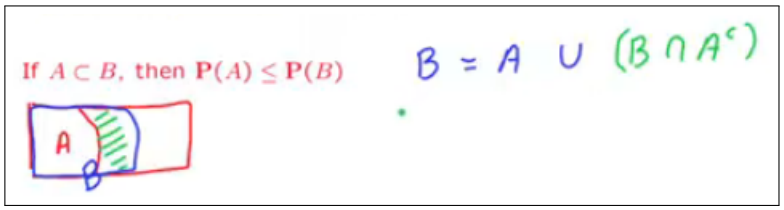
Video: More properties of probabilities (transcripts)

We will now continue and derive some additional properties of probability laws which are, again, consequences of the axioms that we have introduced.

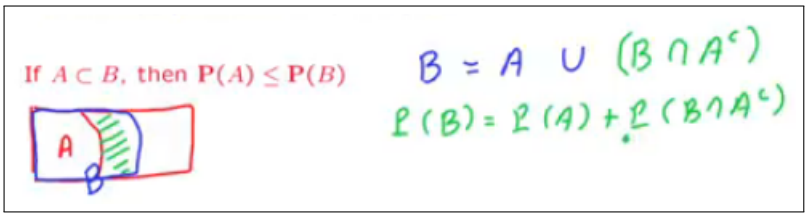
The first property is the following. If we have two sets and one set is inside the other, $A \subset B$, or $B \supset A$ — so we have a picture as follows. We have our sample space. And we have a certain set, A. And then we have a certain set, B, which is even bigger. So the set B is the bigger blue set. So if B is a set which is larger than A, then, naturally, the probability that the outcome falls inside B should be at least as big as the probability that the outcome falls inside A:



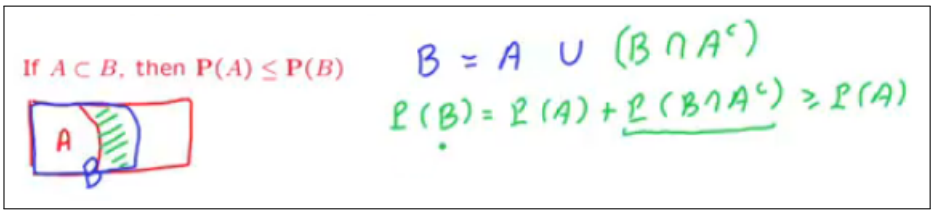
How do we prove this formally? The set B can be expressed as a union of two pieces. One piece is the set A itself. The second piece is whatever elements of B there are, that do not belong in A. What are these elements? They are elements that belong to B. And they do not belong to A, which means that they belong to the complement of A. So we have expressed the set B as the union of two pieces. Now this piece is A. This piece here is outside A. So these two pieces are disjoint:



And so we can apply the additivity axiom, and write that the probability of B is equal to the probability of A plus the probability of the other set:



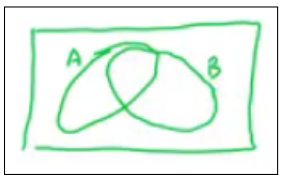
And since probabilities are non-negative, this expression here is at least as large as the probability of A. And this concludes the proof of the property that we wanted to show. Indeed, the probability of A is less than or equal to the probability of B:



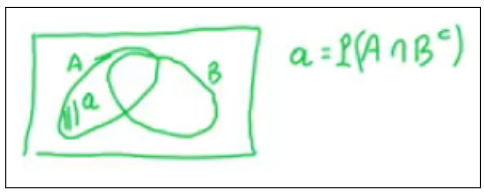
The next property we will show is the following. It allows us to write the probability of the union of two sets for the case now, where the two sets are not necessarily disjoint:

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$$

So the picture is as follows. We have our two sets, A and B. These sets are not necessarily disjoint:

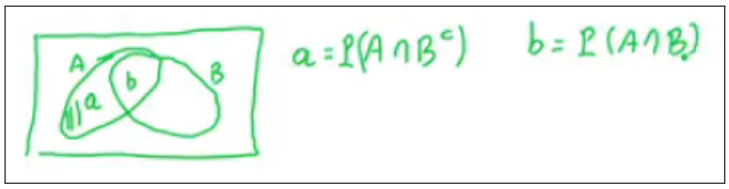


And we want to say something about the probability of the union of A and B. Now the union of A and B consists of three pieces. One piece is this one here. And that piece consists of those elements of A that do not belong to B: $A \cap B^c$. So they belong to B complement. This set has a certain probability, let's call it little "a" and indicate it on this diagram:

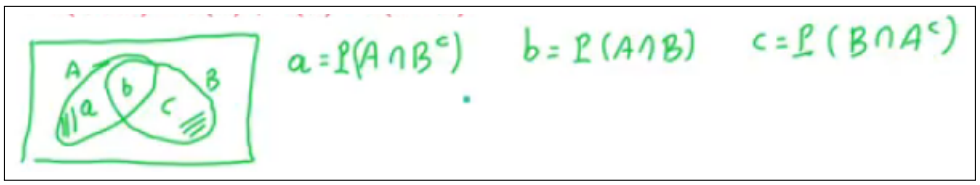


So "a" is the probability of this piece.

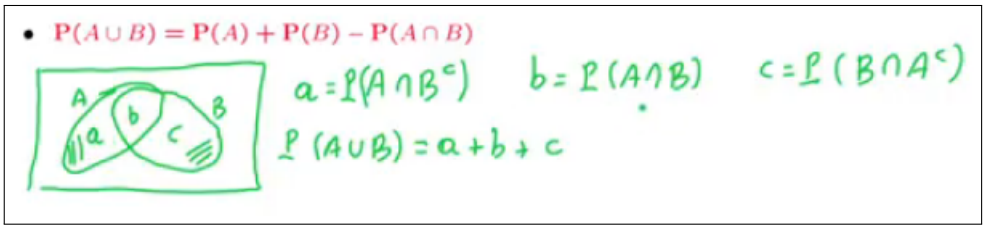
Another piece is this one here, which is the intersection of A and B, $A \cap B$. It has a certain probability that we denote by little "b". This is the probability of A intersection B, $P(A \cap B)$.



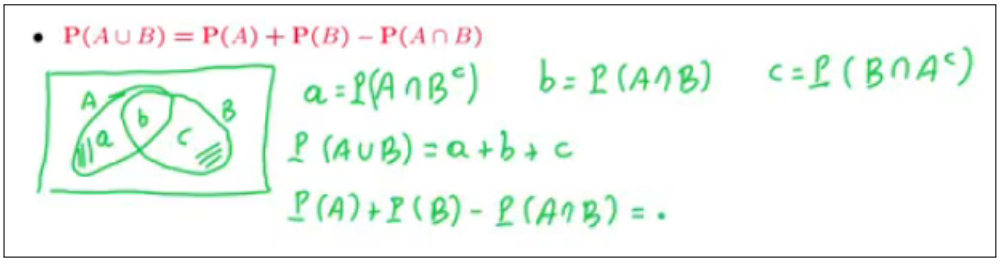
And finally, there's another piece, which is out here. And that piece has a certain probability "c". It is the probability of that set. And what is that set? That set is the following. It's that part of B that consists of elements that do not belong in A. So it's B intersection with the complement of A, $B \cap A^c$:



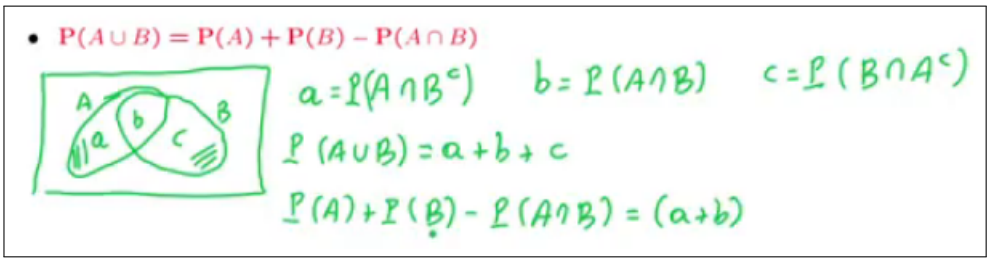
Now let's express the two sides of this equality here in terms of little "a", little "b", and little "c", and see whether we get the same thing. So the probability of A union B. A union B consists of these three pieces that have probabilities little "a", little "b", and little "c", respectively. And by the additivity axiom, the probability of the union of A and B is the sum of the probabilities of these three pieces:



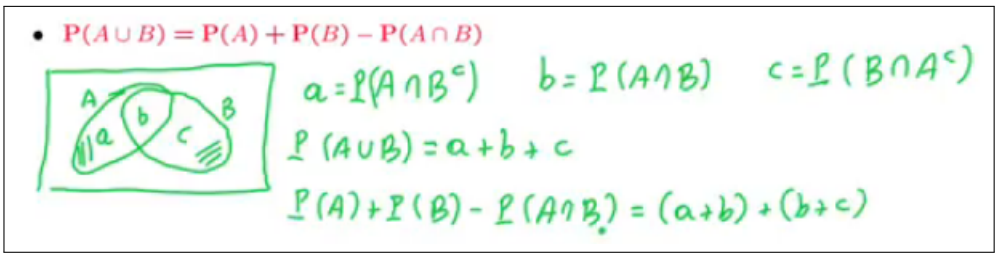
Let's look now at the right hand side of that equation and see whether we get the same thing. The probability of A plus the probability of B, minus the probability of A intersection B is equal to the following:



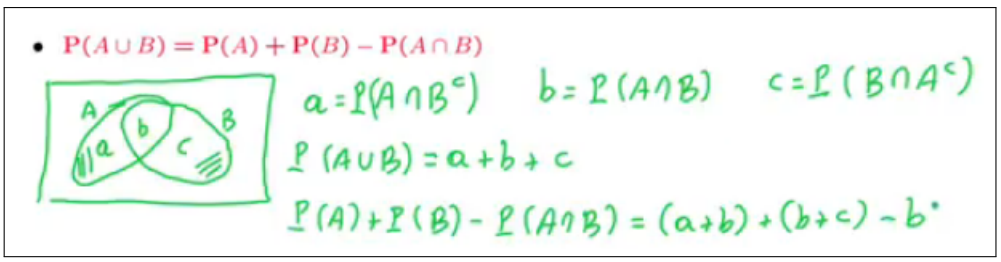
A consists of two pieces that have probabilities little "a" and little "b":



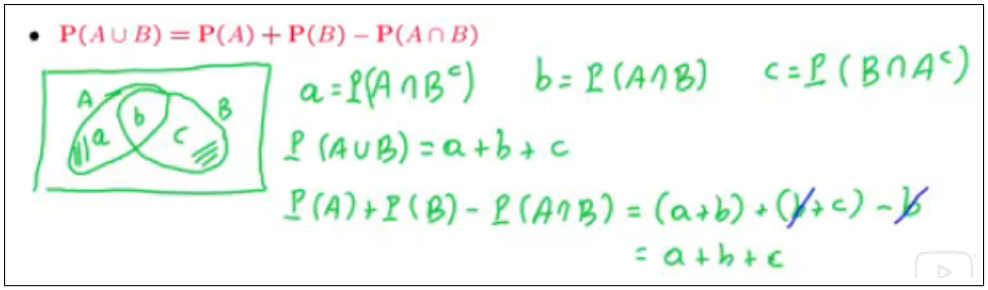
The set B consists of two pieces that have probabilities little "b" and little "c":



And then we subtract the probability of the intersection, which is "b":



And we notice that we can cancel here one "b" with another "b". And what we are left with is "a" plus "b" plus "c":



So this checks. And indeed we have this equality here. We have verified that it is true.

One particular consequence of the equality that we derived is the following. Since this term here is always non-negative,

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

this means that the $P(A \cup B)$ is *always* less than or equal to the $P(A) + P(B)$:

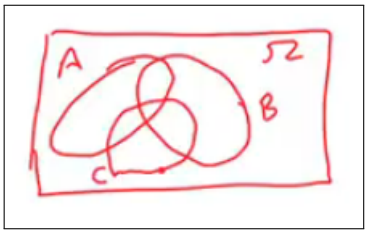
- $\mathrm{P}(A \cup B) \leq \mathrm{P}(A) + \mathrm{P}(B)$

This inequality here is quite useful whenever we want to argue that a certain probability is smaller than something. And it has a name. It's called the **union bound**.

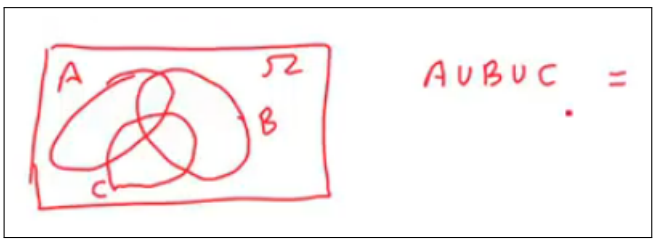
We finally consider one last consequence of our axioms. And namely, we are going to derive an expression, a way of calculating the probability of the union of three sets, not necessarily disjoint:

- $\mathrm{P}(A \cup B \cup C) = \mathrm{P}(A) + \mathrm{P}(A^c \cap B) + \mathrm{P}(A^c \cap B^c \cap C)$

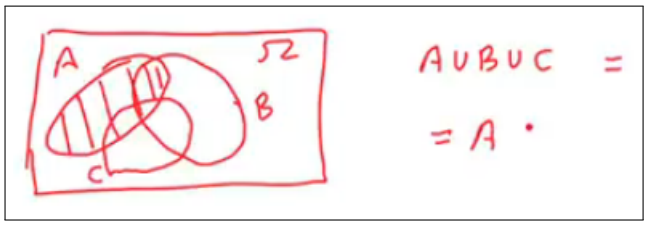
So we have our sample space. And within the sample space there are three sets — set A, set B, and set C:



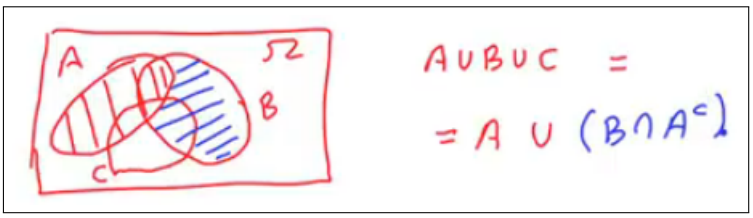
We are going to use a set theoretic relation. We are going to express the union of these three sets as the union of three disjoint pieces:



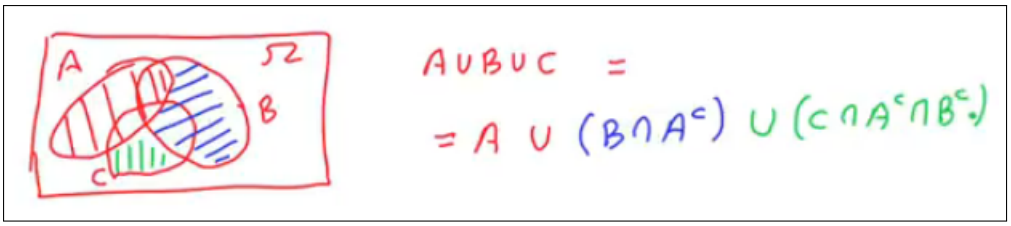
What are these disjoint pieces? One piece is the set A itself:



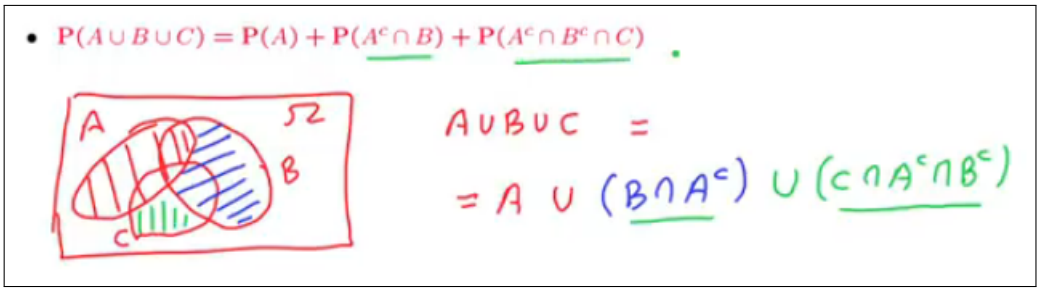
The second piece is going to be that part of B which is outside A. So this is the intersection of B with the complement of A:



The third piece is going to be whatever is left in order to form the union of the three sets. What is left is that part of C that does not belong to A and that does not belong to B. So that part is C intersection with A complement and B complement:



Now this set here, of course, is the same as that set because intersection of two sets is the same no matter in which order we take the two sets. And similarly, the set that we have here is the same one that appears in that expression:



Now we notice that these three pieces, the red, the blue, and the green, are disjoint from each other. So by the additivity axiom, the probability of this union here is going to be the sum of the probabilities of the three pieces. And that's exactly the expression the we have up here.

3.1.12 Exercise: More properties

Exercise 3.1.12-1: More properties
Let A, B, and C be subsets of the sample space, not necessarily disjoint. For each one of the following statements, determine whether it is true or false. Note: "False" means "not guaranteed to be true."

$P\left[(A \cap B) \cup (C \cap A^c)\right] \leq P(A \cup B \cup C)$

- ◯ True
- ◯ False

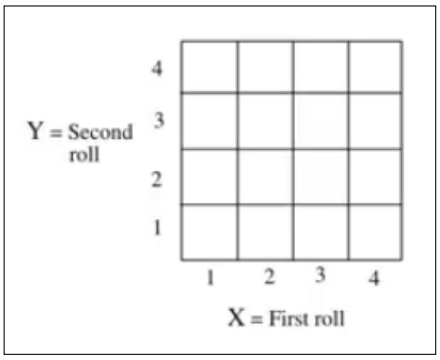
$P(A \cup B \cup C) = P(A \cap C^c) + P(C) + P(B \cap A^c \cap C^c)$

- ◯ True
- ◯ False

3.1.13 A discrete example

Video: A discrete example (transcripts)

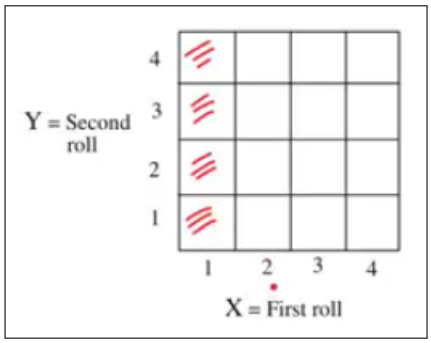
Let us now move from the abstract to the concrete. Recall the example that we discussed earlier where we have two rolls of a tetrahedral die. So there are 16 possible outcomes illustrated in this diagram:



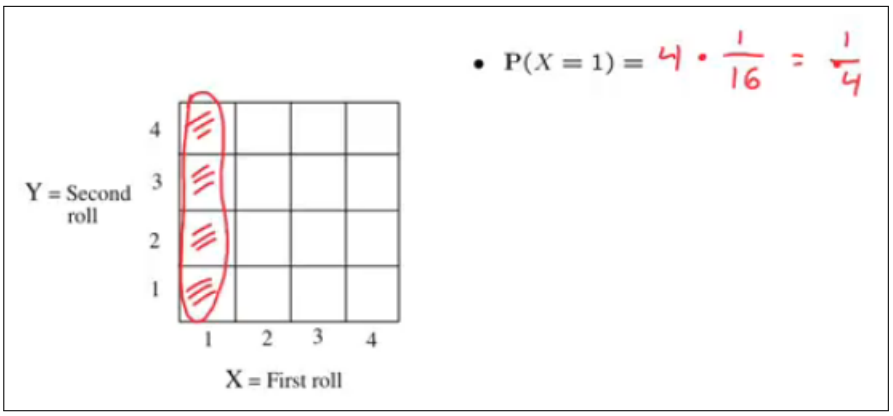
To continue, now we *need to specify a probability law*, some kind of probability assignment. To keep things simple, we're going to make the assumption that the 16 possible outcomes are all equally likely. And each outcome has a probability of 1 over 16 ($1/16$). Given this assumption, we will now proceed to calculate certain probabilities.

Let us look first at the probability that X, which stands the result of the first roll, is equal to 1: $X = 1$. The way to calculate this probability is to identify what exactly that event is in our picture of the sample space, and then

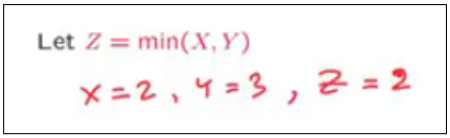
calculate. The event that X is equal to 1 can happen in four different ways that correspond to these four particular outcomes:



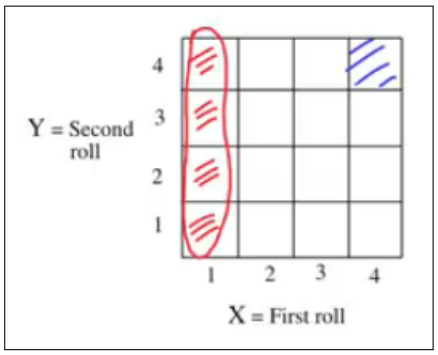
Each one of these outcomes has a probability of 1 over 16. The probability of this event is the sum of the probabilities of the outcomes that it contains. So it is 4 times 1/16, equal to 1/4:



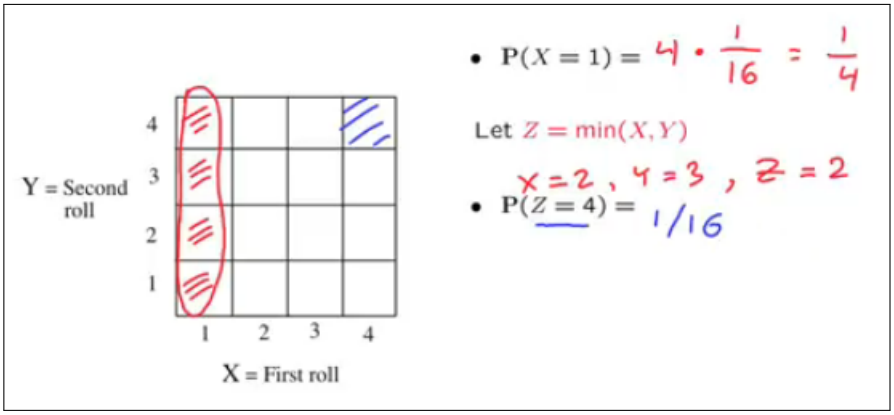
Let now Z stand for the smaller of the two numbers that came up in our two rolls. So for example, if $X = 2$ and $Y = 3$, then $Z = 2$, which is the smaller of the two:



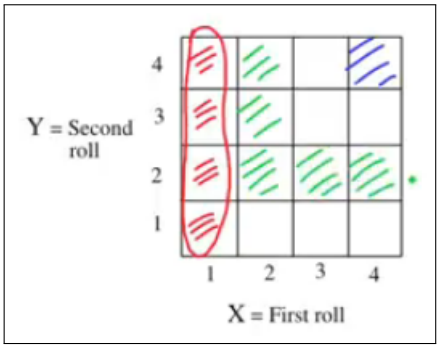
Let us try to calculate the probability that the smaller of the two outcomes is equal to 4, $Z = 4$. Now for the smaller of the two outcomes to be equal to 4, we must have that both X and Y are equal to 4. So this outcome here (in blue) is the only way that this particular event can happen:



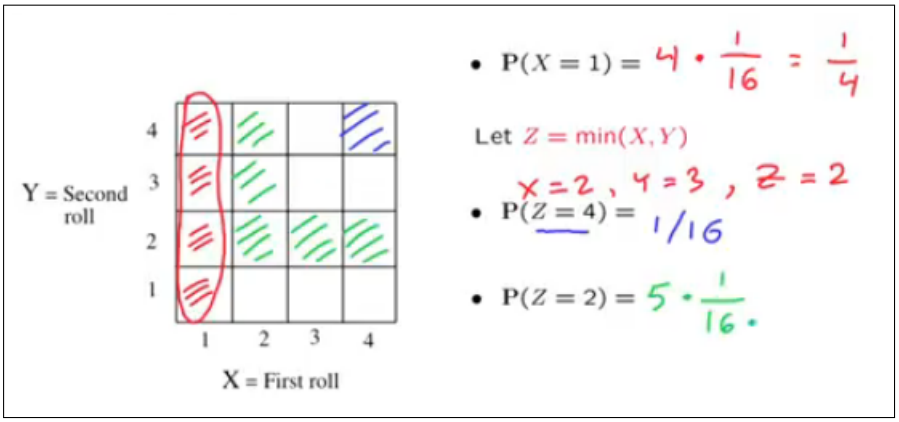
Since there's only one outcome that makes the event happen, the probability of this event is the probability of that outcome and is equal to 1/16.



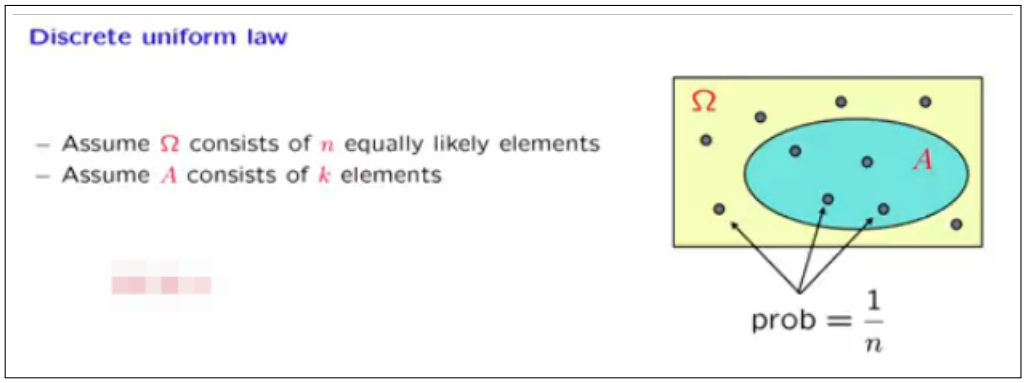
For another example, let's calculate the probability that the minimum is equal to 2. What does it mean that the minimum is equal to 2? It means that one of the dice resulted in a 2, and the other die resulted in a number that's 2 or larger. So we could have both equal to 2. We could have X equal to 2, but Y larger. Or we could have Y equal to 2 and X something larger. This green event, this green set, is the set of all outcomes for which the minimum of the two rolls is equal to 2:



There's a total of five such outcomes. Each one of them has probably 1 over 16. And we have discussed that for finite sets, *the probability of a finite set is the sum of the probabilities of the elements of that set*. So we have five elements here, each one with probability 1 over 16, and get 5 over 16, and this is the answer to this problem:

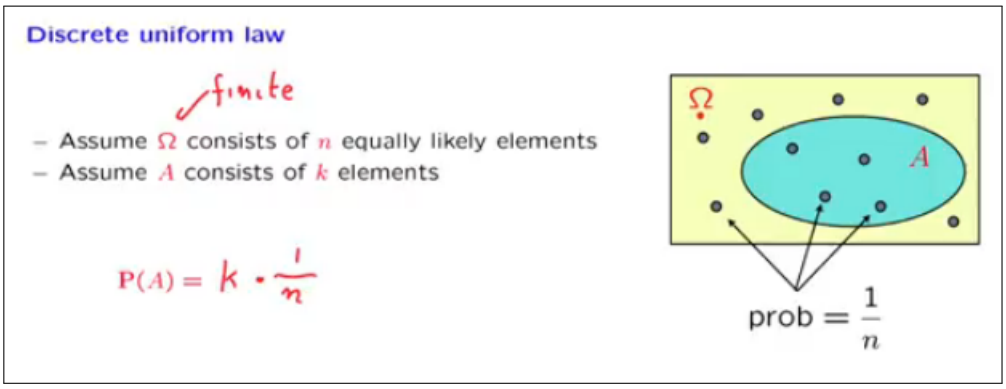


This particular example that we saw here is a special case of what is called a **discrete uniform law**. In a discrete uniform law, we have a *finite sample space*. And it has n elements. And we assume that these n elements are *equally likely*:



Now since the probability of omega, the probability of the entire sample space, is equal to 1, this means that each one of these elements must have probability 1/n. That's the only way that the sum of the probabilities of the different outcomes would be equal to 1 as required by the normalization axiom.

Consider now some subset of the sample space, an event A that, exactly k elements. What is the probability of the set A? It's the *sum of the probabilities of its elements*. There are k elements. And each one of them has a probability of 1/n. And this way we can find the probability of the set A:



So when we have a *discrete uniform probability law*, we can calculate probabilities by simply *counting the number of elements of omega*, which is n, finding the number n, and *counting the number of elements of the set A*. That's the reason why counting will turn out to be an important skill. And there will be a whole lecture devoted to this particular topic.

3.1.14 Exercise: Discrete probability calculations

Exercise 3.1.14-1: Discrete probability calculations
Consider the same model of two rolls of a tetrahedral die, with all 16 outcomes equally likely. Find the probability of the following events:

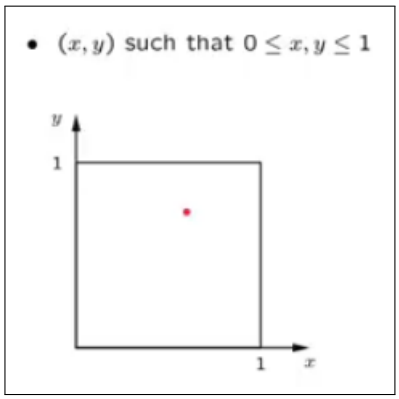
a) The value in the first roll is strictly larger than the value in the second roll.

b) The sum of the values obtained in the two rolls is an even number.

