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) 3

1.1 Welcome to 6.431x: Unit 0 released . . . 3
1.2 Quick info . . . 3
1.3 Updates . . . 4
1.3.1 2018/09/04: Unit 1 Released; Grading Policy; Discussion Forum Guidelines . . . 4
1.3.2 2018/09/07: Unit 1 due next Tuesday; No solutions on forum please . . . 5

2 Unit 0: Overview (2018/08/28) 6

2.1 Course overview . . . 6
2.1.1 Course character and objectives . . . 6
2.1.2 Why study probability? . . . 7
2.1.3 Course contents . . . 7
2.2 Course introduction, objectives and study guide . . . 7
2.2.1 Introduction . . . 7
2.2.2 Course objectives . . . 7
2.2.3 Study guide . . . 8
2.3 Syllabus, calendar, and grading policy . . . 9
2.3.1 Syllabus . . . 9
2.3.2 Calendar . . . 12
2.3.3 Gradingp policy . . . 13
2.4 edX Tutorial . . . 14
2.4.1 Basics . . . 14
2.4.2 Courseware navigation . . . 14
2.4.3 Top-level navigation . . . 14
2.4.4 Discussion forums . . . 14
2.4.5 Summary . . . 14
2.5 Discussion forum and collaboration guidelines . . . 14

2.5.1 Discussion forum guidelines . . . 14
2.5.2 Collaboration guidelines . . . 16
2.6 Homework mechanics and standard notation . . . 16
2.6.1 Checking and submitting an answer . . . 16
2.6.2 Standard notation . . . 17
2.7 Textbook information . . . 21
2.7.1 Textbook . . . 21
2.7.2 Ordering and other information . . . 21
2.8 Micromasters, Certification, and Honor Pledge . . . 21
2.8.1 Micromasters . . . 21
2.8.2 Certification . . . 23
2.8.3 EdX Honor Code Pledge . . . 23
2.9 Entrance survey . . . 24

3 Unit 1: Probability models and axioms (2018/09/03) 25
3.1 Lecture 1: probability models and axioms . . . 25
3.1.1 Motivation . . . 25
3.1.2 Overview and slides . . . 25
3.1.3 Sample space . . . 27
3.1.4 Exercise: Sample Space . . . 29
3.1.5 Sample space examples . . . 30
3.1.6 Exercise: Tree representations . . . 34
3.1.7 Probability axioms . . . 34
3.1.8 Exercises: axioms . . . 39
3.1.9 Simple properties of probabilities . . . 39
3.1.10 Exercise: Simple properties . . . 46
3.1.11 More properties of probabilities . . . 46
3.1.12 Exercise: More properties . . . 53
3.1.13 A discrete example . . . 53
3.1.14 Exercise: Discrete probability calculations . . . 57
3.1.15 A continuous example . . . 57
3.1.16 Exercise: Continuous probability calculations . . . 61
3.1.17 Countable additivity . . . 61
3.1.18 Exercise: Using countable additivity . . . 69
3.1.19 Exercise: Uniform probabilities on the integers . . . 69
3.1.20 Exercise: On countable additivity . . . 70
3.1.21 Interpretations and uses of probabilities . . . 70
3.2 Mathematical background: Sets; sequences, limits, and series; (un)countable sets . . . 73
3.2.1 Mathematical background overview . . . 73

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 Aciom (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.

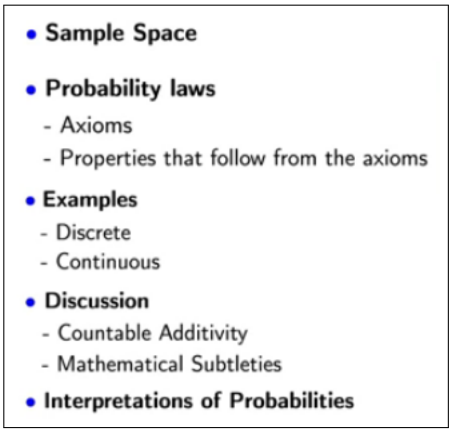
[1] The same material, in live lecture hall format, can be found here and here.