3.1.15 A continuous example

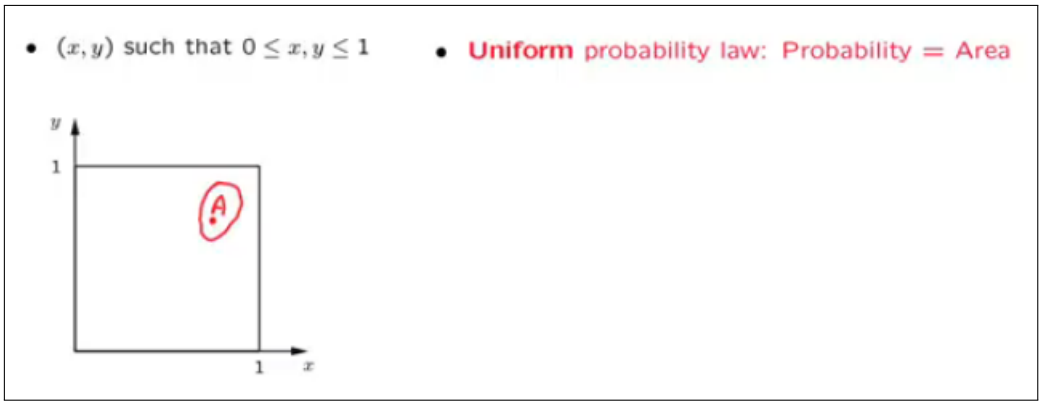
Video: A continuous example (transcripts)

We will now go through a probability calculation for the case where we have a *continuous sample space*. We revisit our earlier example in which we were throwing a dart into a square target, the square target being the unit square. And we were guaranteed that our dart would fall somewhere inside this set.



So our sample space is the unit square itself. We have a description of the sample space, but we do not yet have a probability law. We need to specify one.

The choice of a probability law could be arbitrary. *It's up to us to choose how to model a certain situation*. And to keep things simple, we're going to assume that our probability law is a uniform one, which means that the probability of any particular subset of the sample space is going to be the area of that subset. So if we have some subset lying somewhere here and we ask what is the probability that we fall into that subset? The probability is exactly the area of that particular subset:

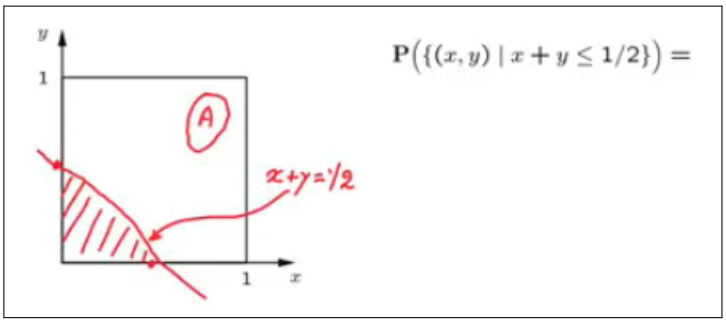


Once more, this is an arbitrary choice of a probability law. There's nothing in our assumptions so far that would force us to make this particular choice. And we just use it for the purposes of this example.

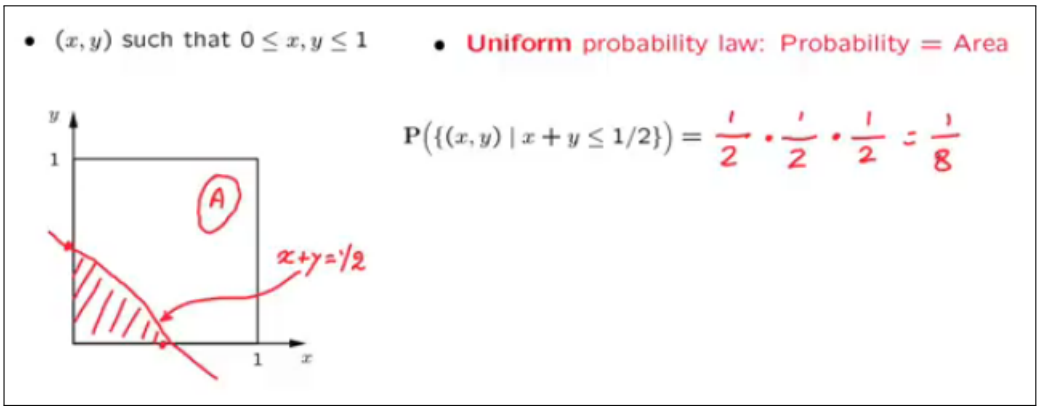
So now let us calculate some probabilities. Let us look at this event. This is the event that the sum of the two numbers that we get in our experiment is less than or equal to 1/2:

$$\mathbf{P}\big(\{(x, y) \mid x + y \leq 1/2\}\big) =$$

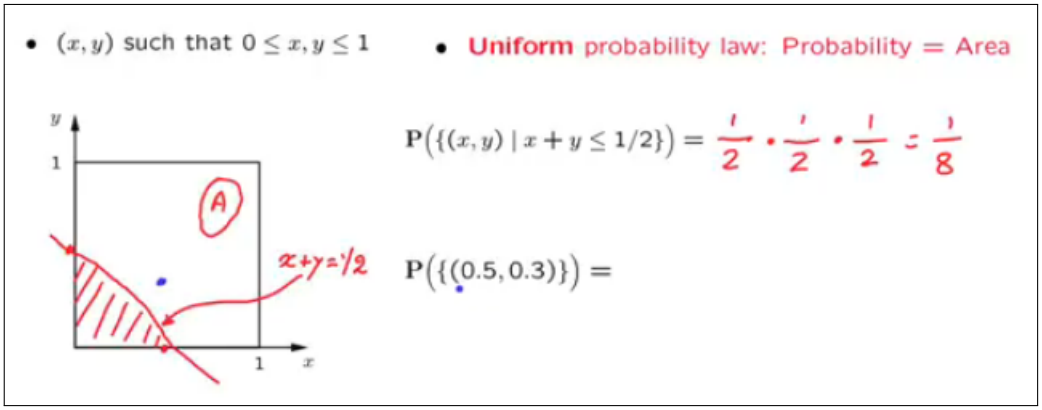
It is always useful to work in terms of a picture and to depict that event in a picture of the sample space. So in terms of that sample space, the points that make this event to be true are just a triangle that lies below the line, where this is the line, that's x plus y equals 1/2:



Anything below that line, these are the outcomes that make this event happen. So we're trying to find the probability of this red event. We have assumed that probability is equal to area. Therefore, the probability we're trying to calculate is the area of a triangle. And the area of a triangle is *1/2* times the *base* of the triangle, which is 1/2 in our case, times the *height* of the triangle, which is again 1/2 in our case. And the end result is 1/8:

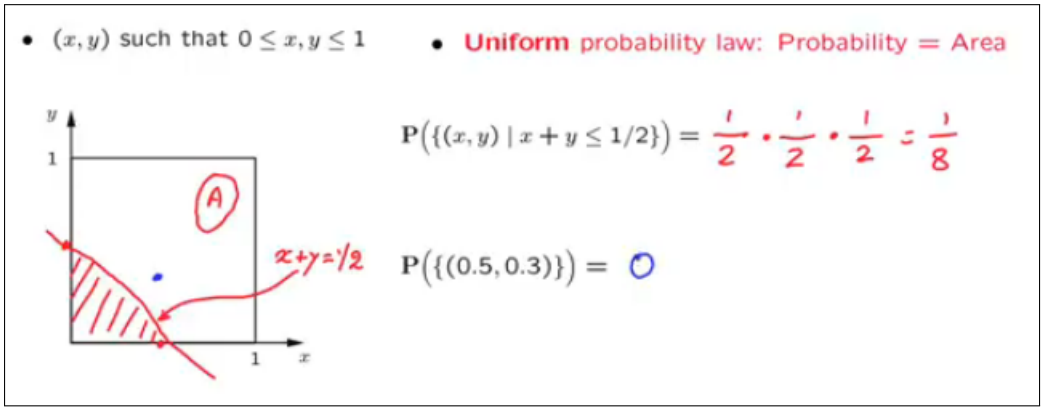


Let us now calculate another probability. Now, this is an event that consists of only a single element. We take the point 0.5, 0.3, which sits somewhere here:



The event of interest is a set, but that set consists of a single point. So we're asking for the probability that our dart falls exactly on top of that point. What

is it? Well, it is the area of a set that consists of a single point. What is the area of a single point? It is 0:



And similarly for any other single point inside that sample space that we might have considered, the answer is going to be 0.

Let us now abstract from this example, as well as the previous one, and note the following. **Probability calculations involve a sequence of four steps**:

1. *Specify the sample space*: starting with a word description of a problem, of a probabilistic experiment, we first write down the sample space.

2. *Specify the probability law*: Then we specify a probability law. Let me emphasize again here that this step has some arbitrariness in it. You can choose any probability law you like, although for your results to be useful it would be good if your probability law captures the real-world phenomenon you're trying to model.

3. *Identify an event of interest*: Typically you're interested in calculating the probability of some event. That event may be described in some loose manner, so you need to describe it mathematically. And if possible, it's always good to describe it in terms of a picture. Pictures are immensely useful when going through this process.

4. *Calculate*: And finally, the last step is to go ahead and calculate the probability of the event of interest.

Now, a probability law in principle specifies the probability of every event, and there's nothing else to do. But quite often the probability law will be given in some *implicit manner*, for example, by specifying the probabilities of only some of the events.

In that case, you may have to do some additional work to find the probability of the particular event that you care about. This last step sometimes will be easy. Sometimes it may be complicated. But in either case, by following this four-step procedure and by being systematic you will always be able to come up with a single correct answer.

3.1.16 Exercise: Continuous probability calculations

Exercise 3.1.16-1: Continuous probability calculations
Consider a sample space that is the rectangular region $[0,1] \times [0,2]$, i.e., the set of all pairs (x,y) that satisfy $0 \leq x \leq 1$ and $0 \leq y \leq 2$. Consider a "uniform"probability law, under which the probability of an event is *half of the area of the event*. Find the probability of the following events:

a) The two components x and y have the same values.

b) The value, x , of the first component is larger than or equal to the value, y, of the second component.

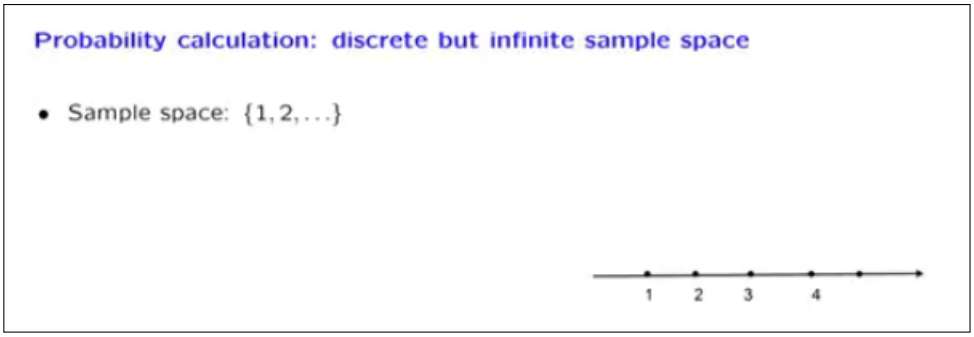
c) The value of x^2 is larger than or equal to the value of y.

3.1.17 Countable additivity

Video: Countable additivity (transcripts)

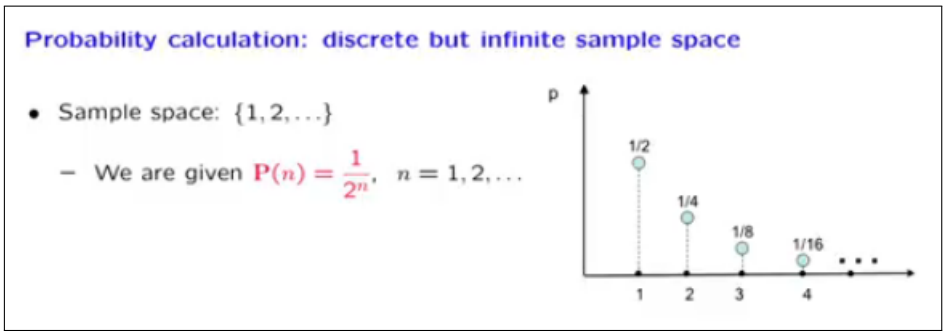
We have seen so far an example of a probability law on a *discrete and finite sample space* as well as an example with an *infinite and continuous* sample space.

Let us now look at an example involving a *discrete but infinite sample space*. We carry out an experiment whose outcome is an arbitrary positive integer. As an example of such an experiment, suppose that we keep tossing a coin and the outcome is the number of tosses until we observe heads for the first time:



The first heads might appear in the first toss or the second or the third, and so on. So in this example, any positive integer is possible. And so our *sample space is infinite*.

Let us now specify a probability law. A *probability law should determine the probability of every event, of every subset of the sample space*. That is, the probability of every set of positive integers. But instead I will just tell you the probability of events that contain a single element. I'm going to tell you that there is probability 1 over 2 to the n that the outcome is equal to n:



Is this good enough? Is this information enough to *determine the probability of any subset*?

Before we look into that question, let us first do a quick sanity check to see whether these numbers that we are given look like legitimate probabilities.

Do they add to 1? Let's do a quick check. So the sum over all the possible values of n of the probabilities that we're given, which is an infinite sum starting from 1, all the way up to infinity, of 1 over 2 to the n, is equal to the following:

$$\sum_{n=1}^{\infty} \frac{1}{2^n} =$$

First we take out a factor of 1/2 from all of these terms, which reduces the exponent from n to n minus 1. This is the same as running the sum from n equals 0 to infinity of 1/2 and to the n:

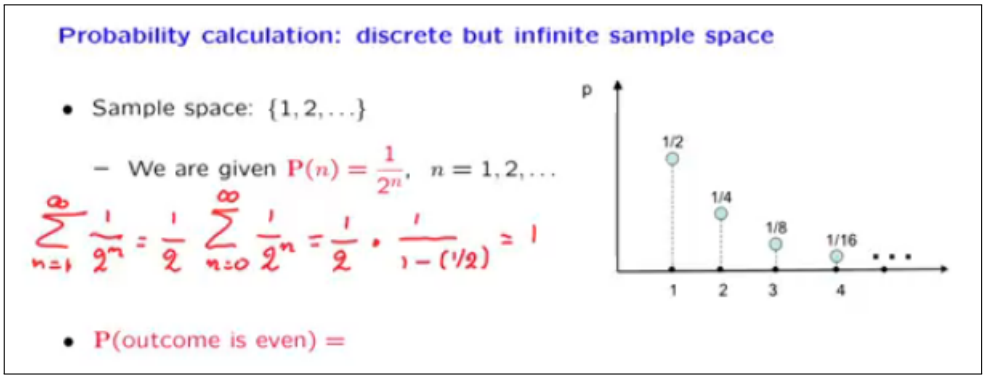
$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} \sum_{n=0}^{\infty} \frac{1}{2^n}$$

And now we have a usual infinite geometric series and we have a formula for this. The geometric series has a value of 1 over 1 minus the number whose power we're taking, which is 1/2. And after we do the arithmetic, this turns out to be equal to 1:

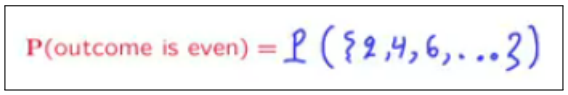
$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} \sum_{n=0}^{\infty} \frac{1}{2^n} = \frac{1}{2} \cdot \frac{1}{1-(1/2)} = 1$$

So indeed, it appears that we have the basic elements of what it would take to have a legitimate probability law.

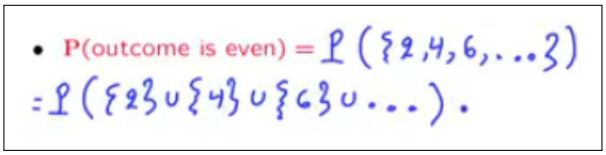
But now let us look into how we might calculate the probability of some general event. For example, the probability that the outcome is even:



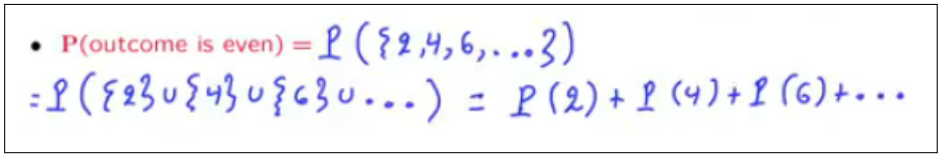
We proceed as follows. The probability that the outcome is even, this is the probability of an infinite set that consists of all the even integers:



We can write this set as the union of lots of little sets that contain a single element each. So it's the set containing the number 2, the set containing the number 4, the set containing the number 6, and so on:

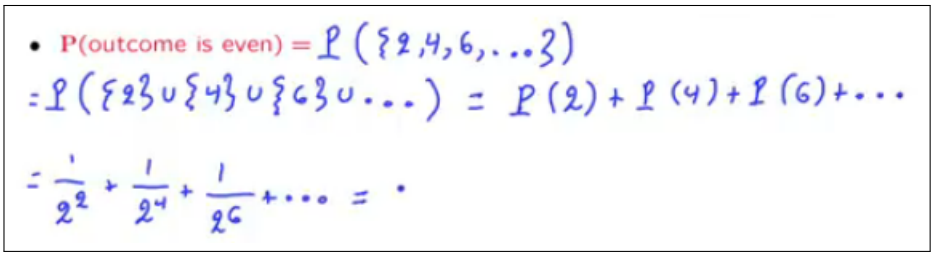


At this point we notice that we're talking about the probability of a union of sets and these sets are disjoint because they contain different elements. So we can use an additivity property and say that this is the probability of obtaining a 2, plus the probability of obtaining a 4, plus the probability of obtaining a 6 and so on:

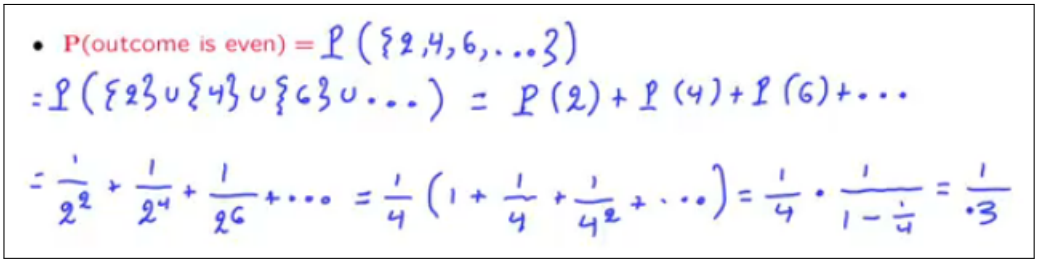


If you're curious about doing this calculation and actually obtaining a numerical answer, you would proceed as follows. You notice that this is 1 over

2 to the second power plus 1 over 2 to the fourth power plus 1 over 2 to the sixth power and so on:

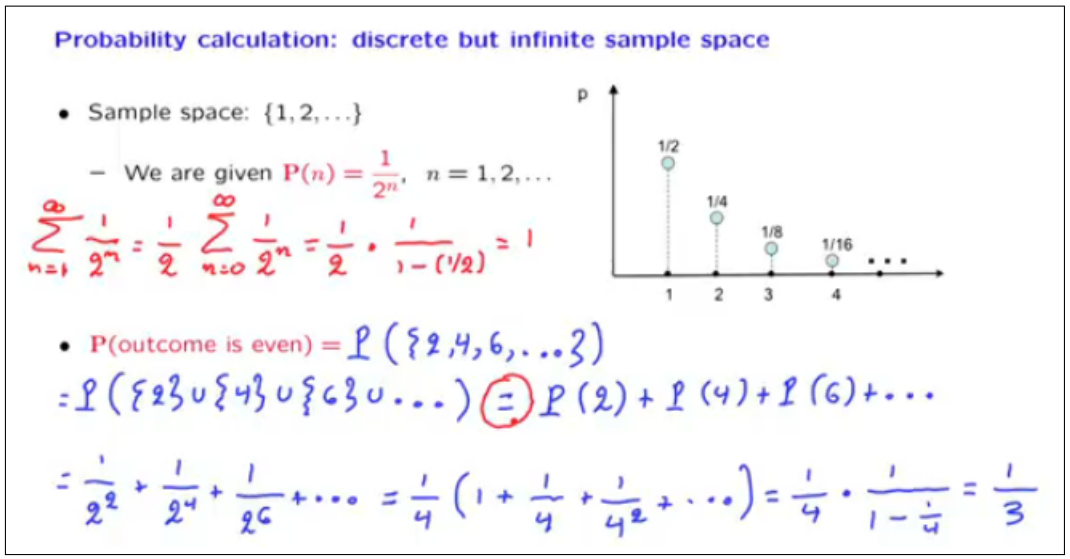


Now you factor out a factor of 1/4 and what you're left is 1 plus 1 over 2 to the second power, which is 1/4, plus 1 over 2 to the fourth power, which is the same as 1/4 to the second power and so on. And now we have 1/4 times the infinite sum of a geometric series, which gives us 1 over 1 minus 1/4. And after you do the algebra you obtain a numerical answer, which is equal to 1/3:

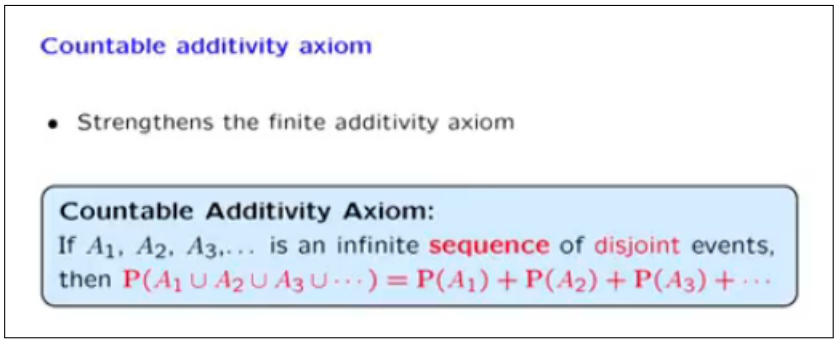


But leaving the details of the calculation aside, the more important question I want to address is the following. Is this calculation correct? We seem to have used an additivity property at this point. But the additivity properties that we have in our hands at this point only talk about *disjoint unions of finitely many subsets*.

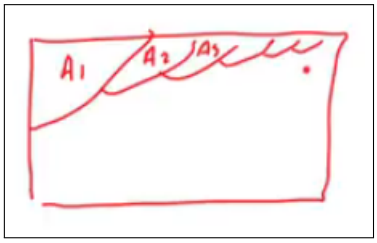
Our initial axiom talked about a disjoint union of two subsets and then later on we established a similar property for a *disjoint union of finitely many subsets*. But here we're talking about the *union of infinitely many subsets*. So this step here is not really allowed by what we have in our hands:



On the other hand, we would like our theory to allow this kind of calculation. The way out of this dilemma is to introduce an additional axiom that will indeed allow this kind of calculation. The axiom that we introduce is the following:

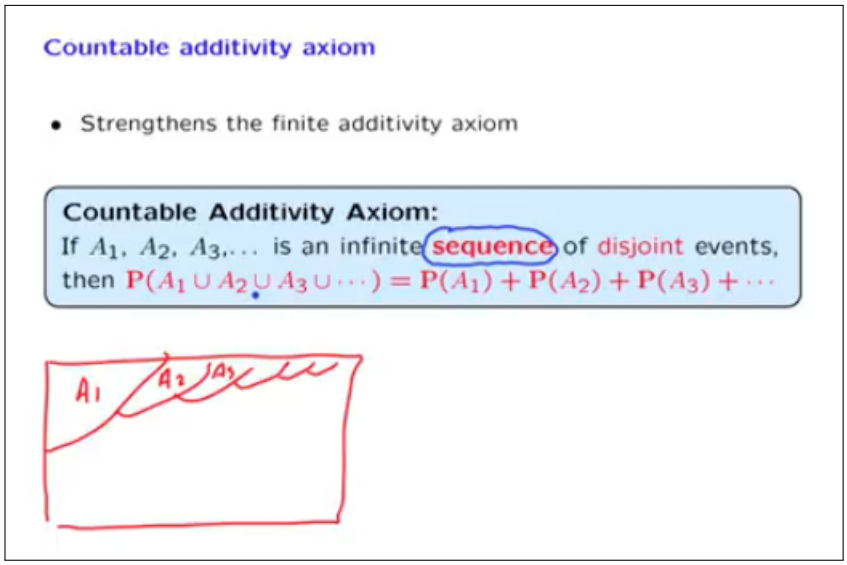


If we have an *infinite sequence of disjoint events*, as for example in this picture. We have our sample space. We have a first event, A1. We have a second event, A2. The third event, A3. And so we keep continuing and we have an infinite sequence of such events:



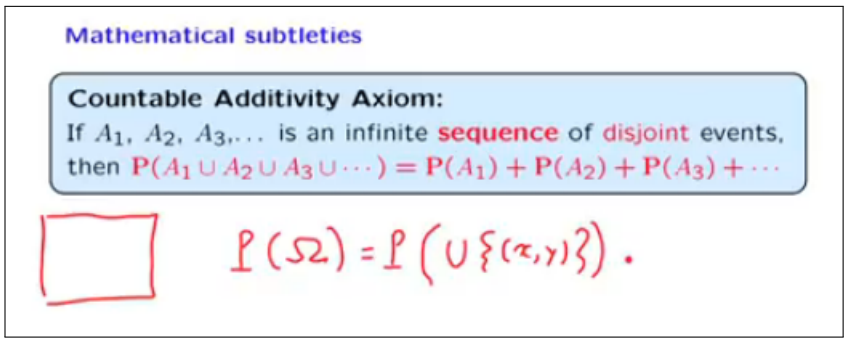
Then *the probability of the union of these events, of these infinitely many events, is*

the sum of their individual probabilities. The key word here is the word **sequence**. Namely, these events, these sets that we're dealing with, can be arranged so that we can talk about the first event, A1, the second event, A2, the third one, A3, and so on:

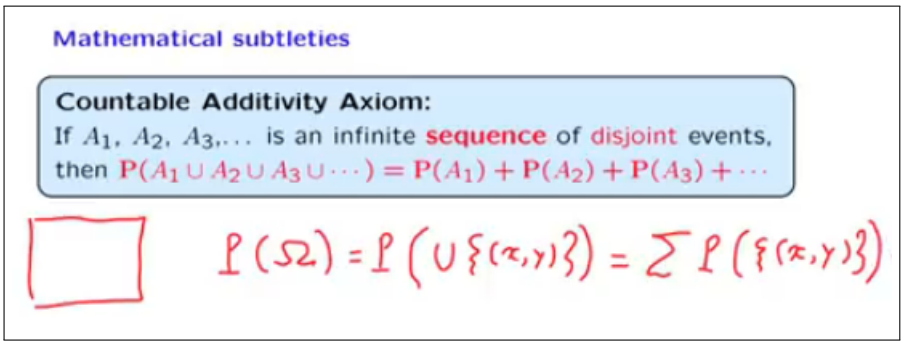


To appreciate the issue that arises here and to see why the word sequence is so important, let us consider the following calculation. Our sample space is the unit square. And we consider a model where the probability of a set is its area, as in the examples that we considered earlier.

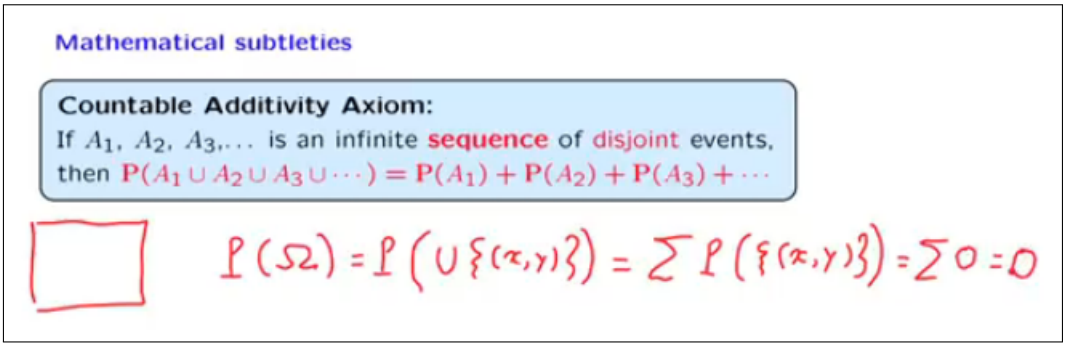
Let us now look at the probability of the overall sample space. Our sample space is the unit square and the unit square can be thought of as the union of various sets that consist of single points. So it's the union of subsets with one element each. And it's a union taken over all the points in the unit square:



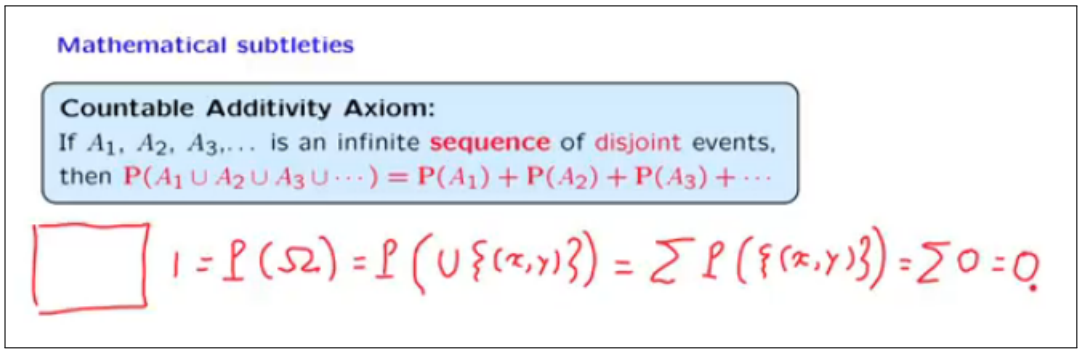
Then we think about additivity. We observe that these subsets are disjoint. If we're considering different points, then we get disjoint single element sets. And then an additivity property would tells us that the probability of this union is the sum of the probabilities of the different single element subsets:



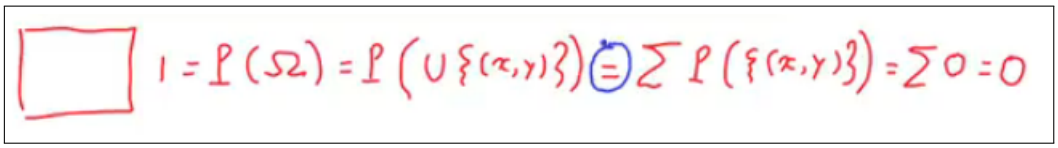
Now, as we discussed before, single element subsets have 0 probability. So we have a sum of lots of 0s and the sum of 0s should be equal to 0:



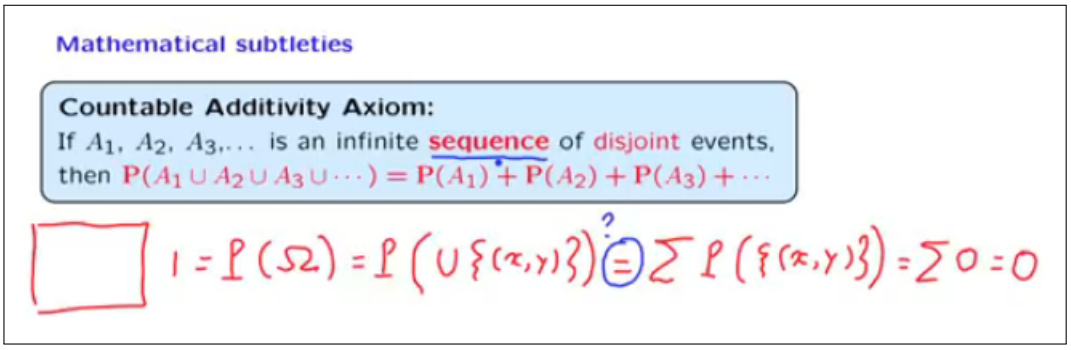
On the other hand, by the probability axioms, the probability of the entire sample space should be equal to 1. And so we have established that 1 is equal to 0:



This looks like a paradox. Is it? The catch is that there is nothing in the axioms we have introduced so far or the properties we have established that would justify this step:

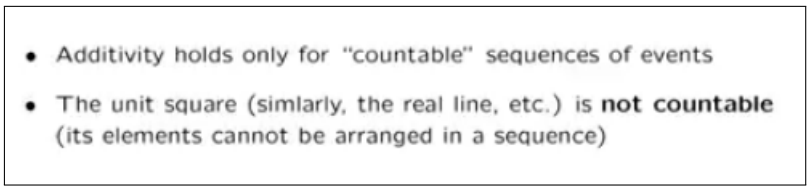


So this step here is questionable. You might argue that the unit square is the union of disjoint single element sets, which is the case that we have in additivity axioms. But the *additivity axiom only applies when we have a sequence of events*:



And this is not what we have here. This is not a union of a sequence of single element sets.

In fact, there is no way that the elements of the unit square can be arranged in a sequence. The unit square is said to be an **uncountable set**:



This is a deep and fundamental mathematical fact. What it essentially says is that *there are two kinds of infinite sets*:

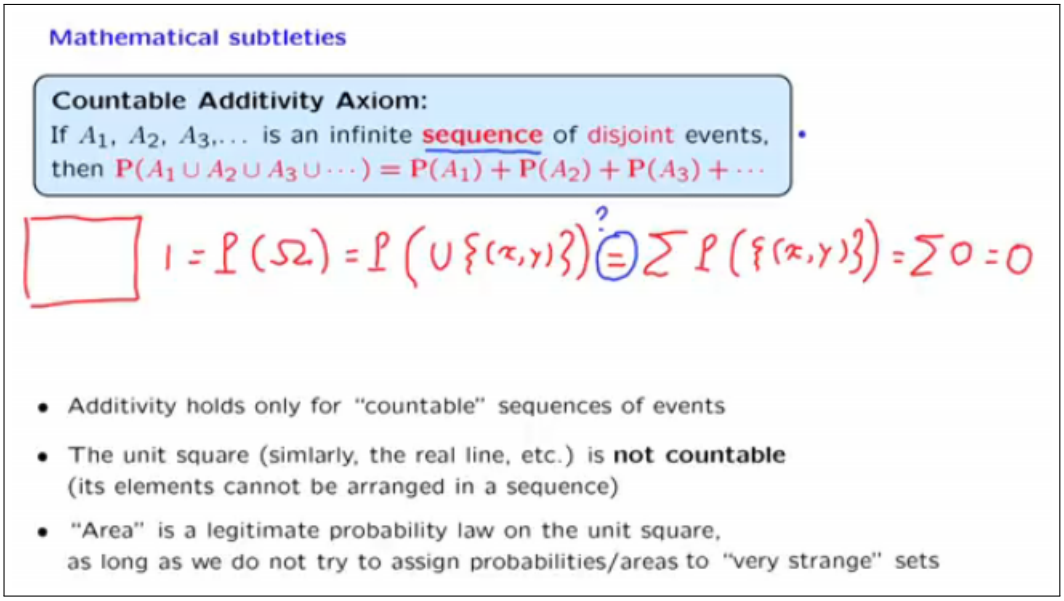
- *Countable*: Discrete ones or in formal terminology countable. These are sets whose elements can be arranged in a sequence, like the integers.
- *Uncontable*: And also uncountable sets, such as the unit square or the real line, whose elements cannot be arranged in a sequence.

If you're curious, you can find the proof of this important fact in the supplementary materials that we are providing.

After all these discussion, you may now have legitimate suspicions about the models we have been looking at. Is area a legitimate probability law? Does

it even satisfy countable additivity? This question takes us into deep waters and has to do with a deep subfield of mathematics called *Measure Theory*.

Fortunately, it turns out that all is well. Area is a legitimate probability law. It does indeed satisfy the countable additivity axiom as long as we only deal with "nice subsets" of the unit square. Fortunately, the subsets that arise in whatever we do in this course will be "nice":



Subsets that are not nice are quite pathological and we will not encounter them.

At this stage we are not in a position to say anything more that would be meaningful about these issues because they're quite complicated and mathematically deep. We can only say that there are some serious mathematical subtleties. But fortunately, they can all be overcome in a rigorous manner. And for the rest of this class, you can just forget about these subtle issues.

3.1.18 Exercise: Using countable additivity

Exercise 3.1.18-1: Using countable additivity
Let the sample space be the set of positive integers and suppose that $P(n) = 1/2^n$, for $n = 1, 2, \ldots, n$. Find the probability of the set $3, 6, 9, \ldots$, that is, of the set of of positive integers that are multiples of 3.

3.1.19 Exercise: Uniform probabilities on the integers

Exercise 3.1.19-1: Uniform probabilities on the integers
Let the sample space be the set of positive integers. Is it possible to have a "uniform"probability law, that is, a probability law that assigns the same probability to each positive integer?

◯ Yes
◯ No

3.1.20 Exercise: On countable additivity

Exercise 3.1.20-1: On countable additivity
Let the sample space be the two-dimensional plane. For any real number x, let A_x be the subset of the plane that consists of all points of the vertical line through the point $(x, 0)$, i.e., $A_x = (x, y) : y \in Re$.

a) Do the axioms of probability theory imply that the probability of the union of the sets A_x (which is the whole plane) is equal to the sum of the probabilities $P(A_x)$?

◯ Yes
◯ No

b) Do the axioms of probability theory imply that $P(A_1 \cup A_2 \cup \cdots) = \sum_{x=1}^{\infty} P(A_x)$? (In other words, we consider only those lines for which the x coordinate is a positive integer.)

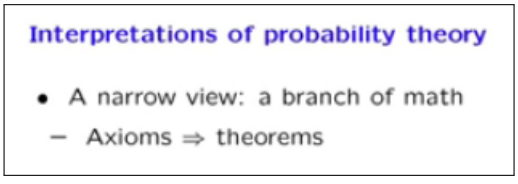
◯ Yes
◯ No

3.1.21 Interpretations and uses of probabilities

Video: Interpretations and uses of probabilities (transcripts)

We end this lecture sequence by stepping back to discuss *what probability theory really is* and *what exactly is the meaning of the word probability*.

In the most narrow view, probability theory is just a branch of mathematics. We start with some axioms. We consider models that satisfy these axioms, and we establish some consequences, which are the theorems of this theory:



You could do all that without ever asking the question of what the word "probability"really means. Yet, one of the theorems of probability theory, that we will see later in this class, is that probabilities can be interpreted as frequencies, very loosely speaking:

Interpretations of probability theory

- A narrow view: a branch of math
  - Axioms ⇒ theorems **"Thm:"** "Frequency" of event A "is" $\mathbf{P}(A)$

If I have a fair coin, and I toss it infinitely many times, then the fraction of heads that I will observe will be one half. In this sense, the probability of an event, A, can be interpreted as the *frequency with which event A will occur in an infinite number of repetitions of the experiment*. But is this all there is? If we're dealing with coin tosses, it makes sense to think of probabilities as frequencies:

Interpretations of probability theory

- A narrow view: a branch of math
  - Axioms ⇒ theorems **"Thm:"** "Frequency" of event A "is" $\mathbf{P}(A)$
- Are probabilities frequencies?
  - $\mathbf{P}$(coin toss yields heads) $= 1/2$

But consider a statement such as the "current president of my country will be reelected in the next election with probability 0.7". It's hard to think of this number, 0.7, as a frequency. It does not make sense to think of infinitely many repetitions of the next election.

In cases like this, and in many others, it is better to think of *probabilities as just some way of describing our beliefs*:

Interpretations of probability theory

- A narrow view: a branch of math
  - Axioms ⇒ theorems **"Thm:"** "Frequency" of event A "is" $\mathbf{P}(A)$
- Are probabilities frequencies?
  - $\mathbf{P}$(coin toss yields heads) $= 1/2$
  - $\mathbf{P}$(the president of ... will be reelected) $= 0.7$
- Probabilities are often intepreted as:
  - Description of beliefs
  - Betting preferences

And if you're a betting person, probabilities can be thought of as some numerical guidance into what kinds of bets you might be willing to make.

But now if we think of probabilities as beliefs, you can run into the argument that, well, beliefs are subjective. Isn't probability theory supposed to be an objective part of math and science? Is probability theory just an exercise in subjectivity? Well, not quite. There's more to it:

The role of probability theory

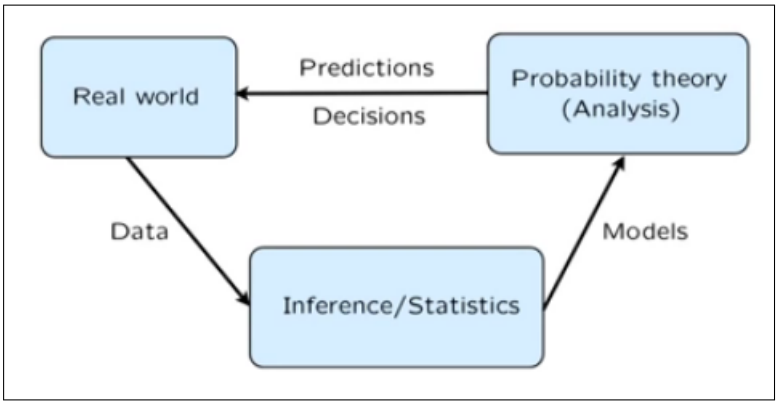
- A framework for analyzing phenomena with uncertain outcomes
  - Rules for consistent reasoning
  - Used for predictions and decisions

Probability, at the minimum, gives us some *rules for thinking systematically about uncertain situations*. And if it happens that *our probability model, our subjective beliefs*, have some relation with the real world, then probability theory can be a very useful tool for making predictions and decisions that apply to the real world.

Now, whether your predictions and decisions will be any good will depend on whether you have chosen a good model. Have you chosen a model that's provides a good enough representation of the real world? How do you make sure that this is the case?

There's a whole field, the field of *statistics, whose purpose is to complement probability theory by using data to come up with good models*.

And so we have the following diagram that summarizes the relation between the real world, statistics, and probability:



The real world generates data. The field of statistics and inference uses these data to come up with probabilistic models. Once we have a probabilistic model, we use probability theory and the analysis tools that it provides to us. And the results that we get from this analysis lead to predictions and decisions about the real world.

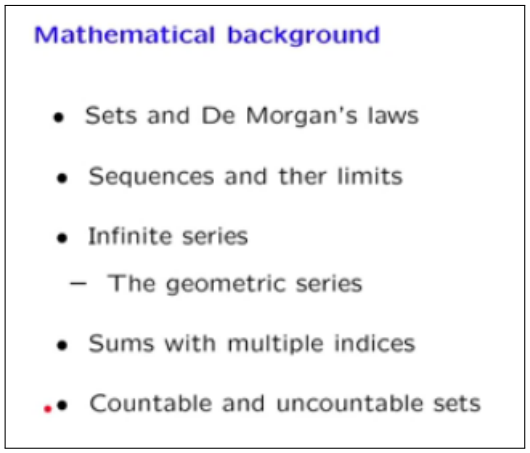
3.2 Mathematical background: Sets; sequences, limits, and series; (un)countable sets

3.2.1 Mathematical background overview

This collection of clips reviews some background material about sets, including De Morgan's laws (see also Section 1.1 of the text), sequences and their convergence, infinite series, infinite series with multiple indices, and uncountable sets.

Video: Mathematical background: Overview (transcripts, annotated slides)

In this sequence of segments, we review some mathematical background that will be useful at various places in this course. Most of what is covered, with the exception of the last segment, is material that you may have seen before. But this could still be an opportunity to refresh some of these concepts:



I should say that this is intended to be just a refresher. Our coverage is not going to be complete in any sense.

What we will talk about is sets, various definitions related to sets, and some basic properties, including De Morgan's laws.

We will talk about what a sequence is and what it means for a sequence to converge to something.

We will talk about infinite series. And as an example, we will look at the geometric series.

Then we will talk about some subtleties that arise when you have sums of terms that are indexed with multiple indices.

And finally, probably the most sophisticated part, will be a discussion of countable versus uncountable sets.

Countable sets are like the integers. Uncountable sets are like the real line. And they're fundamentally different. And this fundamental difference reflects itself into fundamentally different probabilistic models — models that involve discrete experiments and outcomes versus models that involve continuous outcomes.

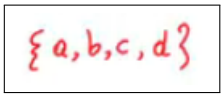
3.2.2 Sets

Video: Sets (transcripts)

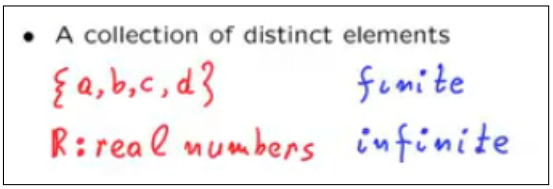
In this segment, we will talk about sets. I'm pretty sure that most of what I will say is material that you have seen before. Nevertheless, it is useful to do a review of some of the concepts, the definitions, and also of the notation that we will be using.

So what is a set? A *set is just a collection of distinct elements*. So we have some elements, and we put them together. And this collection, we call it the set S:

More formally, how do we specify a set? We could specify a set by *listing its elements, and putting them inside braces*. So this is a set that consists of four elements, the letters, a, b, c, d:

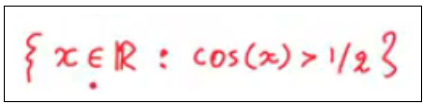


Another set could be the set of all real numbers. Notice a distinction here — the first set is a *finite set*. It has a finite number of elements, whereas the second set is *infinite set*. And in general, sets are of these two kinds. Either they're finite, or their infinite:



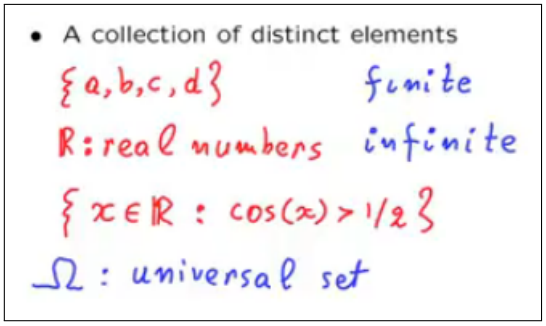
A piece of notation now. We use this notation to indicate that a certain object x is an element of a set S: $x \in S$ (we read that as x belongs to S). If x is not an element of S, then we use this notation to indicate it, $x \notin S$ (we read it as x does not belong to S).

Now, one way of specifying sets is as follows. We start with a bigger se — for example, the set of real numbers — and we consider all of those x's that belong to that big set that have a certain property. For example, that the cosine of this number is, let's say, bigger than 1/2:

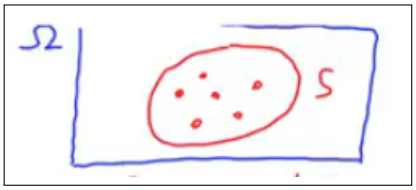


This is a way of specifying a set. We *start with a big set*, but we then *restrict to those elements of that set that satisfy a particular property*.

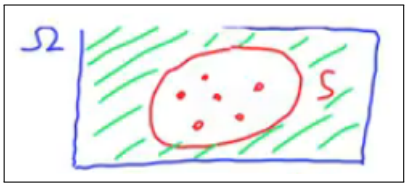
One set of particular interest is the following. Sometimes in some context, we want to fix a collection of all possible objects that we might ever want to consider, and that collection will be a set. We denote it usually by Ω, and we call it the universal set:



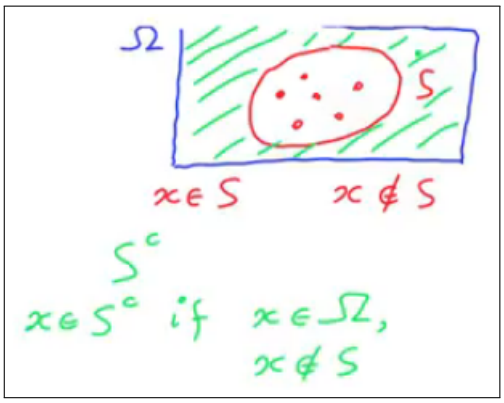
So having fixed a universal set, we will only consider smaller sets that lie inside that big universal set:



And once we have a universal set, we can talk about the collection of all objects, or elements, that belong to our universal set, but do not belong to the set S. So that would be everything outside the set S:

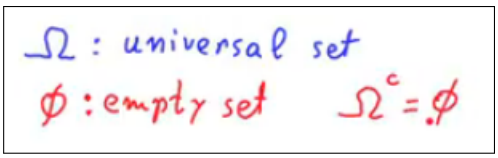


Everything outside the set S, we denote it this way, S^c, and we call it the *complement of the set S*. And it is defined formally as follows — an element belongs to S^c if $x \in \Omega$ and also $x \notin S$:

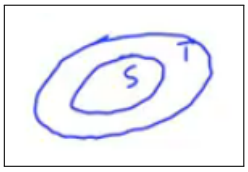


Notice that if we take the complement of the complement — that is, anything that does not belong to the green set — we get back the red set. So what this is saying is that *the complement of the complement of a set is the set itself*: $(S^c)^c = S$.

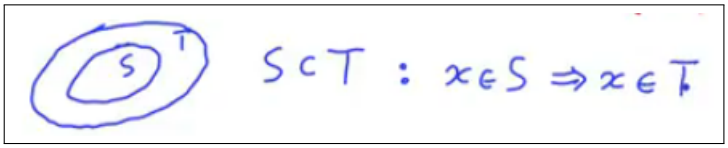
Another set of particular interest is the so-called *empty set*. The empty set is a set that contains no elements. In particular, if we take the complement of the universal set — well, since the universal set contains everything, there is nothing in its complement, so its complement is going to be the empty set, $\Omega^c = \varnothing$:



Finally, one more piece of notation. Suppose that we have two sets, and one set is bigger than the other. So S is the small set here, and T is the bigger set:

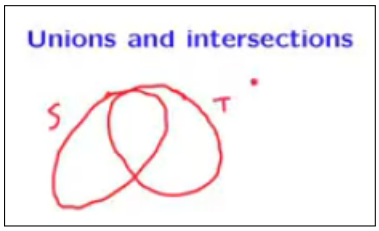


We denote this relation by writing this expression, $S \subset T$, which we read as follows: S is a subset of the set T. And what that means is that if $x \in S$, then such an $x \in T$:

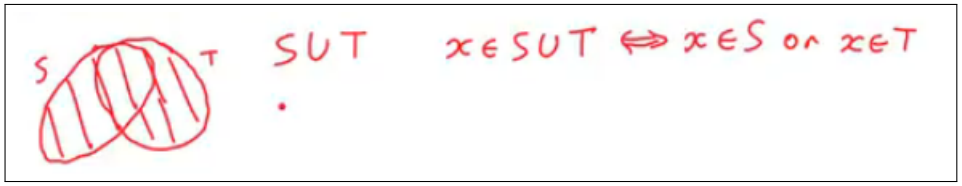


Note that when $S \subset T$, there is also the possibility that $S = T$. One word of caution here — the notation that we're using here is the same as what in some textbooks is written this way: $S \subseteq T$, that is, S is a subset of T, but can also be equal to T. We do not use this notation, but that's how we understand it. That is, we allow for the possibility that the subset is equal to the larger set.

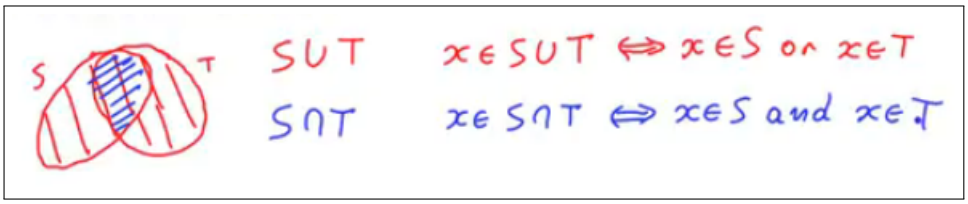
Now when we have two sets, we can talk about their *union* and their *intersection*. Let's say that this is set S, and this is set T:



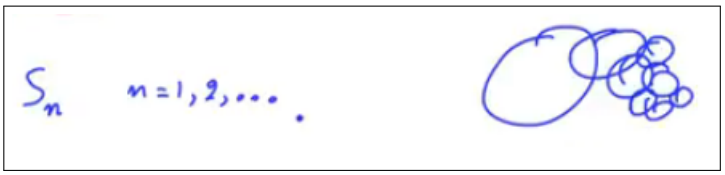
The union of the two sets consists of all elements that belong to one set or the other, or in both. The union is denoted this way, $S \cup T$, and the formal definition is that some element belongs to the union if and only if this element belongs to one of the sets, or it belongs to the other one of the sets, $x \in (S \cup T) \iff (x \in S) \vee (x \in T)$:



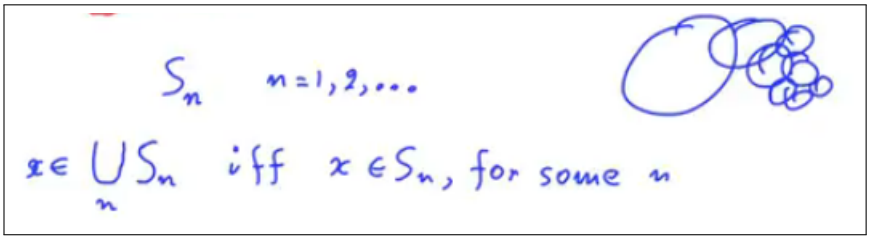
We can also form the intersection of two sets, which we denote this way, $S \cap T$, and which stands for the collection of elements that belong to both of the sets. So formally, an element belongs to the intersection of two sets if and only if that element belongs to both of them. So x must be an element of S, and it must also be an element of T, $x \in (S \cap T) \iff (x \in S) \wedge (x \in T)$:



By the way, we can also define unions and intersections of more than two sets, even of infinitely many sets. So suppose that we have an *infinite collection of sets*. Let's denote them by S_n. So n ranges over, let's say, all of the positive integers. So pictorially, you might think of having one set, another set, a third set, a fourth set, and so on, and we have an infinite collection of such sets:

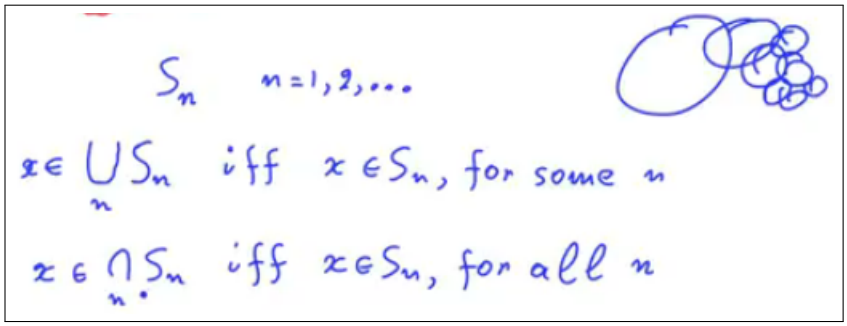


Given this infinite collection, we can still define their union to be the *set of all elements that belong to one of those sets* S_n that we started with:



That is, an element is going to belong to that union if and only if this element belongs to some of the sets that we started with.

We can also define the intersection of an infinite collection of sets. We say that an element x belongs to the intersection of all these sets if and only if x belongs to Sn for all n. So if x belongs to each one of those Sn's, then we say that x belongs to their intersection:



Set operations satisfy certain basic properties:

$$
\begin{aligned}
S \cup T &= T \cup S, & S \cup (T \cup U) &= (S \cup T) \cup U, \\
S \cap (T \cup U) &= (S \cap T) \cup (S \cap U), & S \cup (T \cap U) &= (S \cup T) \cap (S \cup U), \\
(S^c)^c &= S, & S \cap S^c &= \emptyset, \\
S \cup \Omega &= \Omega, & S \cap \Omega &= S.
\end{aligned}
$$

One of these we already discussed: $(S^c)^c = S$. This property, for example, is pretty clear, $S \cup T = T \cup S$, The union of a set with another set is the same as the union if you consider the two sets in different orders.

If you take the union of three sets, you can do it by forming, first, the union of these two sets, and then the union with this one; or, do it in any alternative order, $S \cup (T \cup U) = (S \cup T) \cup U$. Both expressions are equal. Because of this, we do not really need the parentheses, and we often write just this expression here, $S \cup T \cup U$, which is the same as this one.

And the same would be true for intersections. That is, the intersection of three sets is the same no matter how you put parentheses around the different sets: $S \cap (T \cap U) = (S \cap T) \cap U = S \cap T \cap U$.

Now if you take a union of a set with a universal set, you cannot get anything bigger than the universal set, so you just get the universal set: $S \cup \Omega = \Omega$.

On the other hand, if you take the intersection of a set with the universal set, what is left is just the set itself: $S \cap \Omega = S$.

Perhaps the more complicated properties out of this list is this one, $S \cap (T \cup U) = (S \cap T) \cup (S \cap U)$, and this one, $S \cup (T \cap U) = (S \cup T) \cap (S \cup U)$, which are sort of a distributive property of intersections and unions. And I will let you convince yourselves that these are true.

The way that you verify them is by proceeding logically. If $x \cap (T \cup U)$, then x must be an element of S, and it must also be an element of either T or U. Therefore, it's going to belong either to this set — it belongs to S, and it also belongs to T — or it's going to be an element of that set — it belongs to S, and it belongs to U. So this argument shows that this set here is a subset of that set. Anything that belongs here belongs there. Then you need to reverse the argument to convince yourself that anything that belongs here belongs also to the first set, and therefore, the two sets are equal.

Here, I'm using the following fact — that if S is a subset of T, and T is a subset of S, this implies that the two sets are equal. And then you can use a similar argument to convince yourselves about this equality, as well.

So this is it about basic properties of sets. We will be using some of these properties all of the time without making any special comment about them.

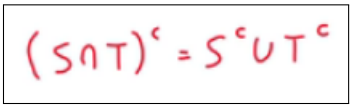
3.2.3 De Morgan's laws

Video: De Morgan's laws (transcripts)

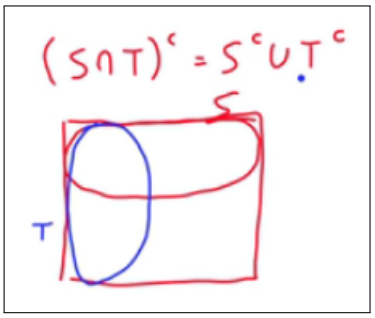
De Morgan's laws: $\left(\bigcup_n S_n\right)^c = \bigcap_n S_n^c, \qquad \left(\bigcap_n S_n\right)^c = \bigcup_n S_n^c$

We will now discuss De Morgan's laws that are some very useful *relations between sets and their complements*. One of the De Morgan's laws takes this

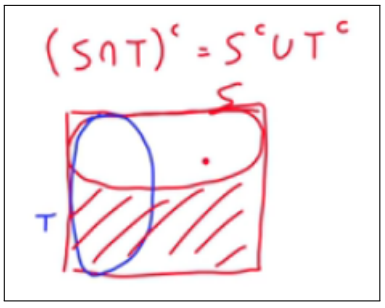
form. If we take the intersection of two sets and then take the complement of this intersection, what we obtain is the union of the complements of the two sets:



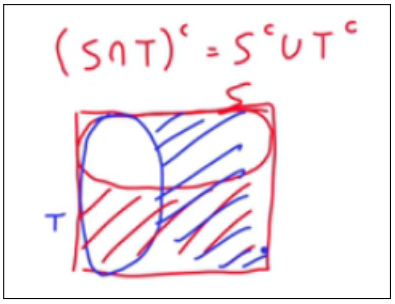
Pictorially, here is the situation. We have our universal set. Inside that set, we have a set, S, which is this one. And we have another set, T, which is this one:



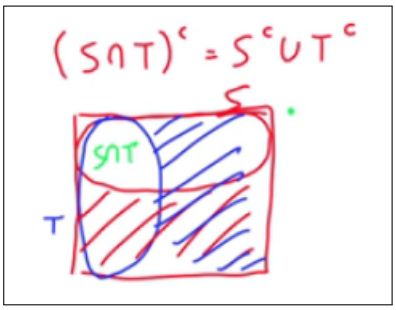
Let us look at this side. The complement of S is this part of the diagram:



The complement of T is this part of the diagram:

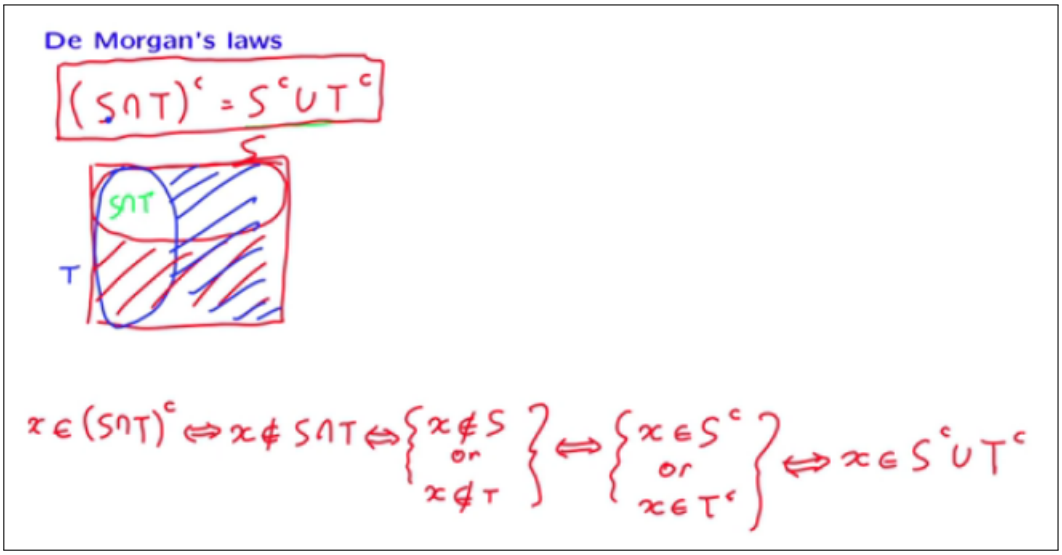


What is left? What is left is just this region here, which is the intersection of S with T:



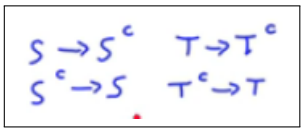
So anything that does not belong here belongs to the intersection. This means that the complement of the intersection is everything out there, which is the set.