[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



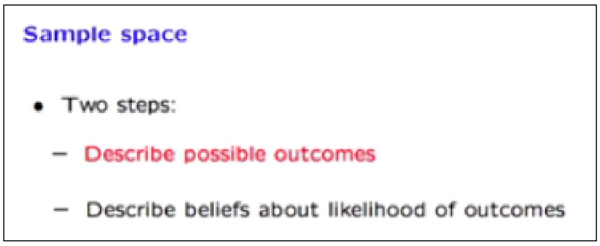
3.1.3 Sample space

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*.

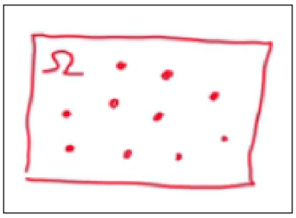
Here, we start by just talking about the first step, namely, the description of the possible outcomes of the experiment.



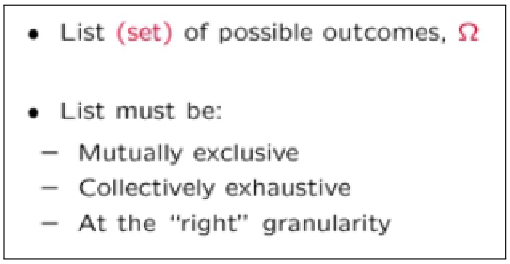
So we carry out an experiment. For example, we flip a coin. Or maybe we flip five coins simultaneously. Or maybe we roll a die.

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**.



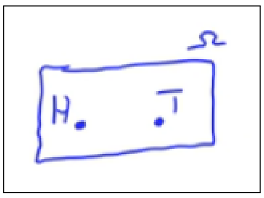
What does that mean? Mutually exclusive means that, if at the end of the experiment, I tell you that this outcome happened, then it should not be possible that this outcome also happened. At the end of the experiment, there can *only be one of the outcomes that has happened*.

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.

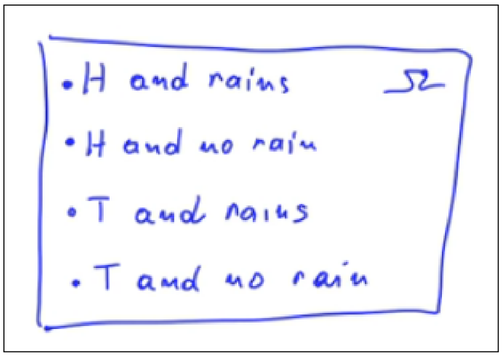
Physically *different outcomes should be distinguished in the sample space* and correspond to distinct points. But when we say physically different outcomes, what do we mean? We really mean *different in all relevant aspects* but perhaps not different in irrelevant aspects. Let's make more precise what 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:

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

◯ Yes
◯ No

$\Omega = \{$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)

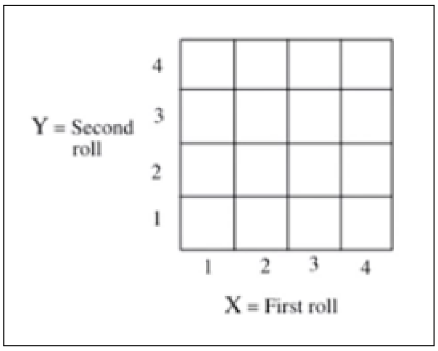
Let us now look at some examples of sample spaces. *Sample spaces are sets*. And a set can be:

- 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].

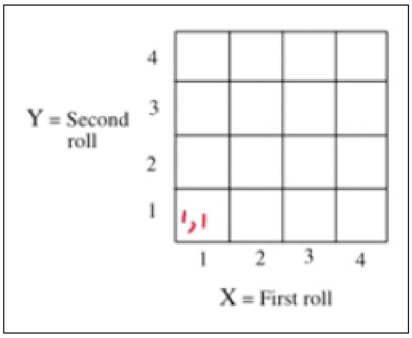
Were not dealing here with two probabilistic experiments. We're dealing with a single probabilistic experiment that involves two rolls of the die within that experiment. What is the sample space of that experiment?