If you're not convinced by this pictorial proof, let us go through an argument that is a little more formal. What does it take for an element to belong to the first set? In order to belong to that set, x belongs to the complement of S intersection T. This is the same as saying that x does not belong to the intersection [of] S with T. What does that mean? Since it is not in the intersection, this is the same as saying that x does not belong to S or x does not belong to T. But this is the same as saying that x belongs to the complement of S or x belongs to the complement of T. And this is equivalent to saying that x belongs to the union of the complement of S with the complement of T:

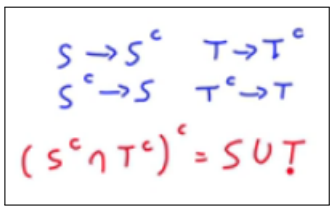


So this establishes this first De Morgan's law. There's another De Morgan's law, which is obtained from this one by a syntactic substitution. We're going to play the following trick. Wherever we see an S, we're going to replace it by S complement. And wherever we see an S complement, we will replace it with

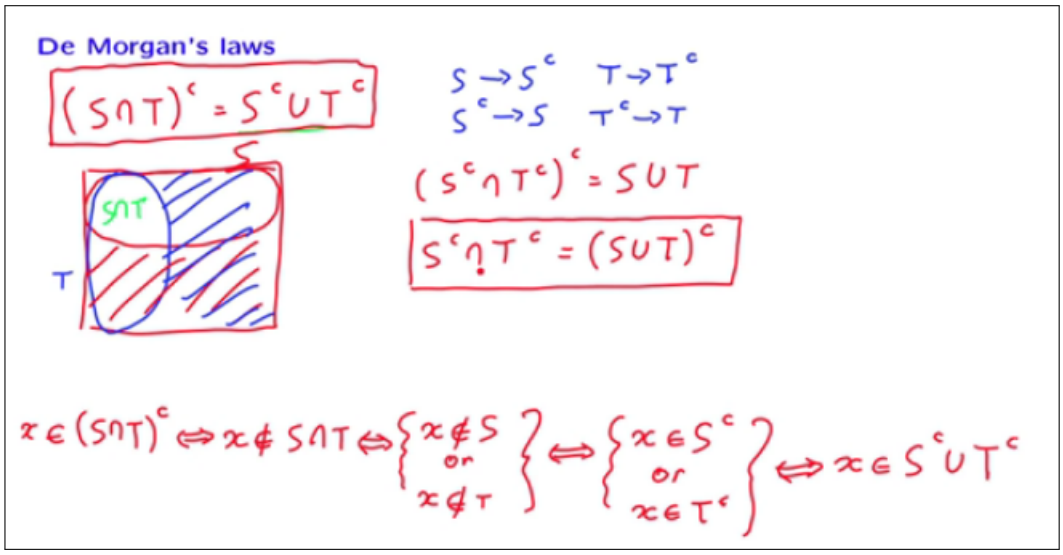
an S. And similarly, whenever we see a T, we'll replace it by T complement. And when we see a T complement, we will replace it by T:



So doing this syntactic substitution, what we obtain is S complement intersection with T complement — everything gets complemented — is the same as S union T:

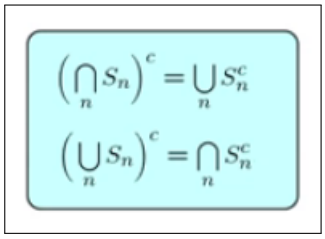


Now, let us take complements of both sides. The complement of a complement is the set itself. So we obtain this. And now, we take the complement of the other side, which is this one. And this is the second De Morgan's law. It tells us that the complement of a union is the same as the intersection of the complements:



We derived it from the first De Morgan's law by a syntactic substitution. If you're not convinced, it would be useful for you to go through an argument of this kind to show that something is an element of this set if and only if it is an element of that set as well.

Finally, it turns out that *De Morgan's laws are valid when we take unions or intersections of more than two sets*. There is a more general form. And the general form is as follows, an analogy with this one:



If we have a collection of sets, S_n, perhaps an infinite collection, we take the intersection of those sets and then the complement, what that is is the union of the complements.

So this is analygous to this law. And this law extends to this one: if we have the union of certain sets and we take the complement of the union, what we obtain is the intersection of the complements.

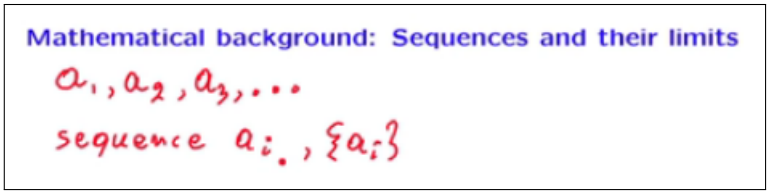
We will have many occasions to use De Morgan's laws. They're actually very useful. They allow us, in general, to go back and forth between unions and intersections.

3.2.4 Sequences and their limits

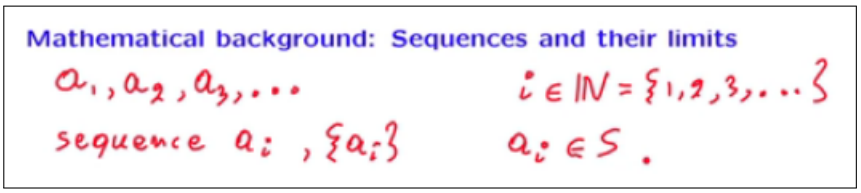
Video: Sequences and their limits (transcripts)

In this segment, we will discuss what a sequence is and what it means for a sequence to converge. So a *sequence is some collection of elements that are coming out of some set, and that collection of elements is indexed by the natural numbers*.

We often use the notation, and we say that we have a sequence a_i, or sometimes we use the notation that we have a sequence of this kind to emphasize the fact that it's a sequence and *not just a single number*:

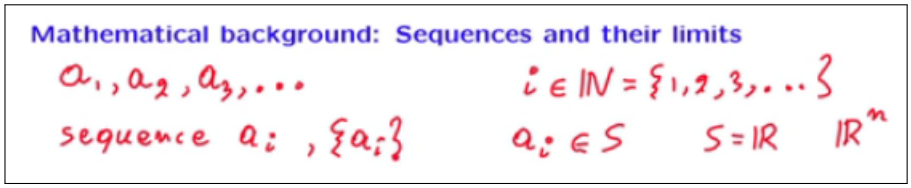


And what we mean by this is that we have i, an index that runs over the natural numbers, which is the set of positive integers, and each a_i is an element of some set:

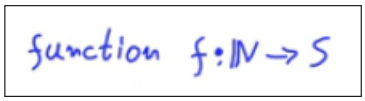


In many cases, the set is going to be just the real line, in which case we're dealing with a sequence of real numbers, $\mathbb{R}$.

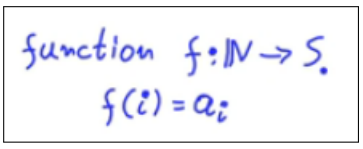
But it is also possible that the set over which our sequence takes values is Euclidean space n-dimensional space, $\mathbb{R}^n$, in which case we're dealing with a sequence of vectors. But it also could be any other kind of set.



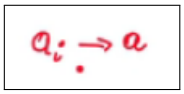
Now, the definition that I gave you may still be a little vague. You may wonder how a mathematician would define formally a sequence. Formally, what a sequence is, is just a *function that, to any natural number, associates an element of S*:



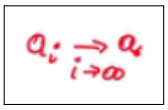
In particular, if we evaluate the $f(i)$, this gives us the i-th element of the sequence:



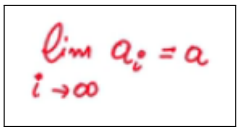
So that's what a sequence is. Now, about sequences, we typically care *whether a sequence converges to some number a*, and we often use this notation:



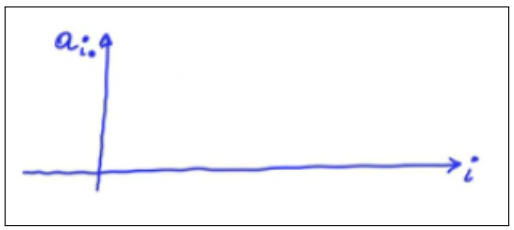
But to make it more precise, you also add this notation here. And we read this as saying that as i converges to infinity, the sequence ai converges to a certain number a:



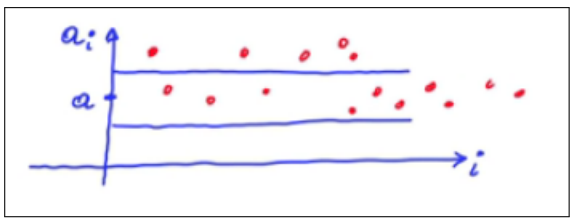
A more formal mathematical notation would be the limit as i goes to infinity of a_i is equal to a certain number, a:



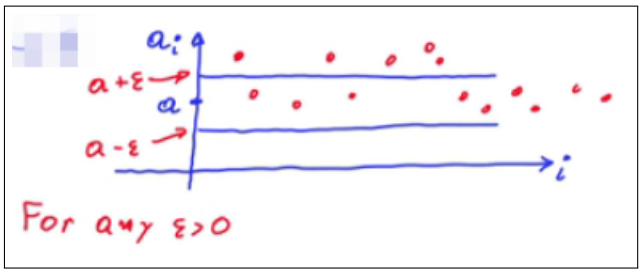
But what exactly does this mean? What does it mean for a sequence to converge? What is the formal definition? It is as follows. Let us plot the sequence as a function of i. So this is the i-axis, and here we plot entries of a_i:



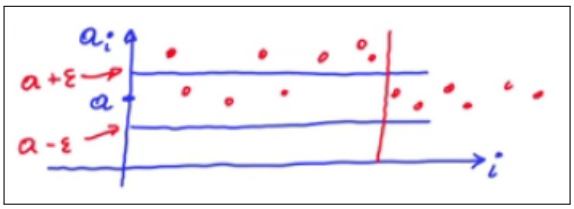
For a sequence to converge to a certain number a, we need to the following to happen. If we draw a small band around that number a, what we want is that the elements of the sequence, as i increases, eventually get inside this band and stay inside that band forever:



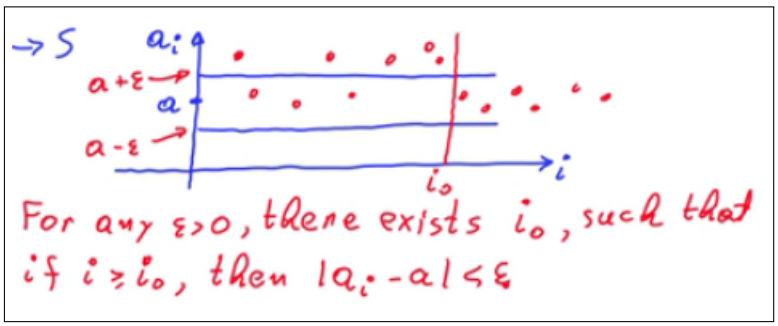
Now, let us turn this into a more precise statement. What we mean is the following. If I give you some positive number epsilon, ϵ, and I'm going to use that positive ϵ to define a band around the number a. So it's this band here:



If I give you a positive number ϵ, and therefore, this way, have defined a certain band, there exists a time after which the entries will get the inside the band. In this picture, it would be this time:

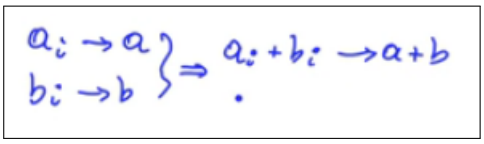


So there exists a time — let's call that time i_0 — so i_0 is here such that after that time, what we have is that the element of the sequence is within ϵ of a:

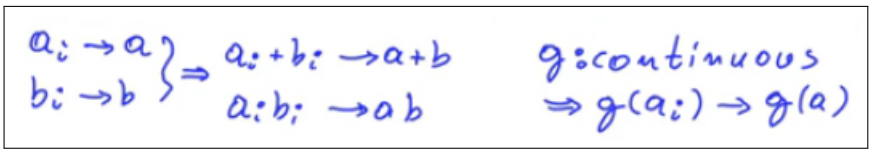


So this is the formal definition of convergence of a sequence to a certain number a. The definition may look formidable and difficult to parse, but what it says in plain English is pretty simple. No matter what kind of band I take around my limit a, eventually, *the sequence will be inside this band and will stay inside there*.

Convergence of sequences has some very nice properties that you're probably familiar with. For example, if I tell you that a certain sequence converges to a number a and another sequence converges to a number b, then we will have that a_i plus b_i, which is a new sequence — the ith element of the sequence is this sum — will converge to a plus b.



Or similarly, a_i times b_i, which is another sequence, converges to a times b. And if, in addition, g is a continuous function, then $g(a_i)$ will converge to $g(a)$.



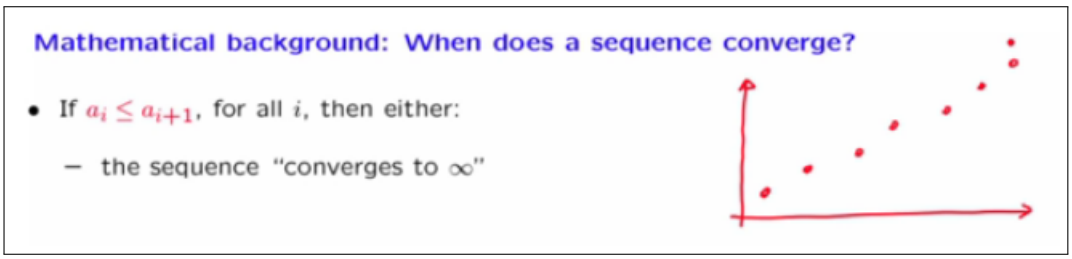
So for example, if the a_i converge to a, then the sequence a_i^2 is going to converge to a^2.

3.2.5 When does a sequence converge?

Video: When does a sequence converge? (transcripts)

So we looked at the formal definition of what it means for a sequence to converge, but as a practical matter, how can we tell whether a given sequence converges or not? There are two criteria that are the most commonly used for that purpose, and it's useful to be aware of them.

The first one deals with the case where we have a *sequence of numbers that keep increasing, or at least, they do not go down*. In that case, those numbers may go up forever without any bound, so if you look at any particular value, there's going to be a time at which the sequence has exceeded that value. In that case, we say that the sequence converges to infinity:



But if this is not the case, if it does not converge to infinity, which means that the entries of the sequence are bounded — they do not grow arbitrarily large — then, in that case, it is guaranteed that the sequence will converge to a certain number:

Mathematical background: When does a sequence converge?

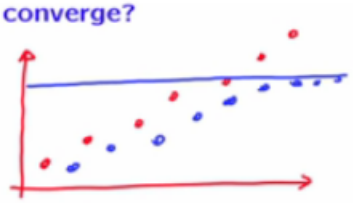
- If $a_i \leq a_{i+1}$, for all i, then either:
  - the sequence "converges to ∞"
  - the sequence converges to some real number a

This is not something that we will attempt to prove, but it is a useful fact to know.

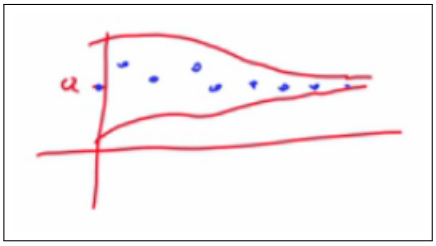
Another way of establishing convergence is to *derive some bound on the distance or our sequence from the number that we suspect to be the limit*. If that distance becomes smaller and smaller, if we can manage to bound that distance by a certain number and that number goes down to 0, then it is guaranteed that since this distance goes down to 0, that the sequence, ai, converges to a:

Mathematical background: When does a sequence converge?

- If $a_i \leq a_{i+1}$, for all i, then either:
  - the sequence "converges to ∞"
  - the sequence converges to some real number a
- If $|a_i - a| \leq b_i$, for all i, and $b_i \to 0$, then $a_i \to a$



And there's a variation of this argument, which is the so-called *sandwich argument*, and it goes as follows. If we have a certain sequence that converges to some number, a, and we have another sequence that converges to that same number, a, and our sequence is somewhere in-between, then our sequence must also converge to that particular number, a:



So these are the usual ways of quickly saying something about the convergence of a given sequence, and we will be often using this type of argument in this class, but without making a big fuss about them, or without even referring to these facts in an explicit manner.

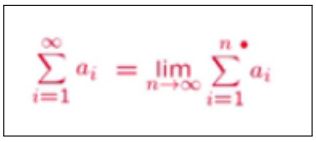
3.2.6 Infinite series

Video: Infinite series (transcripts)

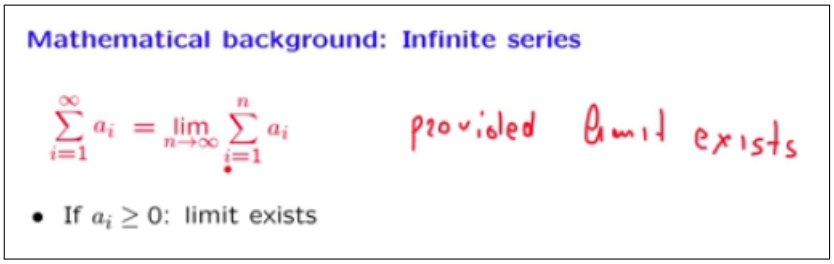
This will be a short tutorial on infinite series, their definition and their basic properties.

What is an infinite series? We're given a sequence of numbers a_i, indexed by i, where i ranges from 1 to ∞. So it's an infinite sequence. And we want to add the terms of that sequence together. We denote the resulting sum of that infinity of terms using this notation:

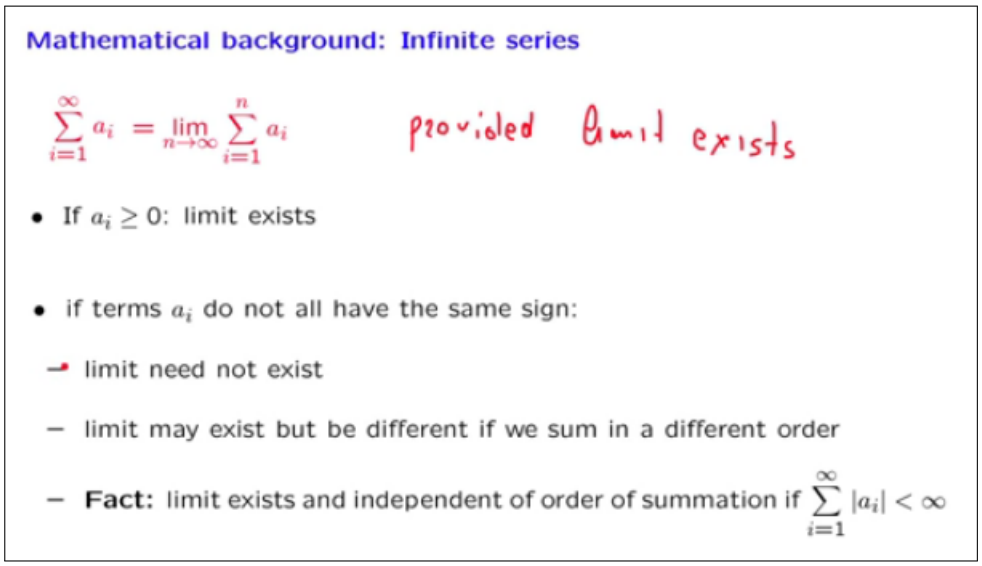
But what does that mean exactly? What is the formal definition of an infinite series? Well, the infinite series is defined as the limit, as n goes to infinity, of the finite series in which we add only the first n terms in the series:



However, this definition makes sense only as long as the limit exists. This brings up the question, when does this limit exist? The nicest case arises when all the terms are non-negative. If all the terms are non-negative, here's what's happening. We consider the partial sum of the first n terms. And then we increase n. This means that we add more terms. So the partial sum keeps becoming bigger and bigger. The sequence of partial sums is a *monotonic sequence*. Now *monotonic sequences always converge either to a finite number or to infinity*. In either case, this limit will exist, and therefore, the series is well defined:



The situation is more complicated if the terms a_i can have different signs:



In that case, it's possible that the limit does not exist. And so the series is not well defined.

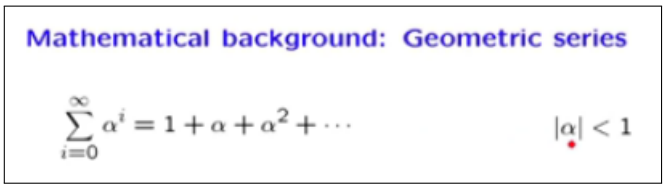
The more interesting and complicated case is the following. It's possible that this limit exists. However, if we *rearrange the terms in the sequence, we might get a different limit*. When can we avoid those complicated situations? We can avoid them if it turns out that the sum of the absolute value of the numbers sums to a finite number.

Now this series that we have here is an infinite series in which we add non-negative numbers. And by the fact that we mentioned earlier, this infinite series is always well defined. And it's going to be either finite or infinite. If it turns out to be finite, then the original series is guaranteed to be well defined, to have a finite limit when we define it that way, and furthermore, that finite limit is the same even if we rearrange the different terms, if we rearrange the sequence with which we sum the different terms.

3.2.7 The geometric series

Video: The geometric series (transcripts)

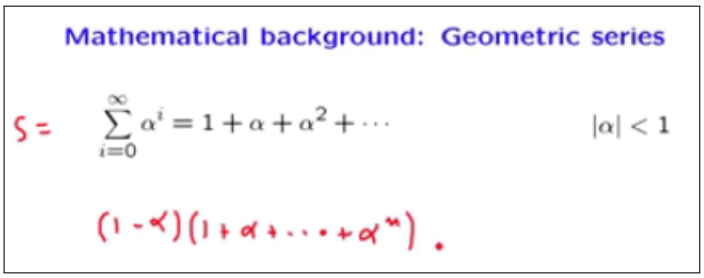
One particular series that shows up in many applications, examples, or problems is the geometric series. [In] the geometric series, we are given a certain number, alpha, and we want to sum all the powers of alpha, starting from the 0th power, which is equal to 1, the first power, and so on, and this gives us an infinite series. It's the sum of alpha to the i where i ranges from 0 to infinity:



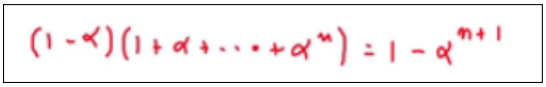
Now, for this series to converge, we need subsequent terms, the different terms in the series, to become smaller and smaller. And for this reason, we're going to make the assumption that the number alpha is less than 1 in magnitude, which implies that consecutive terms go to zero.

Let us introduce some notation. Let us denote the infinite sum by S, and we're going to use that notation shortly. One way of evaluating this series is to start from an algebraic identity, namely the following.

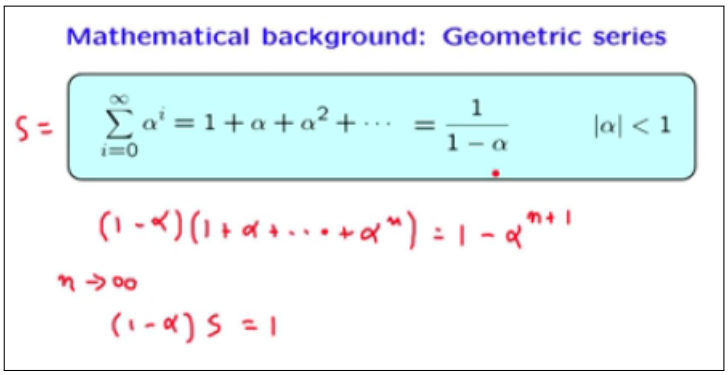
Let us take 1 minus alpha and multiply it by the terms in the series, but going only up to the term alpha to the n. So it's a finite series:



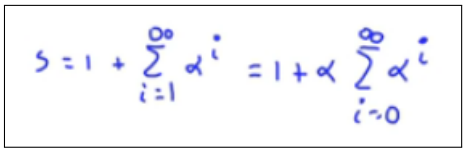
We do this multiplication, we get a bunch of terms, we do the cancellations, and what is left at the end is 1 minus alpha to the power n plus 1:



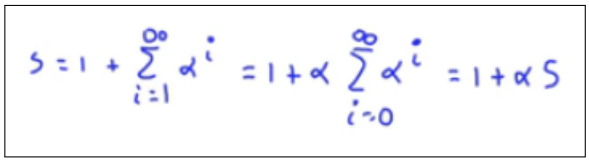
What we do next is we take the limit as n goes to infinity. On the left hand side, we have the term 1 minus alpha, and then the limit of this finite series is by definition the infinite series, which we're denoting by s. On the right hand side, we have the term 1. How about this term? Since alpha is less than 1 in magnitude, this converges to 0 as alpha goes to infinity, so that term disappears. We can now solve this relation, and we obtain that s is equal to 1 over 1 minus alpha, and this is the formula for the infinite geometric series:



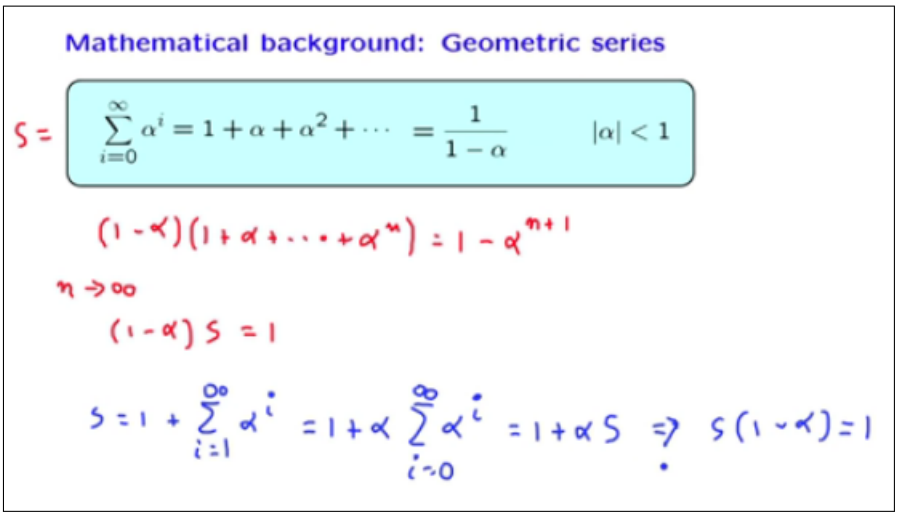
There's another way of deriving the same result, which is interesting, so let us go through it as well. The infinite geometric series has one first term and then the remaining terms, which is a sum for i going from 1 to infinity of alpha to the i. Now, we can take a factor of alpha out of this infinite sum and write it as 1 plus alpha, the sum of alpha to the i, but because we took out one factor of alpha, here, we're going to have smaller powers. So now the sum starts from 0 and goes up to infinity:



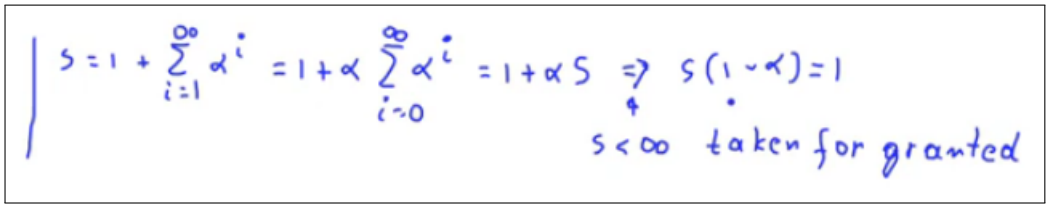
Now, this is just 1 plus alpha times s because here, we have the infinite geometric series:



Therefore, if we subtract alpha s from both sides of this equality, we get s times 1 minus alpha equal to 1. And now by moving 1 minus alpha to the denominator, we get again the same expression:



So this is an alternative way of deriving the same result. However, there's one word of caution. In this step, we subtracted alpha s from both sides of the equation. And in order to do that, this is only possible if we take for granted that s is a finite number. So this is taken for granted in order to carry out this derivation:

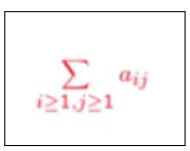


This is to be contrasted with the first derivation, in which we didn't have to make any such assumption. So strictly speaking, for this derivation here to be correct, we need to have some independent way of verifying that s is less than infinity. But other than that, it's an interesting algebraic trick.

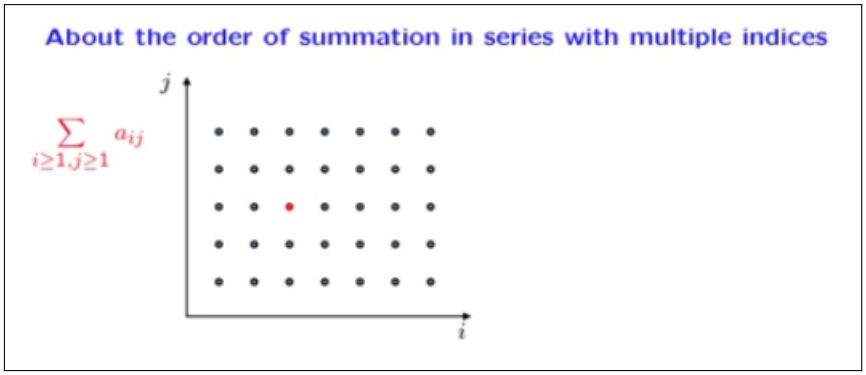
3.2.8 About the order of summation in series with multiple indices

Video: About the order of summation in series with multiple indices (transcripts)

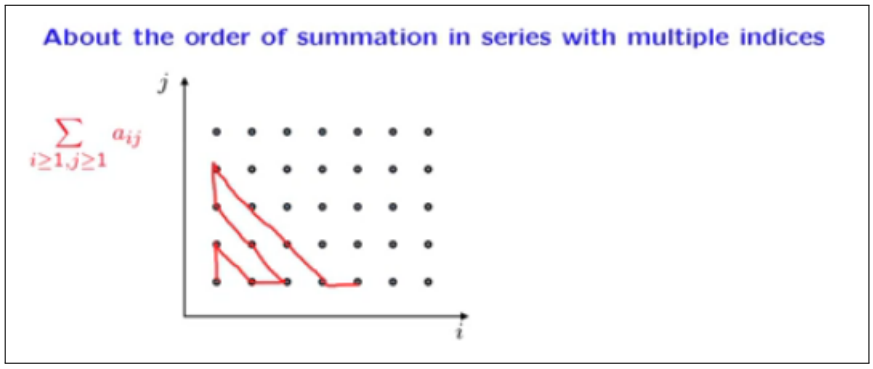
We now continue our discussion of infinite series. Sometimes we have to deal with series where the terms being added are indexed by multiple indices, as in this example here:



We're given numbers, aij, and i ranges over all the positive integers. j also ranges over all the positive integers. So what does this sum represent? We can think of it as follows. We have here a two-dimensional grid that corresponds to all the pairs (i,j):



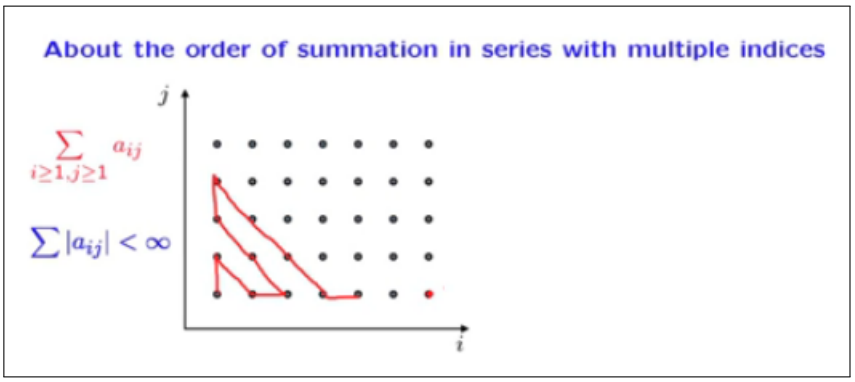
And in essence, each one of those points corresponds to one of the terms that we want to add. So we can sum the different terms in some arbitrary order. Let's say we start from here. Take that term, add this term, then add this term here, then add this term, then the next term, next term, and so on. And we can keep going that way, adding the different terms according to some sequence:



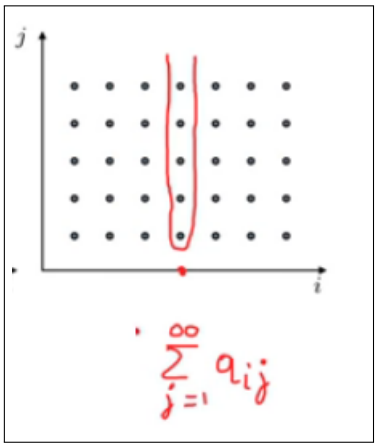
So essentially, what we're doing here is we're taking this two-dimensional grid and arranging the terms associated with that grid, in some particular linear order. And we're summing those terms in sequence. As long as this sum converges to something as we keep adding more and more terms, then this double series will be well defined.

Notice, however, we can add those terms in many different orders. And *in principle, those different orders might give us different kinds of results*.

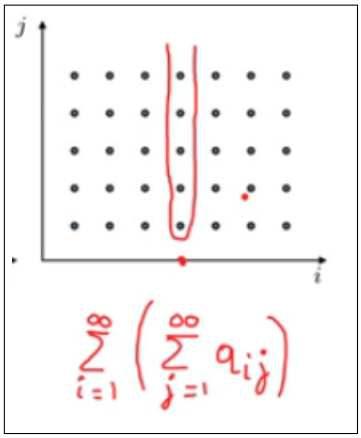
On the other hand, as long as the sum of the absolute values of all the terms turns out to be finite, then the particular order in which we're adding the different terms will turn out that it doesn't matter:



There's another way that we can add the terms together, and this is the following. Let us consider fixing a particular choice of i, and adding all of the terms that are associated with this particular choice of i, as j ranges from 1 to infinity:

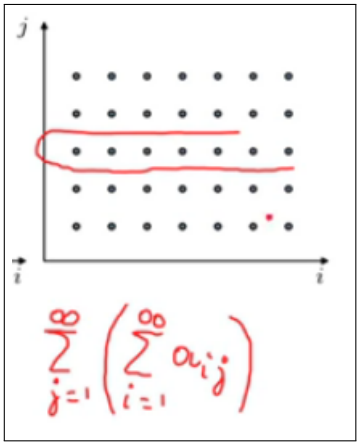


So what we're doing is we're taking the summation from j equal to 1 to infinity, while keeping the value of i fixed. We do this for every possible i:

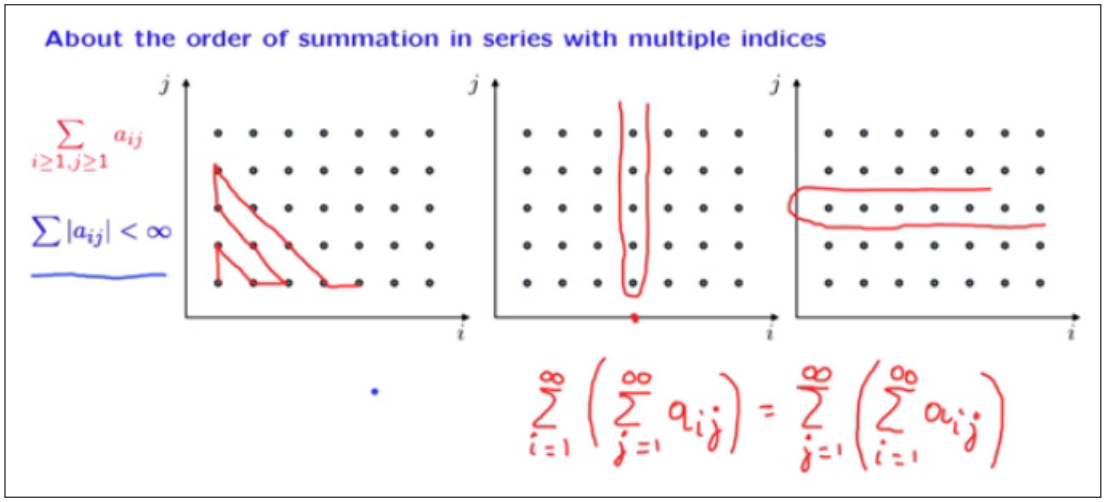


So for every possible i, we're going to get a particular number. And then we take the numbers that we obtain for the different choices if i, so i ranges from 1 to infinity. And we add all those terms together. So this is one particular order, one particular way of doing the infinite summation.

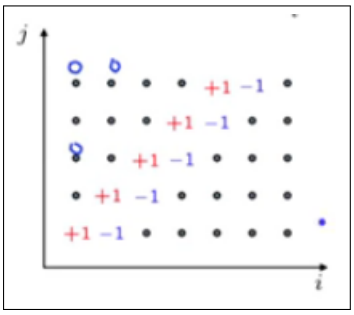
Now, why start with the summation over j's while keeping i fixed? There's no reason for that. We could also carry out the summation by fixing a particular choice of j and summing over all i's. So now it is i that ranges from 1 to infinity. And we look at this infinite sum. This is the infinite sum of those terms. We obtain one such infinite sum for every choice of j. And then we take that sum that we obtain for any particular choice of j, and add over the different possible values of j. So j goes from 1 to infinity:



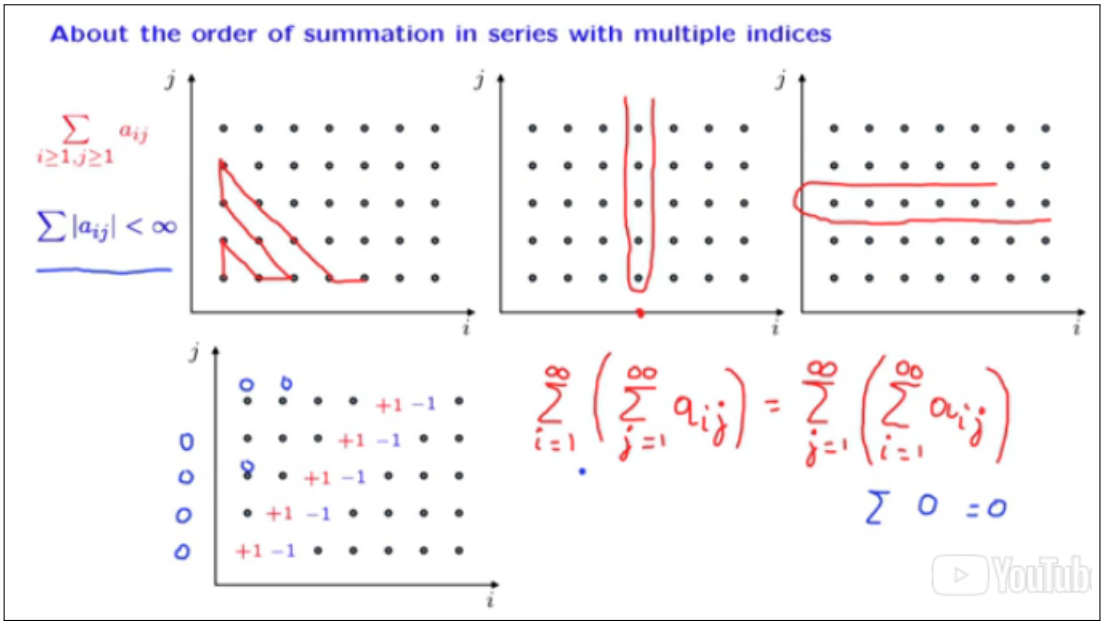
This is a different way of carrying out the summation. And these are going to give us the same result, and the same result that we would also obtain if we were to add the terms in this particular order, as long as *the double series is well-defined*, in the following sense: *If we have a guarantee that the sum of the absolute values of those numbers is finite, no matter which way we add them, then it turns out that we can use any particular order to add the terms in the series*. We're going to get the same result. And we can also carry out the double summation by doing — by adding over one index at a time:



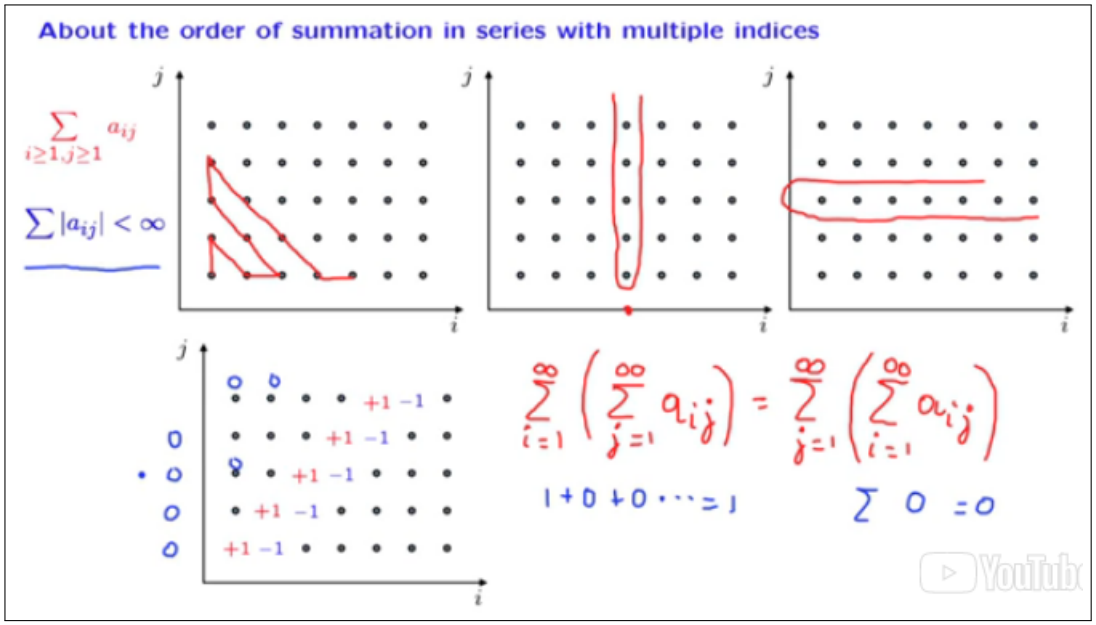
A word of caution: this condition is not always satisfied. And in those cases, strange things can happen. Suppose that the sequences we're dealing with, the aij's, take those particular values indicated in this picture. And all the remaining terms, the aij's associated with the other dots, are all 0's. So all these terms out there will be 0's:



If we carry out the summation by fixing a j and adding over all i's, what we get here is 0, and a 0, and a 0, and a 0. That's because in each row we have a 1 and a minus 1, which cancel out and give us 0's. So if we carry out the summation in this manner, we get a sum of 0's, which is 0:



But if we carry out the summation in this order, fix an i, and then add over all j's, the first term that we get here is going to be 1, because in this column, this is the only non-zero number. And then in the remaining columns, as we add the terms, we're going to get 0's, and 0's, and so on. And so if we carry out the summation in this way, we obtain a 1:

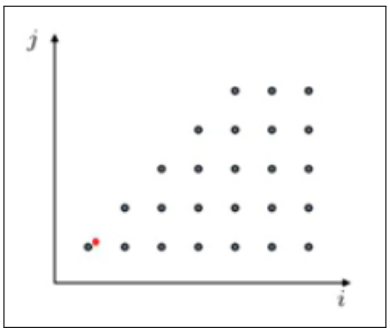


So this is an example that shows you that the order of summation actually may matter. In this example, the sum of the absolute values of all of the terms that are involved is infinity, because we have infinitely many plus or minus 1's, so this condition here is not satisfied in this example.

Let us now consider the case where we want to add the terms of a double sequence, but over a limited range of indices as in this example, where we have coefficients aij, which we want to add, but only for those i's and j's for which j is less than or equal to i:

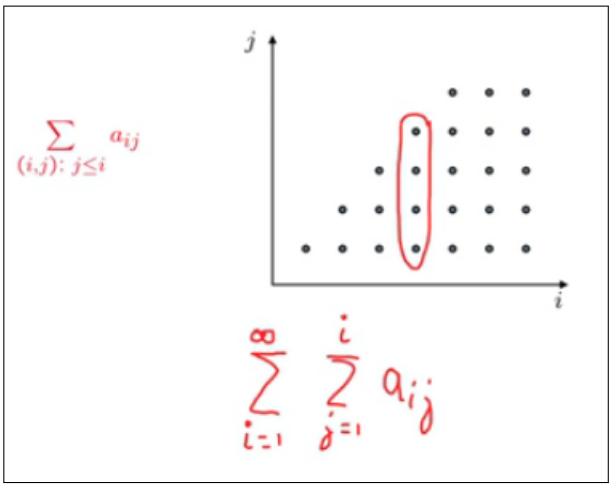
$$\sum_{(i,j):\, j \le i} a_{ij}$$

Graphically, this means that we only want to consider the pairs shown in this picture:



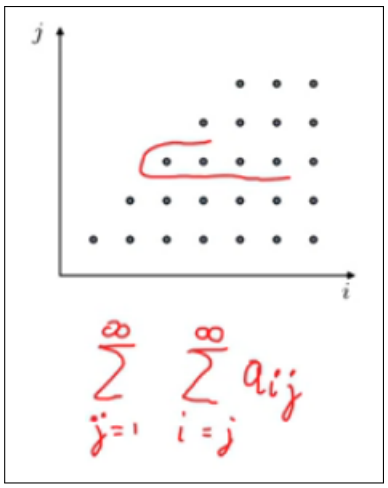
So these points here correspond to i,j pairs for which i is equal to j. Terms on the right, or points to the right, correspond to i,j pairs for which i is at least as large as j.

We can carry out this summation in two ways. One way is the following. We fix a value of i, and we consider all of the corresponding terms, that correspond to different choices of j. But we only go up to the point where i is equal to j. This is the largest term. So what are we doing here? We're taking the coefficients aij, and we are adding over all j's, starting from 1, which corresponds to this term. And j goes up to the point where it becomes equal to i. We do this for every value of i. And so we get a number for the sum of each one of the columns, and then we add those numbers together. So we're adding over all i's, and i ranges from 1 up to infinity:

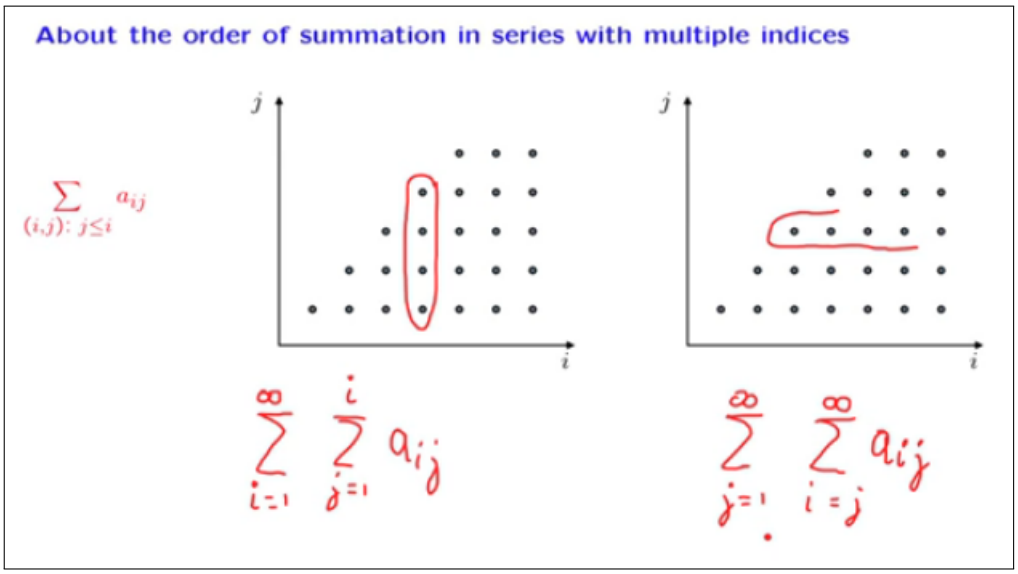


This is one way of carrying out the summation.

Alternatively, we could fix a value of j, and consider doing the summation over all choices of i. So this corresponds to the sum over all choices of i, from where? The smallest term, the first term, happens when i is equal to the value of j. And then we have larger choices of i, numbers for which i is bigger than the corresponding value of j. And so i ranges from j all the way to infinity. And this is the sum over one of the rows in this diagram. We do this for every j. We get a result, and then we need to add all those results together. So we're summing for all j's from 1 up to infinity:



So these are two different ways that we can evaluate this series associated with a double sequence. We can either add over all j's first and then over i's, or we can sum over all i's first, and then over all j's:



The two ways of approaching this problem, this summation, should give us the same answer. And this is going to be, again, subject to the usual qualification. As long as the *sum of the absolute values of the terms that we're trying to add is less than infinity* — if this condition is true, then the two ways of carrying out the summation are equal, and they're just two different alternatives.

In this segment, we discuss the following fact. We are given a collection of real numbers a_{ij}, indexed by positive integers i and j. When we write $\sum_{i\geq 1, j\geq 1} a_{ij}$, what we mean is the limit of the sum of the terms a_{ij}, where the summation is carried out by arranging these terms in a sequence and then adding. How do we know that different orderings will give the same result?

There is the following important fact. If the infinite sum $\sum_{i\geq 1, j\geq 1} |a_{ij}|$ is *finite for some particular order* in which the summation is carried out, then the infinite series $\sum_{i\geq 1, j\geq 1} a_{ij}$ is well-defined and takes a value which is the same, no matter in which order the terms are added.

Furthermore, in that case, we have

$$\sum_{i\geq 1, j\geq 1} a_{ij} = \sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} a_{ij} \right) = \sum_{j=1}^{\infty} \left(\sum_{i=1}^{\infty} a_{ij} \right)$$

As a special case, if you are dealing with a finite sequence and sum, you can always interchange the order of summation.

On the other hand, if $\sum_{i\geq 1, j\geq 1} |a_{ij}| = \infty$, and if we carry out the summation of the terms in the expression $\sum_{i\geq 1, j\geq 1} a_{ij}$ in different orders, we may get either different values, or failure to converge, in which case the expression $\sum_{i\geq 1, j\geq 1} a_{ij}$ is not well-defined. This is demonstrated by the example in the video.

Note that perhaps this was not stated explicitly at that point in the video, but the example deals with an infinite sequence a_{ij}, which follows a certain pattern. The diagram shows only part of the sequence, but you should think

of it as continuing forever with the pattern indicated in the diagram.

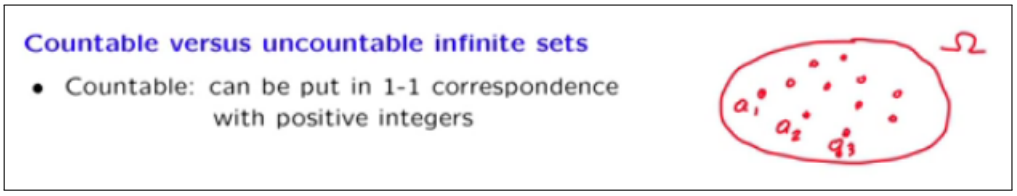
3.2.9 Countable and uncountable sets

Video: Countable and uncountable sets (transcripts)

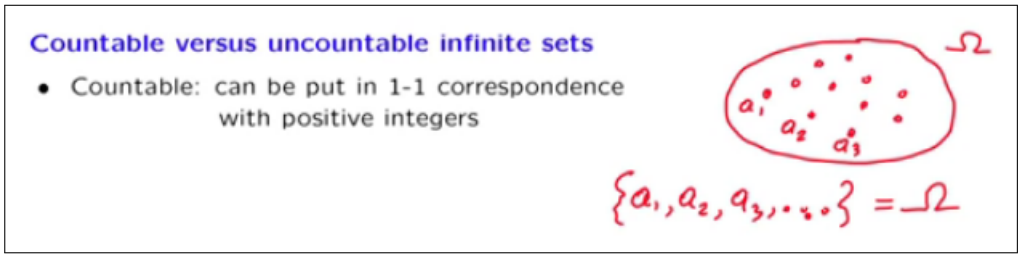
Probability models often involve *infinite sample spaces*, that is, infinite sets. But not all sets are of the same kind.

Some sets are *discrete* and we call them **countable**, and some are *continuous* and we call them **uncountable**. But what exactly is the difference between these two types of sets? How can we define it precisely?

Well, let us start by first giving a definition of what it means to have a countable set. A set will be called **countable** *if its elements can be put into a 1-to-1 correspondence with the positive integers*. This means that we look at the elements of that set, and we take one element — we call it the first element. We take another element — we call it the second. Another, we call the third element, and so on. And this way we will eventually exhaust all of the elements of the set, so that each one of those elements corresponds to a particular positive integer, namely the index that appears underneath:



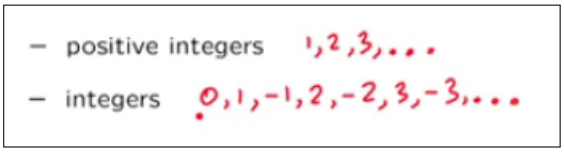
More formally, what's happening is that we take elements of that set that are *arranged in a sequence*. We look at the set, which is the entire range of values of that sequence, and we want that sequence *to exhaust the entire set omega*. Or in other words, in simpler terms, we want to be able to arrange all of the elements of omega in a sequence:



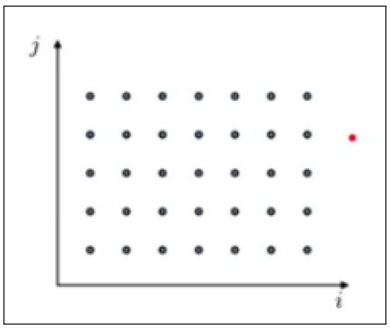
So what are some examples of countable sets? In a trivial sense, the *positive integers* themselves are countable, because we can arrange them in a sequence. This is almost tautological, by the definition.

For a more interesting example, let's look at the set of *all integers*. Can we arrange them in a sequence? Yes, we can, and we can do it in this manner,

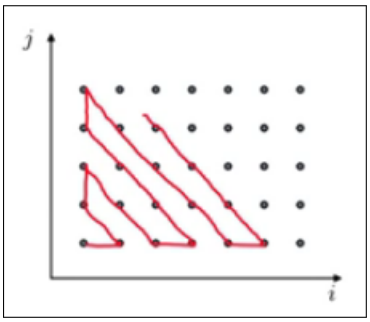
where we alternate between positive and negative numbers. And this way, we're going to cover all of the integers, and we have arranged them in a sequence.



How about the set of all pairs of *positive integers*? This is less clear. Let us look at this picture:

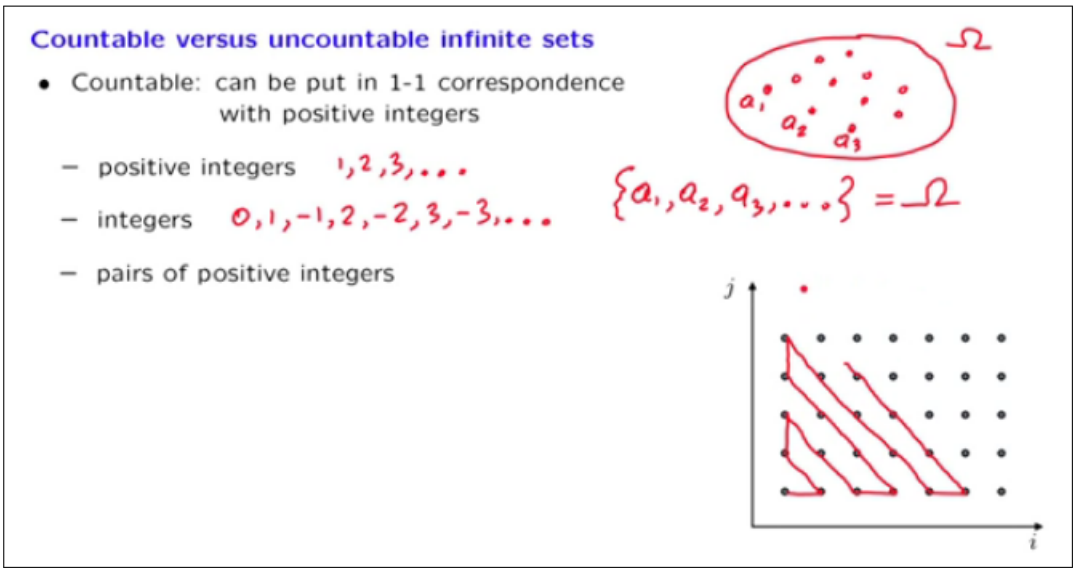


This is the set of all pairs of positive integers, which we understand to continue indefinitely. Can we arrange this sets in a sequence? It turns out that we can. And we can do it by tracing a path of this kind:



So you can probably get the sense of how this path is going. And by continuing this way, over and over, we're going to cover the entire set of all pairs of positive integers. So we have managed to arrange them in a sequence. So the set of all such pairs is indeed a countable set.

And the same argument can be extended to argue for the set of all *triples of positive integers*, or the set of all *quadruples of positive integers*, and so on. This is actually not just a trivial mathematical point that we discuss for some curious reason, but it is because we will often have sample spaces that are of this kind. And it's important to know that they're countable:



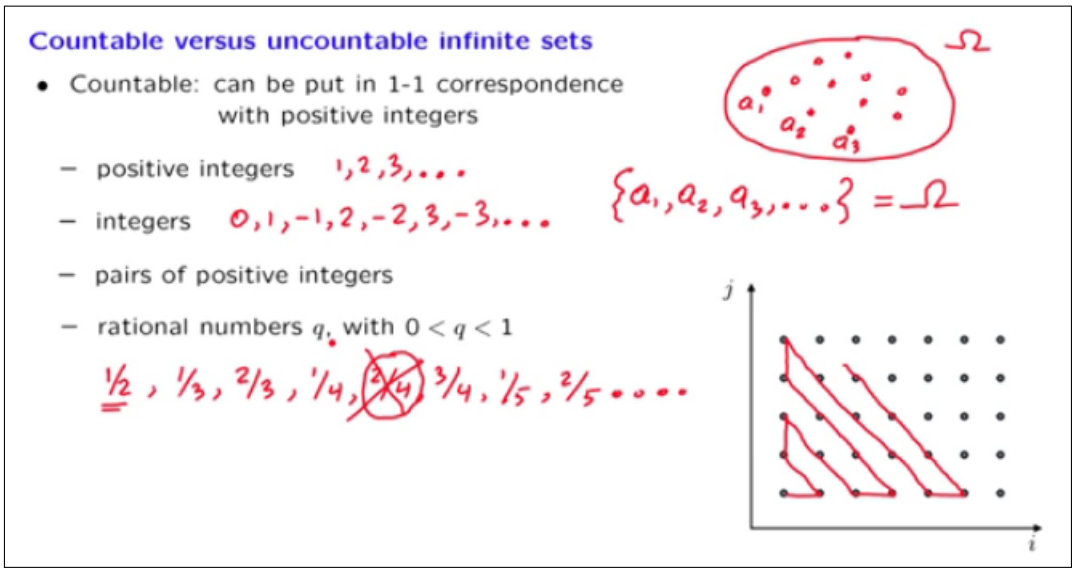
Now for a more subtle example. Let us look at all *rational numbers* within the range between 0 and 1. What do we mean by rational numbers? We mean those numbers that can be expressed as a ratio of two integers. It turns out that we can arrange them in a sequence, and we can do it as follows. Let us first look at rational numbers that have a denominator term of 2. Then, look at the rational numbers that have a denominator term of 3. Then, look at the rational numbers, always within this range of interest, that have a denominator of 4. And then we continue similarly — rational numbers that have a denominator of 5, and so on. This way, we're going to exhaust all of the rational numbers:

$$1/2,\ 1/3,\ 2/3,\ 1/4,\ 2/4,\ 3/4,\ 1/5,\ 2/5,\ \ldots$$

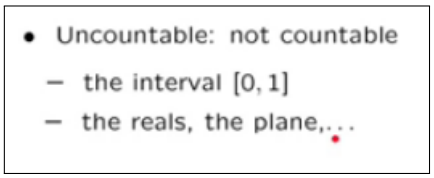
Actually, this number here already appeared there. It's the same number. So we do not need to include this in a sequence, but that's not an issue. Whenever we see a rational number that has already been encountered before, we just delete it:

$$1/2,\ 1/3,\ 2/3,\ 1/4,\ \cancel{2/4},\ 3/4,\ 1/5,\ 2/5,\ \ldots$$

In the end, we end up with a sequence that goes over all of the possible rational numbers. And so we conclude that the set of all rational numbers is itself a countable set:



So what kind of set would be uncountable? *An uncountable set, by definition, is a set that is not countable*. And there are examples of uncountable sets, most prominent, *continuous subsets of the real line*. Whenever we have an interval, the unit interval, or any other interval that has positive length, that interval is an uncountable set. And the same is true if, instead of an interval, we look at the entire real line, or we look at the two-dimensional plane, or three-dimensional space, and so on. So all the usual sets that we think of as continuous sets turn out to be uncountable.



How do we know that they are uncountable? There is actually a brilliant argument that establishes that the unit interval is uncountable. And then the argument is easily extended to other cases, like the reals and the plane.

We do not need to know how this argument goes, for the purposes of this course. But just because it is so beautiful, we will actually be presenting it to you.

3.2.10 Proof that the set of real numbers is uncountable

Video: Proof that the set of real numbers is uncountable (transcripts)

For those of you who are curious, we will go through an argument that establishes that the set of real numbers is an uncountable set. It's a famous argument known as *Cantor's diagonalization argument*.

Actually, instead of looking at the set of all real numbers, we will first look at the set of all numbers, x, that belong to the open unit interval — so numbers between 0 and 1 — and such that their decimal expansion involves only threes and fours.

$\{ x \in (0,1) : \text{decimal expansion only has 3,4} \}$

Now, the choice of three and four is somewhat arbitrary. It doesn't matter. What really matters is that we do not have long strings of nines.

So suppose that this set was countable. If the set was countable, then that set could be written as equal to a set of this form, x1, x2, x3 and so on, where each one of these is a real number inside that set:

$\{ x \in (0,1) : \text{decimal expansion only has 3,4} \}$

If countable "$\{x_1, x_2, x_3, \ldots\}$

Now, suppose that this is the case. Let us take those numbers and write them down in decimal notation. For example, one number could be this one, and it continues forever. Since we're talking about real numbers, their decimal expansion will go on forever. Suppose that the second number is of this kind, and it has its own decimal expansion. Suppose that the third number is, again, with some decimal expansion and so on:

$\{ x \in (0,1) : \text{decimal expansion only has 3,4} \}$

If countable "$\{x_1, x_2, x_3, \ldots\}$

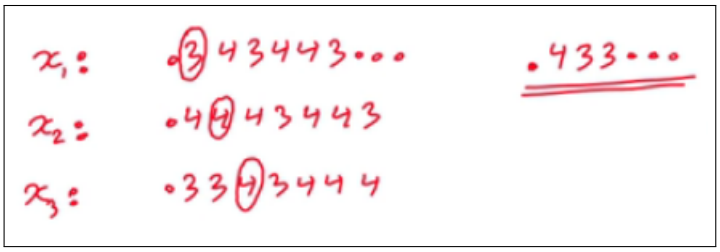
x_1: .343443...

x_2: .4443443

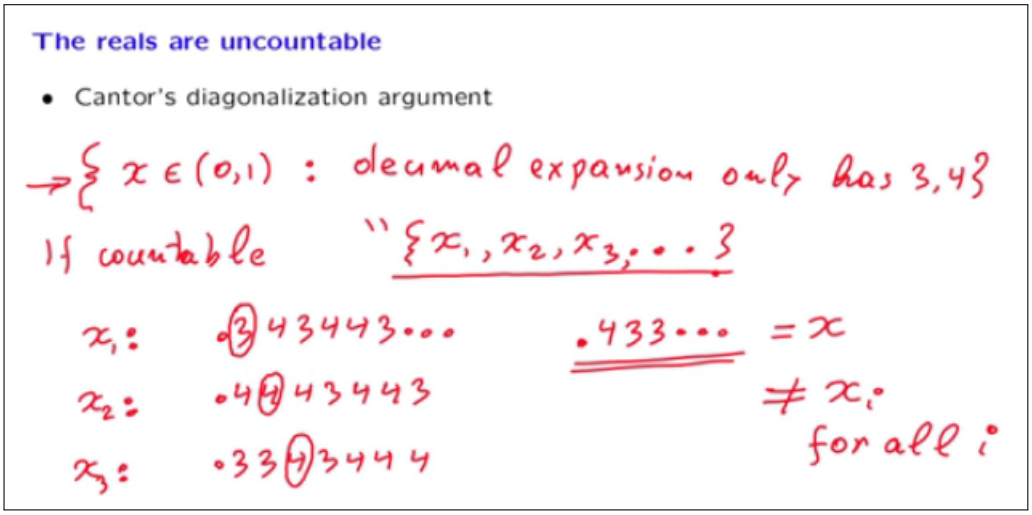
x_3: .3343444

So we *have assumed that our set is countable* and therefore, the set is equal to that sequence. So this sequence exhausts all the numbers in that set. Can it do that?

Let's construct a new number in the following fashion. The new number looks at this digit and does something different. Looks at this digit, the second digit of the second number, and does something different. Looks at the third digit of the third number and does something different. And we continue this way:



This number that we have constructed here is different from the first number. They differ in the first digit. It's different from the second number. They differ in the second digit. It's different from the third number because it's different in the third digit and so on. So this is a number, and this number is different from x_i for all i:



So we have an element of this set which does not belong to this sequence. Therefore, it cannot be true that this set is equal to the set formed by that sequence. And so this is a contradiction to the initial assumption that this set could be written in this form, and this contradiction establishes that since this is not possible, that the set that we have here is an uncountable set.

Now, this set is a subset of the set of real numbers. Since this one is uncountable, it is not hard to show that the set of real numbers, which is a bigger set, will also be uncountable.

And so this is this particular famous argument. We will not need it or make any arguments of this type in this class, but it's so beautiful that it's worth for everyone to see it once in their lifetime.

3.3 Solved problems

3.3.1 The probability of the difference of two events

Video: The probability of the difference of two events (transcripts)

The probability of the difference of two events. Give a mathematical derivation of the formula

$$P\left((A \cap B^c) \cup (A^c \cap B)\right) = P(A) + P(B) - 2P(A \cap B)$$

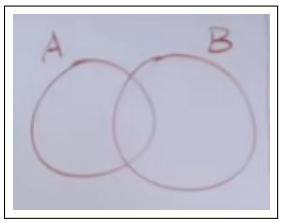
for the probability that exactly one of the events A and B will occur. Your derivation should be a sequence of steps, with each step justified by appealing to one of the probability axioms.

Teaching Assistant: Kuang Xu

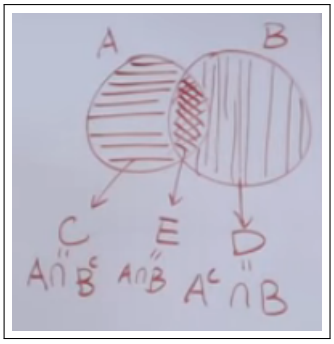
In this problem, we're going to use the set of probability axioms to *derive the probability of the difference of two events*.

Now, before we get started, there's one thing you might notice, that the equation we're trying to prove is actually quite complicated. And I don't like it either, so the first thing we're going to do will be to find a simpler notation for the events that we're interested in.

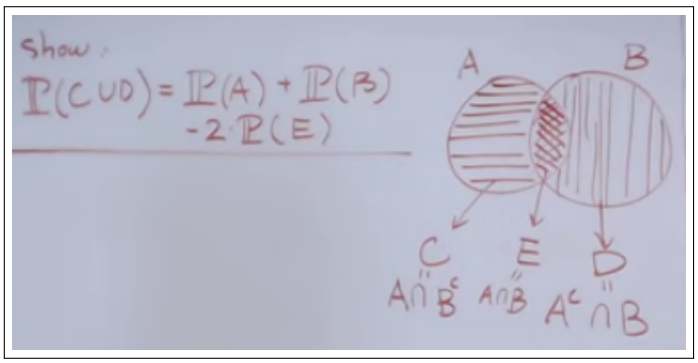
So we start with two events, A and B, and there might be some intersection between the two events:



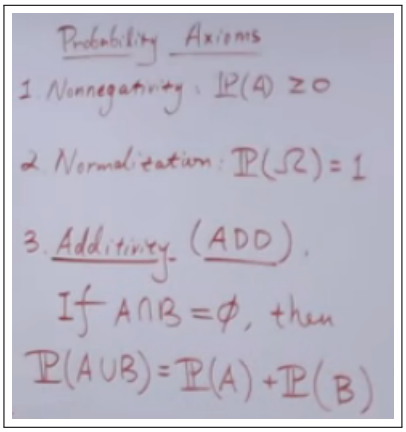
We'll label the set of points or samples in A that are not in B, as a set C. So C will be A intersection B complement. Similarly, for all points that are in B but not in A, this area, we'll call it D. And D will be the set A complement intersection B. And finally, for points that are in the intersection of A and B, we'll call it E. So E is A intersection B. And for the rest of our problem, we're going to be using the notation C, D, and E instead of whatever's down below:



If we use this notation, we can rewrite our objective as the following. We want to show that the probability of C union D is equal to the probability of the event A plus the probability of B minus twice the probability of E. And that will be our goal for the problem:

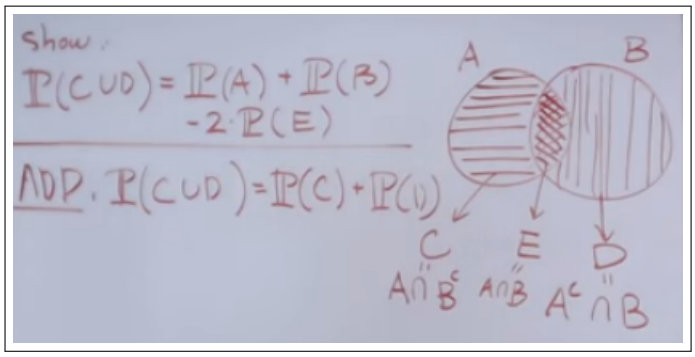


Now, let's take a minute to review what the axioms are, what the probability axioms are. The first one says non-negativity. We take any event A, then the probability of A must be at least 0. The second, normalization, says the probability of the entire space, the entire sample space omega, must be equal to 1. And finally, the additivity axiom, which will be the axiom that we're going to use for this problem says, if there are two events, A and B that are disjoint — which means they don't have anything in common, therefore. the intersection is the empty set. Then the probability of their union will be equal to the probabilty A plus the probability of B:

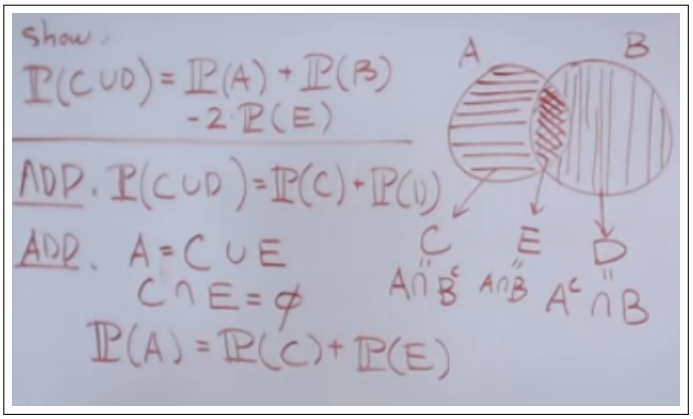


For the rest of the problem, I will refer to this axiom as ADD. So whenever we invoke this axiom, I'll write "ADD"on the board.

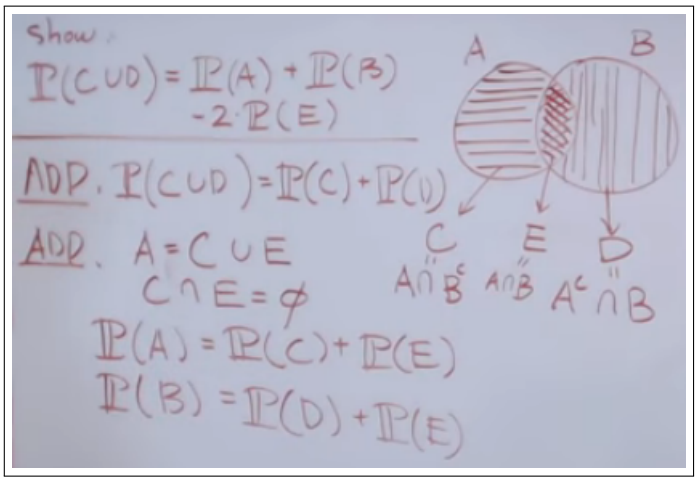
Let's get started. First, we'll invoke the additivity axioms to argue that the probability of C union D is simply the sum of probability of C plus probability of D:



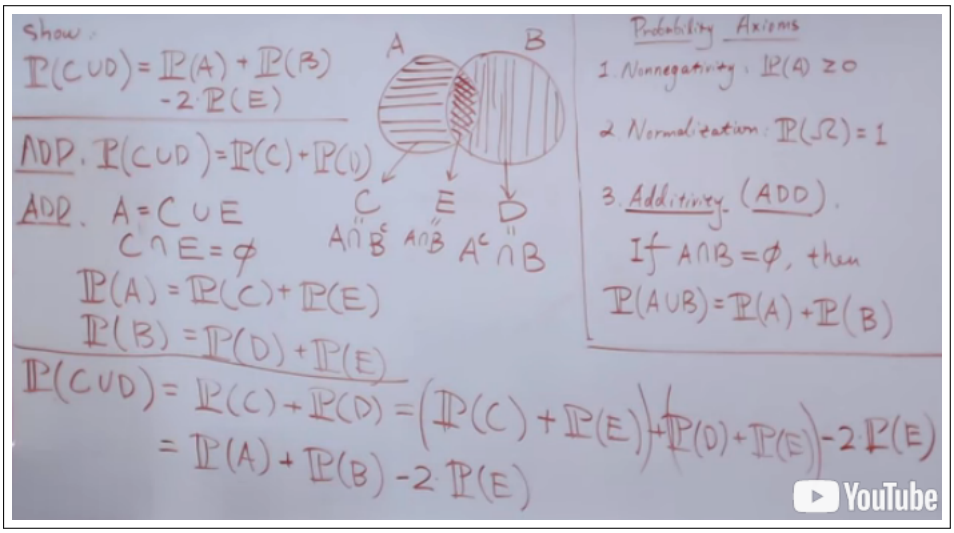
Why is this true? We can apply this axiom, because the set C here and the set D here, they're completely disjoint from each other. And in a similar way, we also notice the following. We see that A is equal to the union of the set C and E. And also, C and E, they're disjoint with each other, because C and E by definition don't share any points. And therefore, we have probability of A is equal to probability of C plus the probability of E:



Now, in a similar way, the probability of event B can also be written as a probability of D plus the probability of E, because event B is the union of D and E. And D and E are disjoint from each other. So we again invoke the additivity axiom:



Now, this should be enough to prove our final claim. We have the probability of C union D. By the very first line, we see this is simply probability of C plus the probability of D. Now, I'm going to insert two terms here to make the connection with a second part of the equation more obvious. That is, I will write probability C plus probability E plus probability D plus probability of E. Now, I've just added two terms here — probability E. So to make the equality valid we'll subtract out two times, the probability of E. Hence this equality is valid:



So if we look at this equation, we see that there are two parts here that we've already seen before, right here. The very first parenthesis is equal to the probability of A. And the value of the second parenthesis is equal to the probability of B. We just derived these here. And finally, we have the minus 2 probability of E. This line plus this line gives us the final equation. And that will be the answer for the problem.

3.3.2 Geniuses and chocolates

Video: Geniuses and chocolates (transcripts)

Geniuses and chocolates. Out of the students in a class, 60% are geniuses, 70% love chocolate, and 40% fall into both categories. Determine the probability that a randomly selected student is neither a genius nor a chocolate lover.

Teaching Assistant: Katie Szeto

Hi. Today, we're going to do a really fun problem called geniuses and chocolates. And what this problem is exercising is your knowledge of properties of probability laws. So let me just clarify what I mean by that.

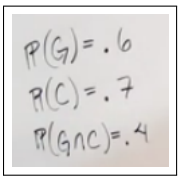
Hopefully, by this point, you have already learned what the axioms of probability are. And properties of probability laws are essentially any rules that you can derive from those axioms.

So take for example the fact that the probability of A union B is equal to the probability of A plus the probability of B minus the probability of the intersection. That's an example of a property of a probability law.

So enough with the preamble. Let's see what the problem is asking us. In this problem, we have a class of students. And we're told that 60% of the students are geniuses. 70% of the students love chocolate. So I would be in that category. And 40% fall into both categories.

And our job is to determine the probability that a randomly selected student is neither a genius nor a chocolate lover.

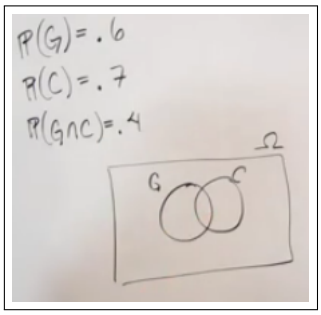
So first I just want to write down the information that we're given in the problem statement. So if you let G denote the event that a randomly selected student is a genius then the problem statement tells us that the probability of G is equal to 0.6. Similarly, if we let C denote the event that a randomly selected student is a chocolate lover, then we have that the probability of C is equal to 0.7. Lastly, we are told that the probability a randomly selected student falls into both categories is 0.4. And the way we can express that using the notation already on the board is probability of G intersect C is equal to 0.4:



OK, now one way of approaching this problem is to essentially use this information and sort of massage it using properties of probability laws to get to our answer. Instead, I'm going to take a different approach, which I think will be helpful. So namely, we're going to use something called a Venn diagram.

Now a Venn diagram is just a tool that's really useful for telling you how different sets relate to each other and how their corresponding probabilities relate to each other. So the way you usually draw this is you draw a rectangle, which denotes your sample space, which of course, we call omega. And then

you draw two intersecting circles. So one to represent our geniuses and one to represent our chocolate lovers:

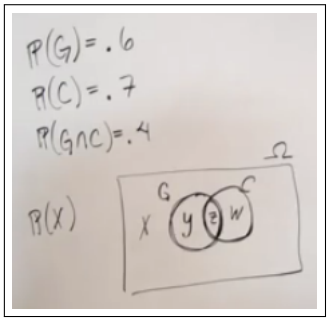


And the reason why I drew them intersecting is because we know that there are 40% of the students in our class are both geniuses and chocolate lovers.

OK, and the way you sort of interpret this diagram is the space outside these two circles correspond to students who are neither geniuses nor chocolate lovers. And so just keep in mind that the probability corresponding to these students on the outside, that's actually what we're looking for. Similarly, students in this little shape, this tear drop in the middle, those would correspond to geniuses and chocolate lovers. You probably get the idea. So this is our Venn diagram.

Now I'm going to give you guys a second trick if you will. And that is to work with *partitions*. So I believe you've seen partitions in lecture by now. And a partition is essentially a way of cutting up the sample space into pieces. But you need two properties to be true. So the pieces that you cut up your sample space into, they need to be *disjoint*, so they can't overlap. So for instance, G and C are not disjoint because they overlap in this tear drop region.

Now the second thing that a partition has to satisfy is that if you put all the pieces together, they have to *comprise the entire sample space*. So I'm just going to put these labels down on my graph. X, Y, Z, and W. So X is everything outside the two circles but inside the rectangle. And just note, again, that what we're actually trying to solve in this problem is the probability of X, the probability that you're neither genius, because you're not in this circle, and you're not a chocolate lover, because you're not in this circle:



So Y I'm using to refer to this sort of crescent moon shape. Z, I'm using to refer to this tear drop. And W, I'm using to refer to this shape.

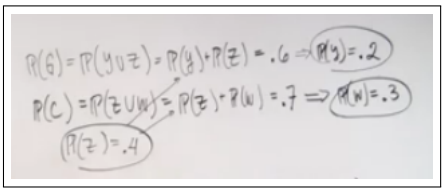
So, hopefully, you agree that X, Y, Z, and W form a partition because they don't overlap. So they are disjoint. And together they form omega. So now we're ready to do some computation.

The first step is to sort of get the information we have written down here in terms of these new labels. So hopefully, you guys buy that G is just the union of Y and Z. And because Y and Z are disjoint, we get that the probability of the union is the sum of the probabilities. And, of course, we have from before that this is 0.6.

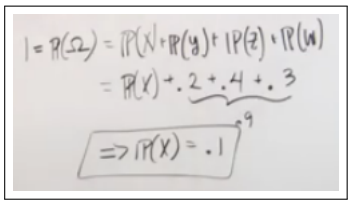
Similarly, we have that the probability of C is equal to the probability of Z union W. And, again, using the fact that these two guys are disjoint, you get this expression. And that is equal to 0.7.

OK, and the last piece of information, G intersects C corresponds to Z, or our tear drop, and so we have that the probability of Z is equal to 0.4.

And now, if you notice, probability of Z shows up in these two equations. So we can just plug it in. So plug in 0.4 into this equation. We get P of Y plus 0.4 is 0.6. So that implies that P of Y is 0.2. That's just algebra. And similarly we have point. 0.4 plus P of W is equal to 0.7. So that implies that P of W is 0.3. Again, that's just algebra.



So now we're doing really well because we have a lot of information. We know the probability of Y, the probability of Z, the probability of W. But remember we're going for, we're trying to find the probability of X. So the way we finally put all this information together to solve for X is we use the axiom that tells us that 1 is equal to the probability of the sample space. And then, again, we're going to use sort of this really helpful fact that X, Y, Z, and W form a partition of omega to go ahead and write this as probability of X plus probability of Y plus probability, oops, I made a mistake. Hopefully, you guys caught that. It's really, oh, no. I'm right. Never mind. Probability of X plus probability of Y plus probability of Z plus probability of W. And now we can go ahead and plug-in the values that we solved for previously. So we get probability of X plus 0.2 plus 0.4 plus 0.3. These guys sum to 0.9. So, again, just simple arithmetic, we get that the probability of X is equal to 0.1.



So we're done because we've successfully found that the probability that a randomly selected student is neither a genius nor a chocolate lover is 0.1.

So this was a fairly straightforward problem. But there are some important takeaways:

- **Venn Diagrams**: The first one is that Venn diagrams are a really nice tool. Whenever the problem is asking you how different sets relate to each other or how different probabilities relate to each other, you should probably draw a Venn diagram because it will help you.
- **Partitions**: And the second takeaway is that it's frequently useful to divide your sample space into a partition mainly because the pieces that compose a partition are disjoint. So we will be back soon to solve more problems.

3.3.3 Uniform probabilities on a square

Video: Uniform probabilities on a square

Uniform probabilities on a square. Romeo and Juliet have a date at a given time, and each will arrive at the meeting place with a delay between 0 and 1 hour, with all pairs of delays being "equally likely,"that is, according to a uniform probability law on the unit square. The first to arrive will wait for 15 minutes and will leave if the other has not arrived. What is the probability that they will meet?

Teaching Assistant: Jimmy Li

In this problem, we will be helping Romeo and Juliet meet up for a date. And in the process, also we'll review some concepts in basic probability theory, including sample spaces and probability laws.

This problem, the basic setup is that Romeo and Juliet are trying to meet up for a date. And let's say they're trying to meet up for lunch tomorrow at noon. But they're not necessarily punctual. So they may arrive on time with a delay of 0, or they may actually be up to 1 hour late and arrive at 1:00 PM.

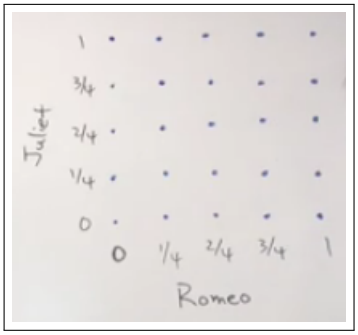
So the other thing that we assume in this problem is that all pairs of arrival times — so the time that Romeo arrives paired with the time they Juliet arrives — all of these pairs are equally likely. And I've put this in quotes, because we haven't really specify exactly what this means. And we'll come back to that in a little bit.

The last important thing is that each person will wait for 15 minutes for the other person to arrive. If within that 15-minute window the other per-

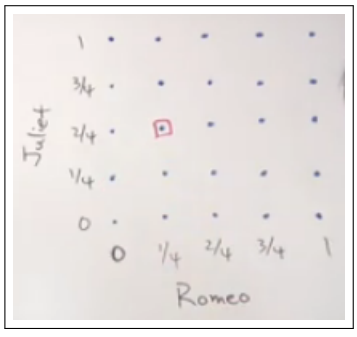
son doesn't arrive, then they'll give up and they'll end up not meeting up for lunch.

So to solve this problem, let's first try to *set up a sample space* and come up with a *probability law* to describe this scenario.

And let's actually start with a simpler version of this problem. And instead of assuming that they can arrive at any delay between 0 and 1 hour, let's pretend instead that Romeo and Juliet can only arrive in 15-minute increments. So Romeo can arrive on time with a delay 0, or be 15 minutes late, 30 minutes late, 45 minutes late, or one hour late. But none of the other times are possible. And the same thing for Juliet:



Let's start out with just the simple case first, because it helps us get the intuition for the problem, and it's an easier case to analyze. So it's actually easy to visualize this. It's a nice visual tool to group this sample space into a grid. So the horizontal axis here represents the arrival time of Romeo, and the vertical axis represents the arrival time of Juliet. And so, for example, this point here would represent Romeo arriving 15 minutes late and Juliet arriving 30 minutes late:

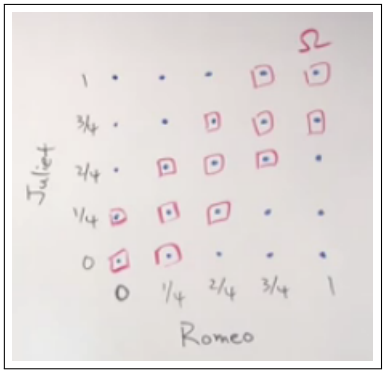


So this is our sample space now. This is our omega. And now let's try to assign a probability law. And we'll continue to assume that all pairs of arrival times are equally likely. And now we can actually specifically specify what this term means. And in particular, we'll be invoking the *discrete uniform law*, which basically says that all of these points, which are just outcomes in our probabilistic experiment — all of these outcomes are equally likely. And so

since there are 25 of them, each one of these outcomes has a probability of 1 over 25.

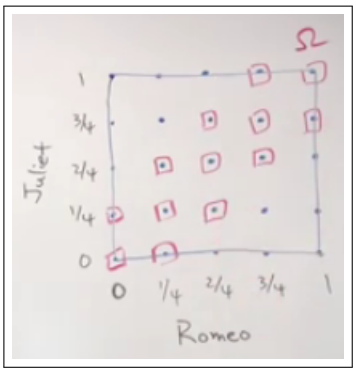
So now we've specified our sample space and our probability law. So now let's try to answer the question, what is the probability that Romeo and Juliet will meet up for their date? So all that amounts to now is just identifying which of these 25 outcomes result in Romeo and Juliet arriving within 15 minutes of each other.

So let's start with this one that I've picked out. If Romeo arrives 15 minutes late and Juliet arrives 30 minutes late, then they will arrive within 15 minutes of each other. So this outcome does result in the two of them meeting. And so we can actually highlight all of these. And it turns out that these outcomes that I'm highlighting result in the two them arriving within 15 minutes of each other:

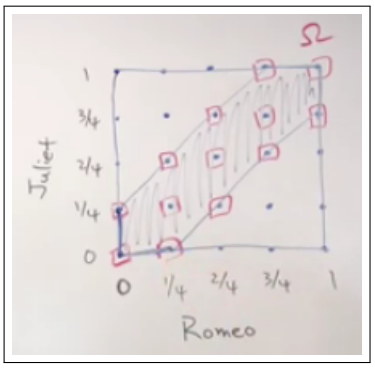


So because each one has a probability of 1 over 25, all we really need to do now is just count how many outcomes there are. So there's 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13. So the probability in the end is for the discrete case. The discrete case — I'm referring to the case where we simplified it and considered only arrival times with increments of 15 minutes. In this case, the probability is 13 over 25.

So now we have an idea of how to solve this problem. It amounts to basically coming up with a sample space, a probability law, and then identifying the event of interest and calculating the probability of that event. So now let's actually solve the problem that we really are interested in, which is that instead of confining Romeo and Juliet to arrive in only 15-minute minute increments, really, time is continuous, and Romeo and Juliet can arrive at any time. So they don't necessarily have to arrive 15 minutes late. Romeo could arrive 15 minutes and 37 seconds late if he wanted to. So now our new sample space is actually just, instead of only these 25 points in the grid, it's this entire square. So any point within the square could be a possible pair of meeting times between Romeo and Juliet:

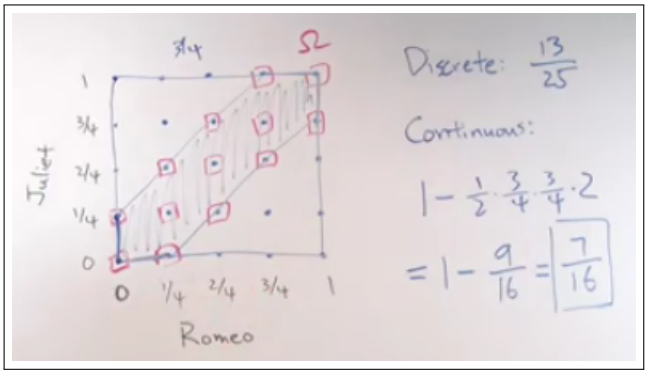


So that is our new sample space, our new omega. And now let's assign a new probability law. And now, instead of being in the discrete world, we're in the continuous world. And the analogy here is to consider probabilities as areas. So the area of this entire square is one. And that also corresponds to the probability of omega, the sample space. And imagine just spreading probability evenly across this square so that the probability of any event — which in this case would just be any shape within this square — is exactly equal to the area of that shape. So now that is our new sample space and our new probability law. So what we have to do now is just to identify the event of interest, which is still the event that Romeo and Juliet arrive within 15 minutes of each other. So let's do that. If Romeo and Juliet arrive both on time, then obviously they'll meet. And if Romeo's on time and Juliet is 15 minutes late, then they will still meet. And in fact, any pairs of meeting times between these would still work, because now Romeo can be on time, and Juliet can arrive at any time between 0 and 15 minutes late. But you notice that if Juliet is even a tiny bit later than 15 minutes, then they won't end up meeting. So this segment here is part of the event of interest. And similarly, this segment here is also part of the event. And if you take this exercise and extend it, you can actually verify that the event of interest is this strip shape in the middle of the square. Which, if you think about it, makes sense, because you want the arrival times between Romeo and Juliet to be close to each other, so you would be expect it to be somewhere close to a diagonal in this square:



So now we have our event of interest. We have our sample space and our probability law. So all we have to do now is just calculate what this probability is. And we've already said that the probability in this probability law is just areas. So now it actually just boils down to not a probability problem, but a problem in geometry.

So to calculate this area, you can do it in lots of ways. One way is to calculate the area of the square, which is 1, and subtract the areas of these two triangles. So let's do that. So in the continuous case, the probability of meeting is going to be 1 minus the area of this triangle. The base here is 3/4 and 3/4, so it's 1/2 times 3/4 times 3/4. That's the area of one of these triangles. There's two of them, so we'll multiply by two. And we end up with 1 minus 9/16, or 7/16 as our final answer:



So in this problem, we've reviewed some basic concepts of probability, and that's also helped us solve this problem of helping Romeo and Juliet meet up for a date.

And if you wanted to, you could even extend this problem even further and turn it on its head. And instead of calculating given that they arrive within 15 minutes of each other, what is the probability that they'll meet, let's say that Romeo really wants to meet up with Juliet, and he wants to assure himself a least, say, a 90% chance of meeting Juliet. Then you can ask, if he wants to have at least a 90% chance of meeting her, how long should he be willing to wait? And so that's the flip side of the problem.

And you can see that with just some basic concepts of probability, you can answer some already pretty interesting problems. So I hope this problem was interesting, and we'll see you next time.

3.3.4 Bonferroni's inequality

Video: Bonferroni's inequality (transcripts)

Bonferroni's inequality:

(a) Prove that for any two events A_1 and A_2, we have

$$P(A_1 \cap A_2) \geq P(A_1) + P(A_2) - 1$$

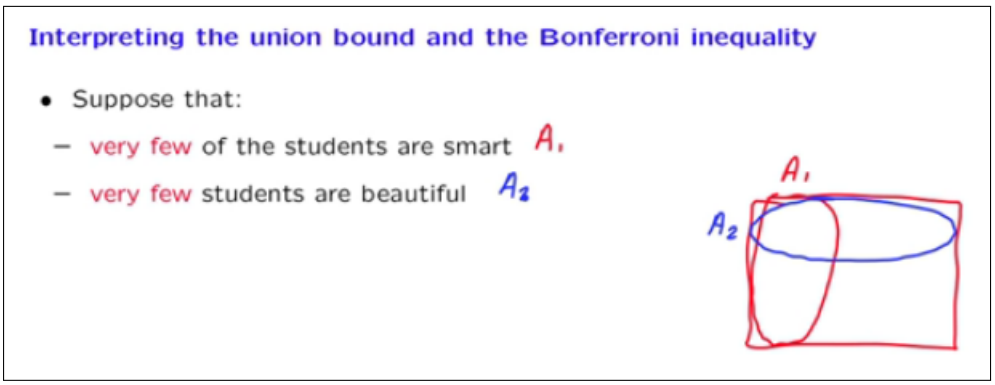
(b) Generalize to the case of n events $A_1, A_2, \ldots, A_n$, showing that

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) + P(A_2) + \cdots + P(A_n) - (n-1)$$

Instructor: John Tsitsiklis

In this segment, we will discuss a little bit the *union bound*. And then discuss a counterpart, which is known as the *Bonferroni inequality*. Let us start with a story.

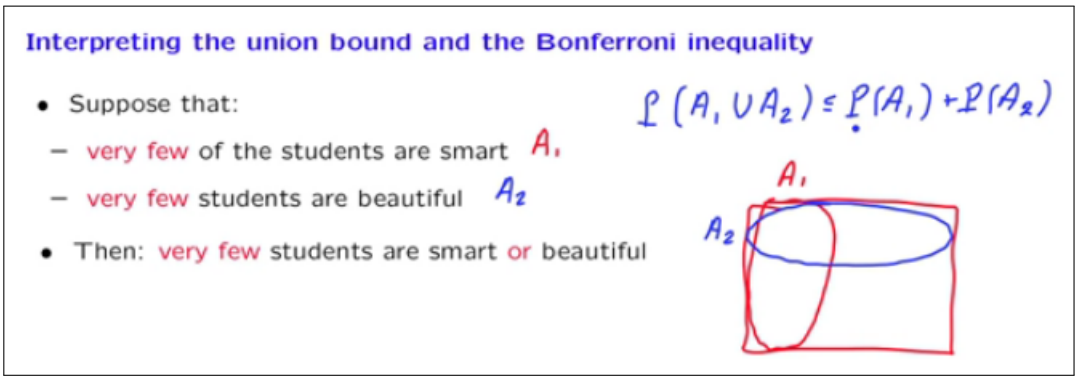
Suppose that we have a number of students in some class. And we have a set of students that are smart, let's call that set A_1. So this is the set of smart students. And we have a set of students that are beautiful. And let's call that set A_2. So A_2 is the set of beautiful students.



If I tell you that the set of smart students is small, and the set of beautiful students [is] small, then you can probably conclude that there are very few students that are either smart or beautiful.

What does this have to do with probability? Well, when we say very few are smart, we might mean that if I pick a student at random, there's only a small probability that I pick a smart student. And similarly for beautiful students.

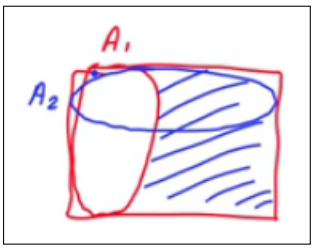
Can we make this statement more precise? Indeed we can. We have the *union bound* that tells us that the probability that I pick a student that is either smart or beautiful is less than or equal to the probability of picking a smart student plus the probability of picking a beautiful student:



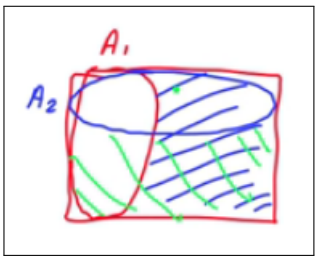
So if this probability is small, and that probability is small, then this probability will also be small. Which means that there is only a small number of students that are either smart or beautiful.

Now let us try to turn this statement around its head. Suppose that most of the students are smart. And most of the students are beautiful. So in this case, I'm telling you that these sets, A1 and A2 are big.

Now if the set A1 is big, then it means that this set here, the complement of A1, is a small set:



And if I tell you that the set A2 is big, then it means that this set here, which is a complement of A2, is also small:

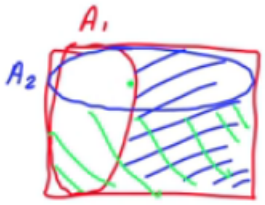


So everything outside here is a small set. Which means that whatever is left — which is the intersection of A1 and A2 — should be a big set. So we should be able to conclude that in this case, most of the students belong to the intersection. So they're both smart and beautiful:

Interpreting the union bound and the Bonferroni inequality

- Suppose that:
  - very few of the students are smart A_1
  - very few students are beautiful A_2
- Then: very few students are smart or beautiful

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



- Suppose that:
  - most of the students are smart
  - most students are beautiful
- Then: most students are smart and beautiful

How can we turn this into a mathematical statement? It's the following inequality that we will prove shortly:

- Then: most students are smart and beautiful

$$\mathbf{P}(A_1 \cap A_2) \geq \mathbf{P}(A_1) + \mathbf{P}(A_2) - 1$$

But what it says is that the probability of the intersection is larger than or equal to something. And if this probability is close to 1, which says that most of the students are smart, and this probability is close to 1, which says that most students are beautiful, then this difference here is going to be close to 1 plus 1 minus 1, which is 1.

Therefore, the probability of the intersection is going to be larger than or equal to some number that's close to 1. So this one will also be close to 1, which is the conclusion that indeed, most students fall into this intersection. And they're both smart and beautiful.

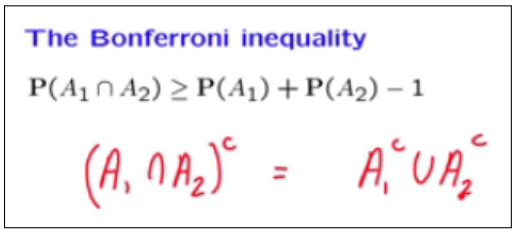
So what we will do next will be to derive this inequality and actually generalize it. So here is the relation that we wish to establish:

The Bonferroni inequality

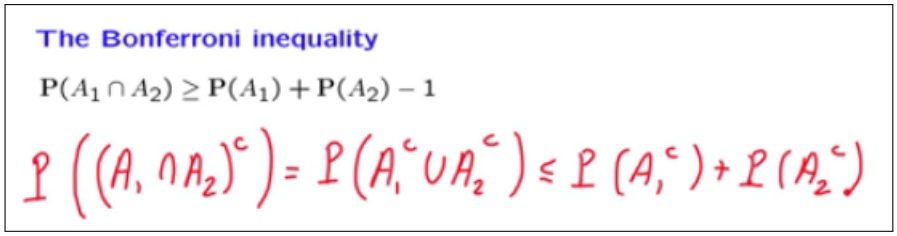
$$\mathbf{P}(A_1 \cap A_2) \geq \mathbf{P}(A_1) + \mathbf{P}(A_2) - 1$$

We want to show that the probability of a certain event is bigger than something. How do we show that? One way is to show that the probability of the complement of this event, namely this event here, $(A_1 \cap A_2)$, we want to show that this event has small probability.

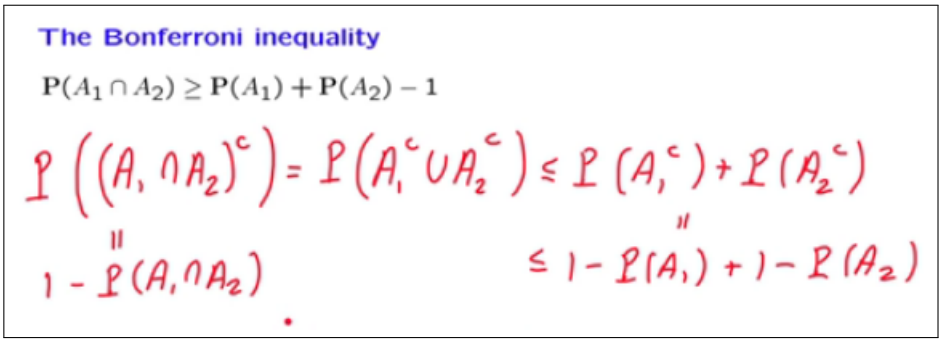
Now what is this event? Here we can use De Morgan's laws, which tell us that this event is the same as this one. That is, the complement of an intersection is the union of the complements:



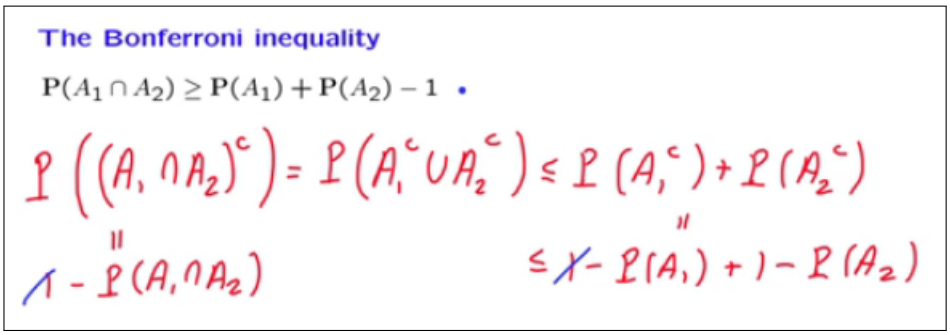
Since these two sets or events are identical, it means that their probabilities will also be equal. And next we will use the union bound to write this probability as being less than or equal to the sum of the probabilities of the two events whose union we are taking:



Now we're getting close, except that here we have complements all over. Whereas up here we do not have any complements. What can we do? Well, the probability of a complement of an event is the same as 1 minus the probability of that event. And we do the same thing for the terms that we have here. This probability here is equal to 1 minus the probability of A1. And this probability here is equal to 1 minus the probability of A2:



And now if we take this inequality, cancel this term with that term, and then move terms around, what we have is exactly this relation that we wanted to prove:



It turns out that this inequality has a generalization to the case where we take the intersection of n events:

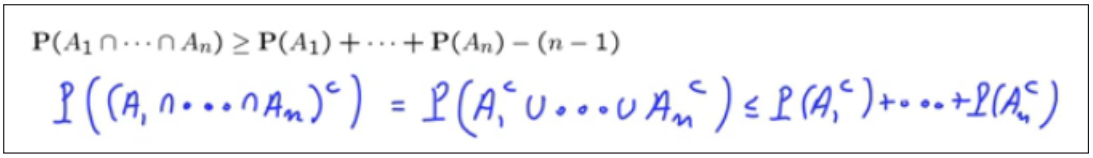
$$P(A_1 \cap \cdots \cap A_n) \geq P(A_1) + \cdots + P(A_n) - (n-1)$$

And this has again the same intuitive content. Suppose that each one of these events, A1 up to An, is almost certain to occur. That is, it has a probability close to 1.

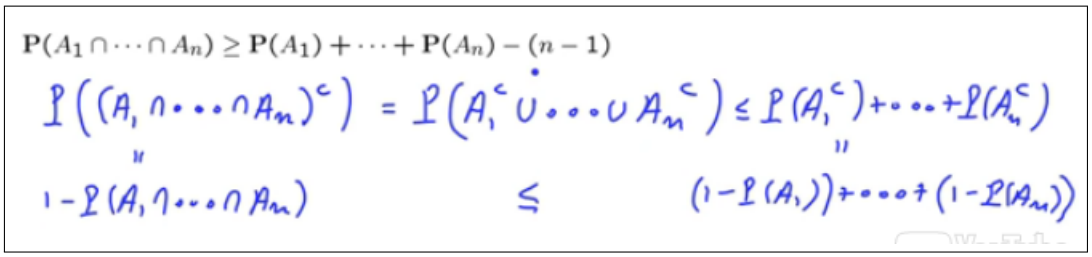
In that case, this term will be close to n. We subtract n minus 1. So this term on the right hand side will be close to 1. Therefore the probability of the intersection will be larger than or equal to something that's close to 1. So this is big.

Essentially, what it's saying is that [if] we have big sets. And we take their intersection. Then that intersection will also be big in terms of having large probability.

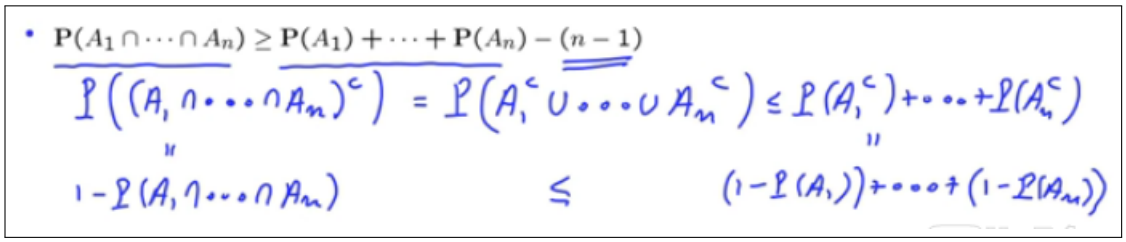
How do we prove this relation? Exactly the same way as it was proved for the case of two sets. Namely, instead of looking at this event, we look at the complement of this event. And we use De Morgan's laws to write this complement as the union of the complements. These two are the same sets or events, so they have the same probability. And then we use the union bound to write this as being less than or equal to the probabilities of all those sets:



Now this is equal to 1 minus the probability of the intersection. This side here is equal to 1 minus the probability of A1. This is one term. We get n such terms, the last one being 1 minus the probability of An. And we still have an inequality going this way:



We collect those ones that we have here. There's n over them. And one here. So we're left with n minus 1 terms that are equal to 1. And this gives rise to this term. We have all of the probabilities of the various events that appear with the same sign. This gives rise to this term. And finally, this term here will correspond to that term. Namely, if we start with this inequality and just rearrange a few terms, we obtain this inequality up here:



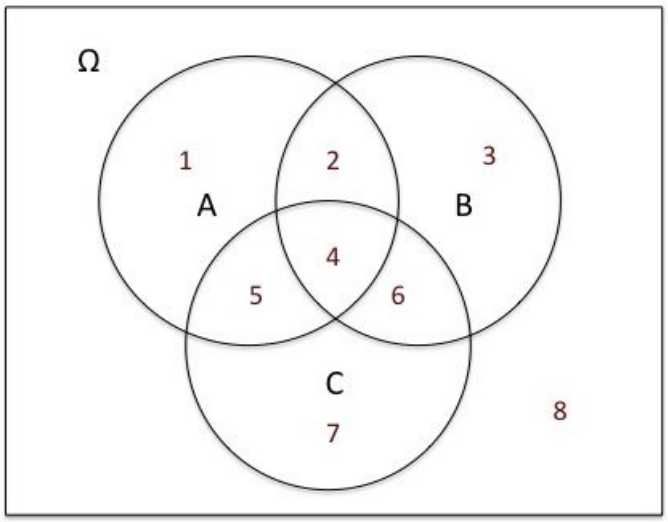
So these Bonferroni inequalities are a nice illustration of how one can combine De Morgan's laws, set -theoretic operations, and the union bound in order to obtain some interesting relations between probabilities.

3.4 Problem Set 1

3.4.1 Venn diagrams

Exercise 3.4.1-1: Venn diagrams

In this problem, you are given descriptions in words of certain events (e.g., "at least one of the events A, B, C occurs"). For each one of these descriptions, identify the correct symbolic description in terms of A, B, C from Events E1-E7 below. Also identify the correct description in terms of regions (i.e., subsets of the sample space Ω) as depicted in the Venn diagram below. (For example, Region 1 is the part A of outside of B and C.)



Symbolic descriptions:

- Event E1: $A \cap B \cap C$
- Event E2: $(A \cap B \cap C)^c$
- Event E3: $A \cap B \cap C^c$
- Event E4: $B \cup (B^c \cap C^c)$
- Event E5: $A^c \cap B^c \cap C^c$
- Event E6: $(A \cap B) \cup (A \cap C) \cup (B \cap C)$
- Event E7: $(A \cap B^c \cap C^c) \cup (A^c \cap B \cap C^c) \cup (A^c \cap B^c \cap C)$

a) At least two of the events A, B, C occur.

◯ Event E1
◯ Event E2
◯ Event E3
◯ Event E4
◯ Event E5
◯ Event E6
◯ Event E7

◯ Regions 1 3 4 7
◯ Regions 2 5 6

◯ Regions 2 4 5 6
◯ Regions 1 2 3 5 6 7

b) At most two of the events A, B, C occur.

◯ Event E1
◯ Event E2
◯ Event E3
◯ Event E4
◯ Event E5
◯ Event E6
◯ Event E7

◯ Regions 1 2 3 5 6 7
◯ Regions 2 5 6
◯ Regions 2 4 5 6
◯ Regions 1 2 3 5 6 7 8

c) None of the events A, B, C occurs.

◯ Event E1
◯ Event E2
◯ Event E3
◯ Event E4
◯ Event E5
◯ Event E6
◯ Event E7

◯ Regions 8
◯ Regions 1 3 7
◯ Regions 4
◯ Regions 2 4 5 6

d) All three events A, B, C occur.

◯ Event E1
◯ Event E2
◯ Event E3
◯ Event E4
◯ Event E5
◯ Event E6
◯ Event E7

◯ Regions 8
◯ Regions 1 3 7
◯ Regions 4
◯ Regions 2 4 5 6

e) Exactly one of the events A, B, C occurs.

◯ Event E1
◯ Event E2
◯ Event E3
◯ Event E4
◯ Event E5
◯ Event E6
◯ Event E7

◯ Regions 2 5 6
◯ Regions 1 3 7
◯ Regions 4
◯ Regions 2 4 5 6

f) Events A and B occur, but C does not occur.

◯ Event E1
◯ Event E2
◯ Event E3
◯ Event E4
◯ Event E5
◯ Event E6
◯ Event E7

◯ Regions 2
◯ Regions 2 4
◯ Regions 4
◯ Regions 1 4 5 6

g) Either (i) event B occurs, or (ii) neither B or C occurs.

◯ Event E1
◯ Event E2
◯ Event E3
◯ Event E4
◯ Event E5
◯ Event E6
◯ Event E7

◯ Regions 2 3 4 6
◯ Regions 2 3 4 6 8
◯ Regions 1 2 3 4 6 8
◯ Regions 1 2 3 4 5 6 8

3.4.2 Set operations and probabilities

Exercise 3.4.2-1: Set operations and probabilities

Find the value of $P(A \cup (B^c \cup C^c)^c)$ for each of the following cases:

a) The events A, B, C are disjoints events and $P(A) = 2/5$.

b) The events A and C are disjoint, and $P(A) = 1/2$ and $P(B \cap C) = 1/4$.

c) $P(A^c \cap (B^c \cup C^c)) = 0.7$.

3.4.3 Three tosses of a fair coin

Exercise 3.4.3-1: Three tosses of a fair coin

You flip a fair coin (i.e., the probability of obtaining Heads is 1/2) three times. Assume that all sequences of coin flip results, of length 3, are equally likely. Determine the probability of each of the following events.

a) $\{HHH\}$: 3 Heads

b) $\{HTH\}$: the sequence Heads, Tails, Heads

c) Any sequence with 2 Heads and 1 Tail (in any order)

d) Any sequence in which the number of Heads is greater than or equal to the number of Tails.

3.4.4 Parking lot problem

Exercise 3.4.4-1: Parking lot problem

Mary and Tom park their cars in an empty parking lot with $n \geq 2$ consecutive parking spaces (i.e, n spaces in a row, where only one car fits in each space). Mary and Tom pick parking spaces at random; of course, they must each choose a different space. (All pairs of distinct parking spaces are equally likely.) What is the probability that there is at most one empty parking space between them? Your answer should be a function of n, entered using standard notation.

3.4.5 Probabilities on a continuous sample space

Exercise 3.4.5-1: Probabilities on a continuous sample space

Alice and Bob each choose at random a real number between zero and one. We assume that the pair of numbers is chosen according to the uniform probability law on the unit square, so that the probability of an event is equal to its area. We define the following events:

$$
\begin{aligned}
A &= \{\text{The magnitude of the difference of the two numbers is greater than } 1/3\} \\
B &= \{\text{At least one of the numbers is greater than } 1/4\} \\
C &= \{\text{The sum of the two numbers is } 1\} \\
D &= \{\text{Alice's number is greater than } 1/4\}
\end{aligned}
$$

Find the following probabilities:
a) $P(A)$
b) $P(B)$
c) $P(A \cap B)$
d) $P(C)$
e) $P(D)$
f) $P(A \cap D)$

3.4.6 Upper and lower bounds on the probability of intersection

Exercise 3.4.6-1: Upper and lower bounds on the probability of intersection
Given two events A, B with $P(A) = 3/4$ and $P(B) = 1/3$, what is the smallest possible value of $P(A \cap B)$? The largest? That is, find a and b such that, $a \leq P(A \cap B) \leq b$, holds and any value in the closed interval [a,b] is possible.

4 Unit 2: Conditioning and independence (2018/09/10)

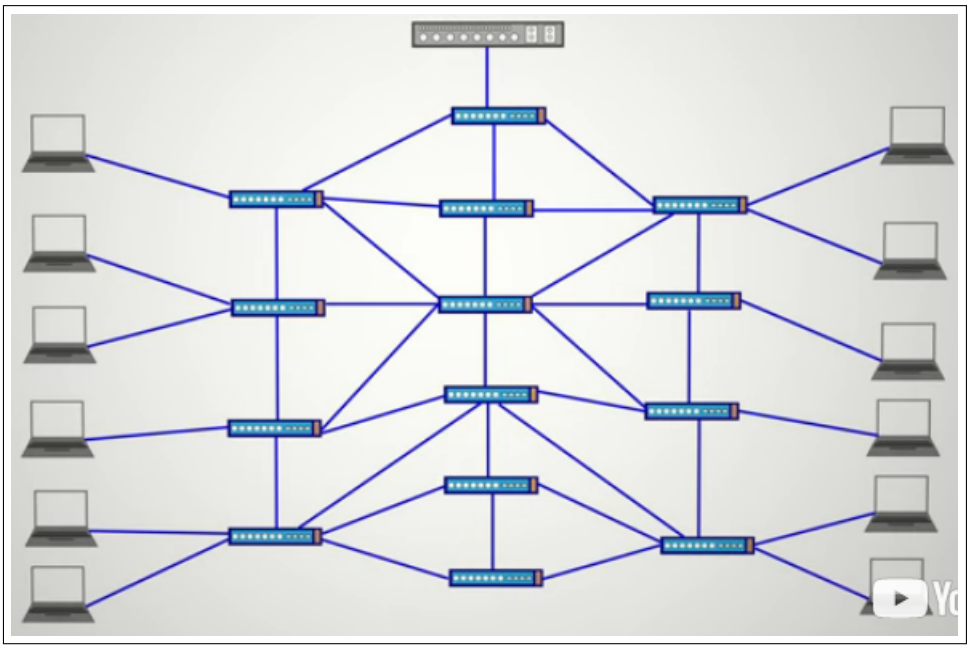
4.1 Unit overview

Atention: Exercises due Sep 18, 2018 20:59:59 -03.

4.1.1 Motivation

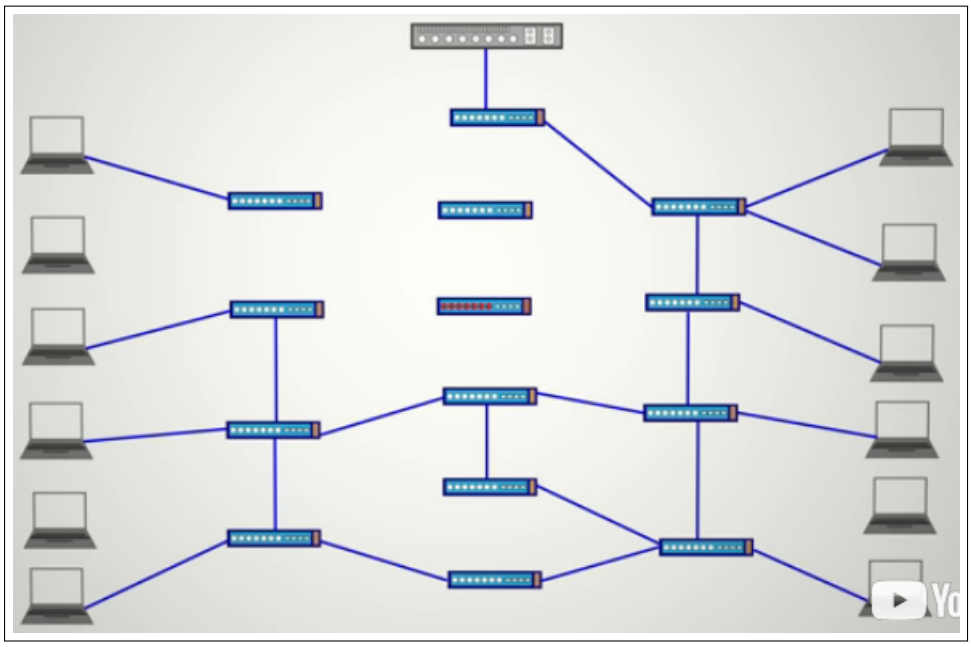
Video: Motivation

You build a local communication network consisting of routers, computers, and some hard-wired links that connect them



However, each link has a small probability of being nonfunctional because of a hardware failure. Assuming that failures at different links happen *independently*, how likely is it that a message from router A will still be able to reach your computer?

Suppose now that you receive notice that one of the routers is down:



Given this information, how would you update the probability that the message from router A will still be able to reach your computer?

By the end of this unit, you will be able to answer questions of this type by using the concept of *conditional probability* and by giving a precise meaning to the concept of *independent* failures.

4.1.2 Unit 2 overview

Video: Unit 2 overview (transcripts, slides)

In this unit we introduce the concepts of *conditioning* and *independence*[4].

Conditioning leads to revised ("conditional") probabilities that take into account partial information on the outcome of a probabilistic experiment. Conditioning is a very useful tool that allows us to "divide and conquer"complex problems.

Independence is used to model situations involving non-interacting probabilistic phenomena and also plays an important role in building complex models from more elementary ones.

In the first lecture, we introduced probabilities as a *way of describing our beliefs about the likelihood that a given event will occur*. But our beliefs will in general *depend on the information that we have*.

Taking into account new information leads us to consider so-called **conditional probabilities**. These are *revised probabilities that take into account the new information*.

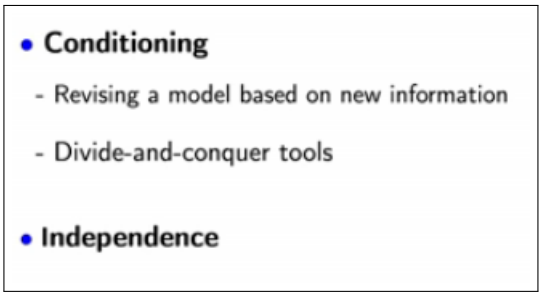
Conditional probabilities are very useful whenever we want to break up a model into simpler pieces using a divide and conquer strategy. This is done using certain tools that we will develop and which we will keep applying

[4]The material in this unit is covered in Sections 1.3-1.5 of the text.

throughout this course in different guises. They are also the foundation of the field of *inference*. And we will see how they arise in that context.

Then, in the second lecture of this unit, we will consider a special case where *one event does not convey useful information about another*, a situation that we call **independence**.

Independence usually describes a situation where the occurrence or non-occurrence of different events is determined by *factors that are completely unrelated*. Independence is what allows us to build complex models out of simple ones. This is because it is often the case that a complex system is made up of several components that are affected by unrelated, that is, independent sources of randomness. And so with the tools to be developed in this unit, we will be ready to calculate probabilities in fairly complex probabilistic models.



4.2 Lecture 2: Conditioning and Bayes' rule

4.2.1 Overview

Video: Lecture 2 Overview (transcripts, slides, annotated slides)

This lecture sequence introduces conditional probabilities and three basic tools: the multiplication rule, the total probability theorem, and Bayes' rule[5] [6].

Suppose I look at the registry of residents of my town and pick a person at random. What is the probability that this person is under 18 years of age? The answer is about 25%.

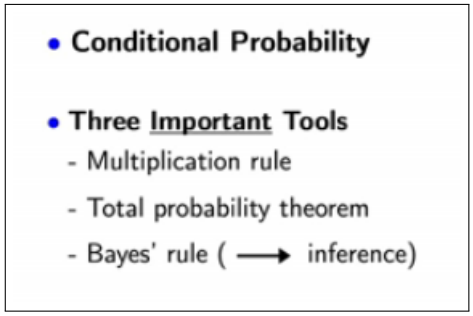
Suppose now that I tell you that this person is married. Will you give the same answer? Of course not. The probability of being less than 18 years old is now much smaller.

What happened here? We *started with some initial probabilities* that reflect what we know or believe about the world. But we then *acquired some additional knowledge*, some new evidence — for example, about this person's family situation. This new knowledge should cause our beliefs to change, and the original probabilities must be replaced with new probabilities that take into account the new information. These revised probabilities are what we call **conditional probabilities**. And this is the subject of this lecture.

[5] The same material, in live lecture hall format, can be found here and here.

[6] More information is given in the text: Conditional probability (Section 1.3) and Total probability theorem and Bayes' rule (Section 1.4)

We will start with a *formal definition* of conditional probabilities together with the motivation behind this particular definition. We will then proceed to develop *three tools that rely on conditional probabilities*, including the Bayes rule, which provides a *systematic way for incorporating new evidence into a probability model*:



The three tools that we introduce in this lecture involve very simple and elementary mathematical formulas, yet they encapsulate some very powerful ideas. It is not an exaggeration to say that much of this class will revolve around the repeated application of variations of these three tools to increasingly complicated situations.

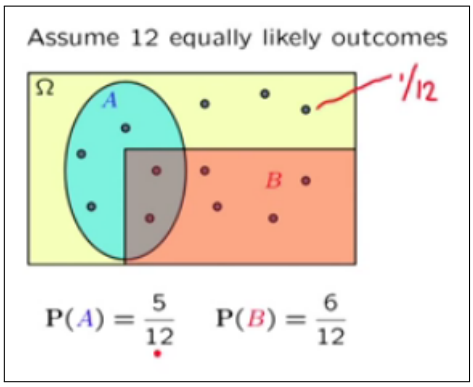
In particular, the *Bayes rule is the foundation for the field of inference*. It is a guide on how to process data and make inferences about unobserved quantities or phenomena. As such, it is a tool that is used all the time, all over science and engineering.

4.2.2 Conditional probabilities

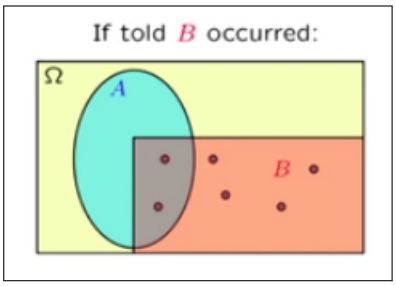
Video: Conditional probabilities (transcripts)

Conditional probabilities are *probabilities associated with a revised model that takes into account some additional information about the outcome of a probabilistic experiment*. The question is how to carry out this revision of our model.

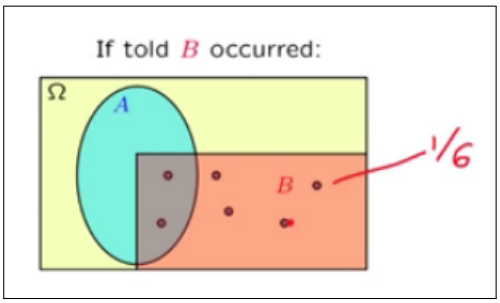
We will give a mathematical definition of conditional probabilities, but first let us motivate this definition by examining a simple concrete example. Consider a probability model with 12 equally likely possible outcomes, and so each one of them has probability equal to 1/12. We will focus on two particular events, event A and B, two subsets of the sample space:



Event A has five elements, so its probability is 5/12, and event B has six elements, so it has probability 6/12. Suppose now that someone tells you that event B has occurred, but tells you nothing more about the outcome. How should the model change?



First, those outcomes that are outside event B are no longer possible. So we can either eliminate them, as was done in this picture, or we might keep them in the picture but assign them 0 probability, so that they cannot occur. How about the outcomes inside the event B? So we're told that one of these has occurred. Now these 6 outcomes inside the event B were equally likely in the original model, and *there is no reason to change their relative probabilities*. So they should remain equally likely in revised model as well, so each one of them should have now probability 1/6 since there's 6 of them:



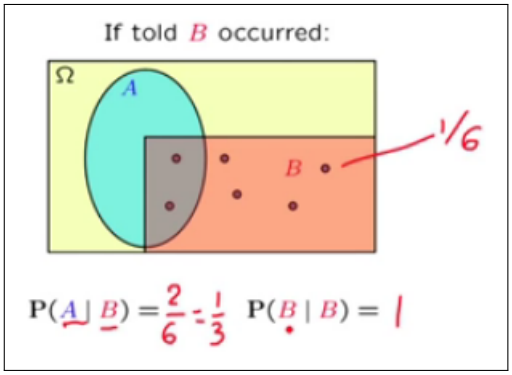
And this is our revised model, the conditional probability law. 0 probability to outcomes outside B, and probability 1/6 to each one of the outcomes

that is inside the event B.

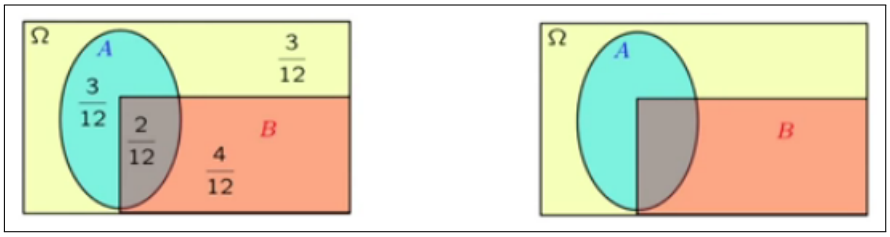
Let us write now this down mathematically. We will use this notation to describe the conditional probability of an event A given that some other event B is known to have occurred: $P(A|B)$. We read this expression as probability of A given B.

So what are these conditional probabilities in our example? So in the new model, where these outcomes are equally likely, we know that event A can occur in two different ways. Each one of them has probability 1/6. So the probability of event A is 2/6 which is the same as 1/3.

How about event B, $P(B|B)$? Well, B consists of 6 possible outcomes each with probability 1/6. So event B in this revised model should have probability equal to 1. Of course, this is just saying the obvious. Given that we already know that B has occurred, the probability that B occurs in this new model should be equal to 1.



How about now, if the sample space does not consist of equally likely outcomes, but instead we're given the probabilities of different pieces of the sample space, as in this example:



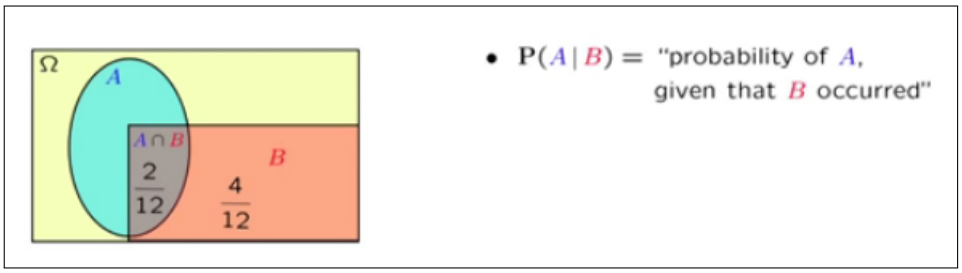
Notice here that the probabilities are consistent with what was used in the original example. So this part of A that lies outside B has probability 3/12, but in this case I'm not telling you how that probability is made up. I'm not telling you that it consists of 3 equally likely outcomes. So all I'm telling you is that the collective probability in this region is 3/12. The total probability of A is, again, 5/12 as before. The total probability of B is 2 plus 4 equals 6/12, exactly as before. So it's a sort of similar situation as before. How should we revise

our probabilities and create — construct — conditional probabilities once we are told that event B has occurred?

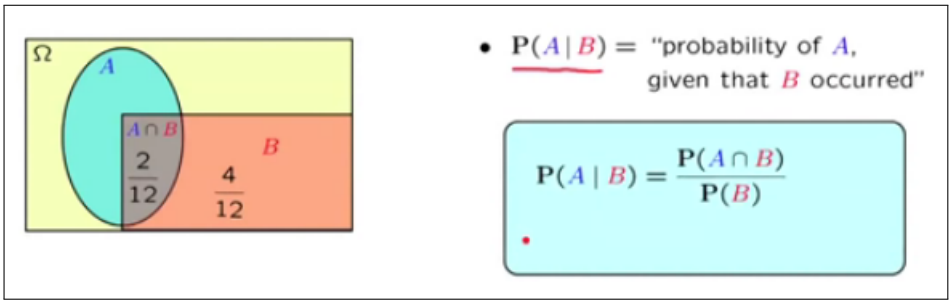
First, this relation, $P(B|B) = 1$, should remain true. Once we are told that B has occurred, then B is certain to occur, so it should have conditional probability equal to 1.

How about the conditional probability of A given that B has occurred? Well, we can reason as follows. In the original model, and if we just look inside event B, those outcomes that make event A happen had a collective probability which was $1/3$ of the total probability assigned to B. So out of the overall probability assigned to B, $1/3$ of that probability corresponds to outcomes in which event A is happening. So therefore, if I tell you that B has occurred, I should assign probability equal to $1/3$ that event A is also going to happen. So that, given that B happened, the conditional probability of A given B should be equal to $1/3$.

By now, we should be satisfied that this approach is a reasonable way of constructing conditional probabilities. But now let us translate our reasoning into a formula. So we wish to come up with a formula that gives us the conditional probability of an event given another event:



The particular formula that captures our way of thinking, as motivated before, is the following:

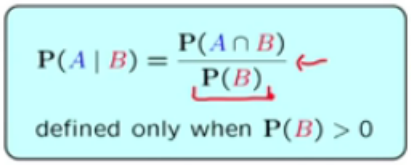


Out of the total probability assigned to B — which is this $P(B)$ — we ask the question, which fraction of that probability is assigned to outcomes under which event A also happens, $P(A \cap B)$? So we are living inside event B, but within that event, we look at those outcomes for which event A also happens. So this is the intersection of A and B. And we ask, out of the total probability

of B, what fraction of that probability is allocated to that intersection of A with B?

So this formula, this definition, captures our intuition of what we did before to construct conditional probabilities in our particular example. Let us check that the definition indeed does what it's supposed to do. In this example, the probability of the intersection was 2/12 and the total probability of B was 6/12, which gives us 1/3, which is the answer that we had gotten intuitively a little earlier.

At this point, let me also make a comment that this definition of conditional probabilities makes sense only if we do not attempt to divide by zero, $P(B) > 0$:



That this, only if the event B on which we're conditioning, has positive probability. If B, *if an event B has 0 probability*, then conditional probabilities given B will be left *undefined*.

And one final comment. This is a *definition*: $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$. It's *not a theorem*. What does that mean? It means that there is no question whether this equality is correct or not. It's just a definition. There's no issue of correctness. The earlier argument that we gave was just a motivation of the definition[7]. We tried to figure out what the definition should be if we want to have a certain intuitive and meaningful interpretation of the conditional probabilities. Let us now continue with a simple example.

4.2.3 Exercise: Conditional probabilities

Exercise 4.2.3-1: Conditional probabilities
Are the following statements true of false?

a) If Ω is finite and we have a discrete uniform probability law, and if $B \neq 0$, then the conditional probability law on B, given that B occurred, is also discrete uniform.

- ◯ True
- ◯ False

b) If Ω is finite and we have a discrete uniform probability law, and if $B \neq 0$, then the conditional probability law on Ω, given that B occurred, is also discrete uniform.

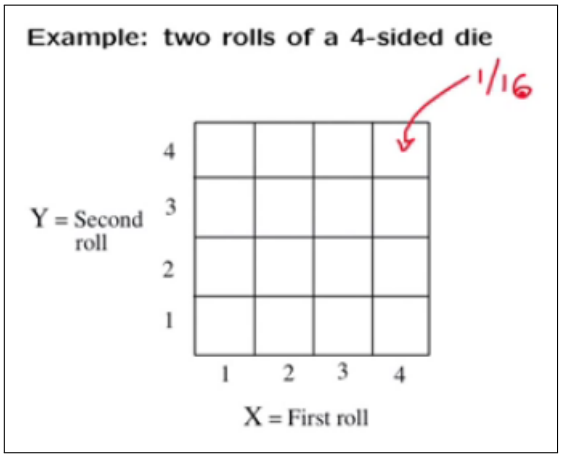
[7] $A \triangleq B$, same as $A \stackrel{\text{def}}{=} B$, means "A is by definition equal to B".

◯ True
◯ False

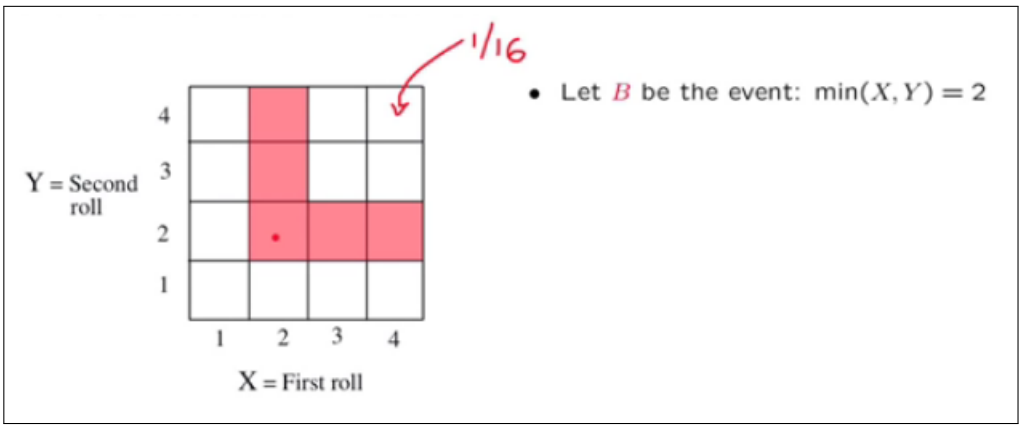
4.2.4 A die roll example

Video: A die roll example (transcripts)

This is a simple example where we want to just apply the formula for conditional probabilities and see what we get. The example involves a four-sided die, if you can imagine such an object, which we roll twice, and we record the first roll, and the second roll. So there are 16 possible outcomes. We assume to keep things simple, that each one of those 16 possible outcomes, each one of them has the same probability, so each outcome has the probability 1/16:

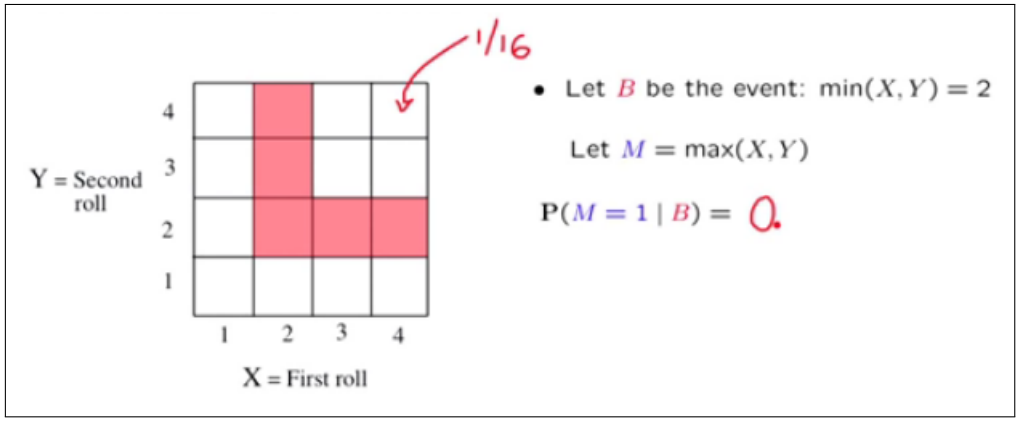


Let us consider now a particular event B on which we're going to condition. This is the event under which the smaller of the two die rolls is equal to 2, $B = \min(X, Y)$, which means that one of the dice must have resulted in two, and the other die has resulted in something which is 2 or larger. So this can happen in multiple ways. And here are the different ways that it can happen:

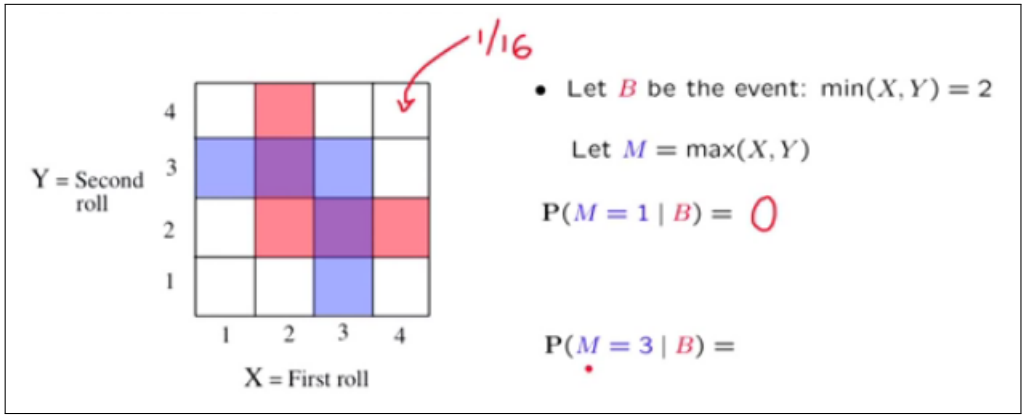


So at 2, 2, or 2, 3, or 2, 4; then a 3, 2 and a 4, 2. All of these are outcomes in which one of the dice has a value equal to 2, and the other die is at least as large. So we condition on this event. This results in a conditional model where each one of those five outcomes are equally likely since they used to be equally likely in the original model.

Now let's look at this quantity. The maximum of the two die rolls — that is, the largest of the results $M = \max(X, Y)$. And let us try to calculate the following quantity — $P(M = 1|B)$, the conditional probability that the maximum is equal to 1 given that the minimum is equal to 2. So this is the conditional probability of this particular outcome. Well, this particular outcome cannot happen. If I tell you that the smaller number is 2, then the larger number cannot be equal to 1, so this outcome is impossible, and therefore this conditional probability is equal to 0:

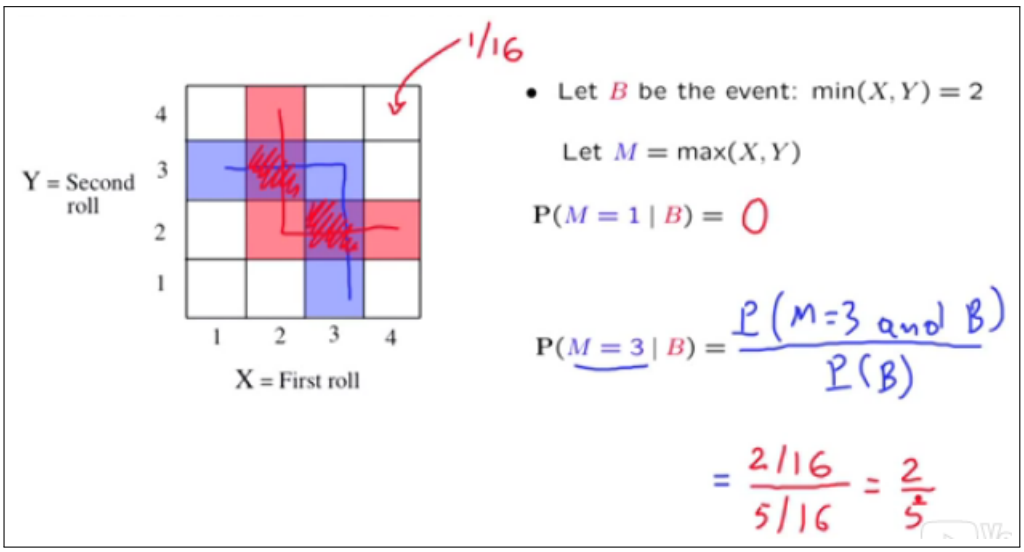


Let's do something a little more interesting. Let us now look at the conditional probability that the maximum is equal to 3 given the information that event B has occurred, $P(M = 3|B)$. It's best to draw a picture and see what that event corresponds to. M is equal to 3 — the maximum is equal to 3 — if one of the dice resulted in a 3, and the other die resulted in something that's 3 or less. So this event here corresponds to the blue region in this diagram:



Now let us try to calculate the conditional probability by just following the definition. The conditional probability of one event given another is the probability that both of them — both of the two events — occur, divided by the probability of the conditioning event. That is, out of the total probability in the conditioning event, we ask, what fraction of that probability is assigned to outcomes in which the event of interest is also happening?

So what is this event? The maximum is equal to 3, which is the blue event. And simultaneously, the red event is happening. These two events intersect only in two places. This is the intersection of the two events:



And the probability of that intersection is 2 out of 16, since there's 16 outcomes and that event happens only with two particular outcomes. So this gives us 2/16 in the numerator. How about the denominator? Event B consists of a total of five possible outcomes. Each one has probability 1/16, so this is 5/16, so the final answer is 2/5.

We could have gotten that same answer in a simple and perhaps more intuitive way. In the original model, all outcomes were equally likely. Therefore, in the conditional model, the five outcomes that belong to B should also be equally likely. Out of those five, there's two of them that make the event of interest to occur. So *given that we live in B*, there's two ways out of five that the event of interest will materialize. So the event of interest has conditional probability [equal to] 2/5.

4.2.5 Exercise: Conditional probabilities in a continuous model

Exercise 4.2.5-1: Conditional probabilities in a continuous model
Let the sample space be the unit square, $\Omega = [0, 1]^2$, and let the probability of a set be the area of the set. Let A be the set of points $(x, y) \in [0, 1]^2$ for which $y \leq x$. Let be B the set of points for which $x \leq 1/2$. Find $P(A|B)$.

4.2.6 Conditional probabilities obey the same axioms

Video: Conditional probabilities obey the same axioms (transcripts)

I now want to emphasize an important point. Conditional probabilities are *just the same as ordinary probabilities applied to a different situation*. They do not taste or smell or behave any differently than ordinary probabilities. What do I mean by that? I mean that they *satisfy the usual probability axioms*.

For example, ordinary probabilities must also be non-negative. Is this true for conditional probabilities? Of course it is true, because conditional probabilities are defined as a ratio of two probabilities. Probabilities are non-negative. So the ratio will also be non-negative, of course as long as it is well-defined. And here we need to remember that we only talk about conditional probabilities when we condition on an event that itself has positive probability:

$$\mathbf{P}(A \mid B) \geq 0 \qquad \text{assuming } \mathbf{P}(B) > 0$$

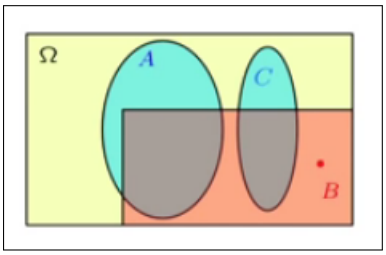
How about another axiom? What is the probability of the entire sample space, given the event B? Let's check it out. By definition, the conditional probability is the probability of the intersection of the two events involved divided by the probability of the conditioning event. Now, what is the intersection of omega with B? B is a subset of omega. So when we intersect the two sets, we're left just with B itself. So the numerator becomes the probability of B. We're dividing by the probability of B, and so the answer is equal to 1. So indeed, the sample space has unit probability, even under the conditional model:

$$\mathbf{P}(\Omega \mid B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

Now, remember that *when we condition on an event B, we could still work with the original sample space*. However, possible outcomes that do not belong to B are considered impossible, so we might as well think of B itself as being our sample space. If we proceed like that and think now of B as being our new sample space, what is the probability of this new sample space in the conditional model? Let's apply the definition once more. It's the probability of the intersection of the two events involved, B intersection B, divided by the probability of the conditioning event. What is the numerator? The intersection of B with itself is just B, so the numerator is the probability of B. We're dividing by the probability of B. So the answer is, again, 1:

$$\mathbf{P}(B \mid B) = \frac{\mathbf{P}(B \cap B)}{\mathbf{P}(B)} = 1$$

Finally, we need to check the additivity axiom. Recall what the additivity axiom says. If we have two events, two subsets of the sample space that are disjoint, then the probability of their union is equal to the sum of their individual probabilities. Is this going to be the case if we now condition on a certain event?



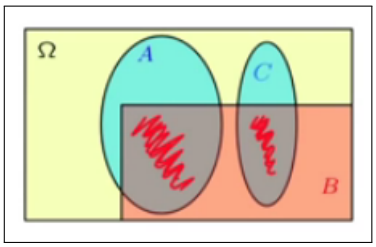
What we want to prove is the following statement:

$$\text{If } A \cap C = \emptyset, \quad \text{then } \mathbf{P}(A \cup C \mid B) = \mathbf{P}(A \mid B) + \mathbf{P}(C \mid B)$$

If we take two events that are disjoint, they have empty intersection, then the probability of the union is the sum of their individual probabilities, but where now the probabilities that we're employing are the conditional probabilities, given the event B. So let us verify whether this relation, this fact is correct or not. Let us take this quantity and use the definition to write it out. By definition, this conditional probability is the probability of the intersection of the first event of interest, the one that appears on this side of the conditioning, intersection with the event on which we are conditioning. And then we divide by the probability of the conditioning event, B:

$$\text{If } A \cap C = \emptyset, \quad \text{then } \mathbf{P}(A \cup C \mid B) = \mathbf{P}(A \mid B) + \mathbf{P}(C \mid B)$$
$$= \frac{\mathbf{P}((A \cup C) \cap B)}{\mathbf{P}(B)}$$

Now, let's look at this quantity, what is it? We're taking the union of A with C, and then intersect it with B. This union consists of these two pieces:



When we intersect with B, what is left is these two pieces here. So A union C intersected with B is the union of two pieces. One piece is A intersection B, this piece here. And another piece, which is C intersection B, this is the second piece here. So here we basically used a set theoretic identity. And now we divide by the same [denominator] as before:

If $A \cap C = \emptyset$, then $\mathbf{P}(A \cup C \mid B) = \mathbf{P}(A \mid B) + \mathbf{P}(C \mid B)$

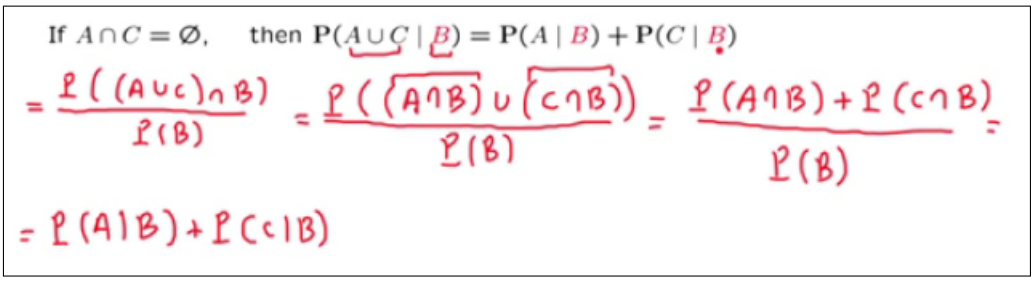
$= \frac{P((A \cup C) \cap B)}{P(B)} = \frac{P((A \cap B) \cup (C \cap B))}{P(B)}$.

And now let us continue. Here's an interesting observation. The events A and C are disjoint. The piece of A that also belongs in B, therefore, is disjoint from the piece of C that also belongs to B. Therefore, this set here and that set here are disjoint. Since they are disjoint, the probability of their union has to be equal to the sum of their individual probabilities. So here we're using the additivity axiom on the original probabilities to break this probability up into two pieces:

If $A \cap C = \emptyset$, then $\mathbf{P}(A \cup C \mid B) = \mathbf{P}(A \mid B) + \mathbf{P}(C \mid B)$

$= \frac{P((A \cup C) \cap B)}{P(B)} = \frac{P((A \cap B) \cup (C \cap B))}{P(B)} = \frac{P(A \cap B) + P(C \cap B)}{P(B)}$.

And now we observe that here we have the ratio of an intersection by the probability of B. This is just the conditional probability of A given B using the definition of conditional probabilities. And the second part is the conditional probability of C given B, where, again, we're using the definition of conditional probabilities:



So we have indeed checked that this additivity property is true for the case of conditional probabilities when we consider two disjoint events. Now, we could repeat the same derivation and verify that it is also true for the case of a disjoint union, of finitely many events, or even for countably many disjoint events. So we do have *finite* and *countable* additivity.

We're not proving it, but the argument is exactly the same as for the case of two events. So conditional probabilities do satisfy all of the standard axioms of probability theory. So conditional probabilities are just like ordinary probabilities. This actually has a very important implication. Since conditional probabilities satisfy all of the probability axioms, any formula or theorem that we ever derive for ordinary probabilities will remain true for conditional probabilities as well.