Well, one possible representation is the following:

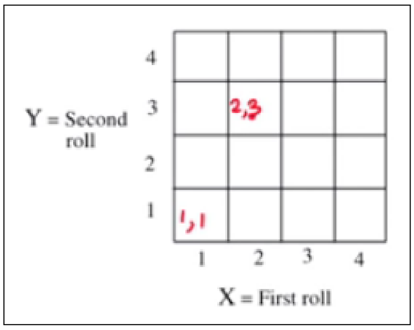


We take note of the result of the first roll. And then we take note of the result of the second roll. And this gives us a pair of numbers. Each one of the possible pairs of numbers corresponds to one of the little squares in this diagram.

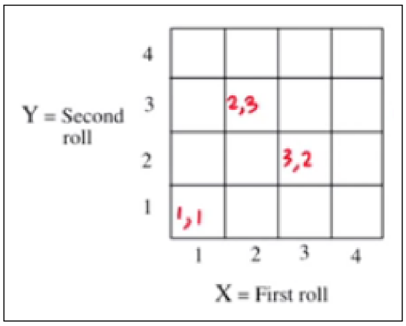
For example, if the first roll is 1 and the second is also 1, then this particular outcome has occurred:



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



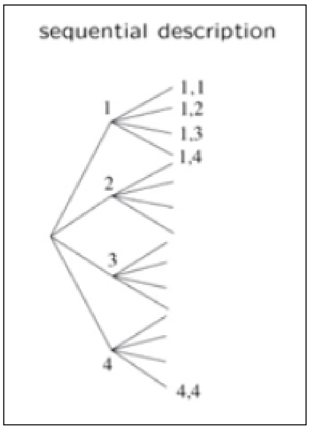
If the first roll is a 3 and then the next one is a 2, then this particular outcome occurs:



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.

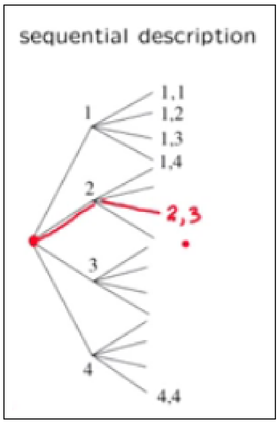
Now this is a case of a model in which *the probabilistic experiment can be described in phases or stages*. We could think about rolling the die once and then going ahead with the second roll. So we have two stages.

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*.

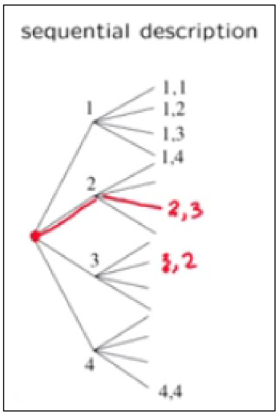


So a diagram of this kind, we call it a tree. You can think of this as the *root* of the tree from which you start. And the endpoints of the tree, we usually call them the *leaves*.

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:



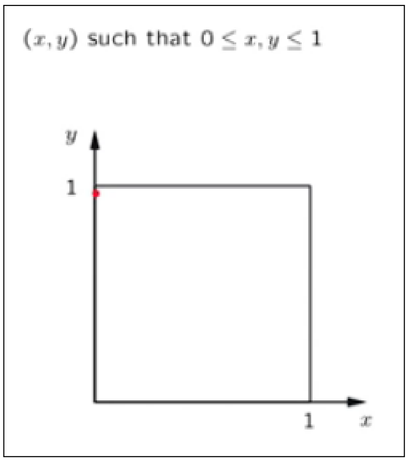
Notice that in this tree we once more have a distinction. The outcome 2 followed by a 3 is different from the outcome 3 followed by a 2, which would correspond to this particular place in the diagram:



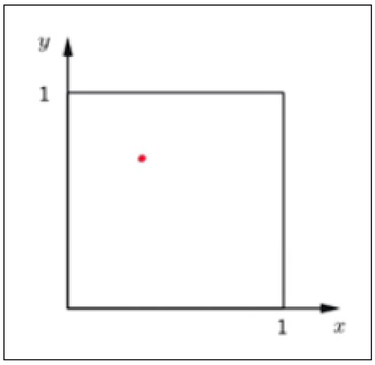
In both cases, we have 16 possible outcomes. 4 times 4 makes 16. And similarly, if you count here, the number of leaves is equal to 16.

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:



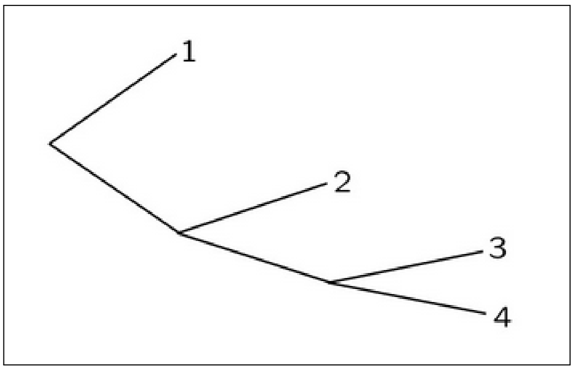
And you throw a dart on that target. And suppose that you are so skilled that no matter what, when you throw the dart, it always falls inside the target:



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

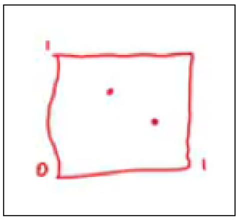
3.1.7 Probability axioms

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.

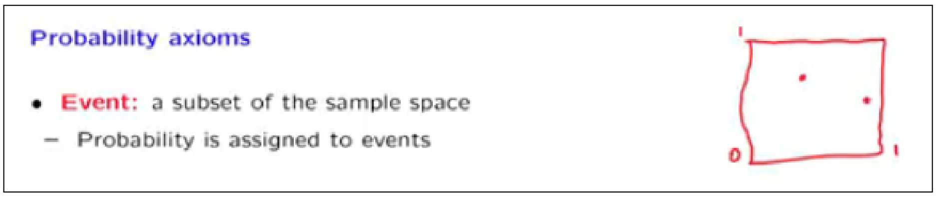
Remember the previous experiment involving a continuous sample space, which was the unit square and in which we throw a dart at random and record the point that occurred:



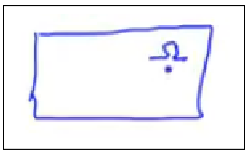
In this experiment, what do you think is the probability of a particular point?

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.

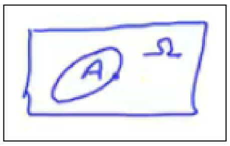
And so it's natural that *in such a continuous model any individual point should have a 0 probability*. For this reason instead of assigning probabilities to individual points, we will instead *assign probabilities to whole sets*, that is, to subsets of the sample space.



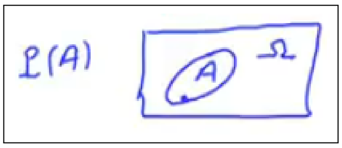
So here we have our sample space, which is some abstract set omega:



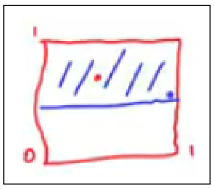
Here is a subset of the sample space, call it capital A:



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:



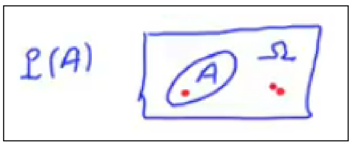
So *probabilities will be assigned to subsets*. And these will not cause us difficulties in the continuous case because even though individual points would have 0 probability, if you ask me what are the odds that my dart falls in the upper half, let's say, of this diagram, then that should be a reasonable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

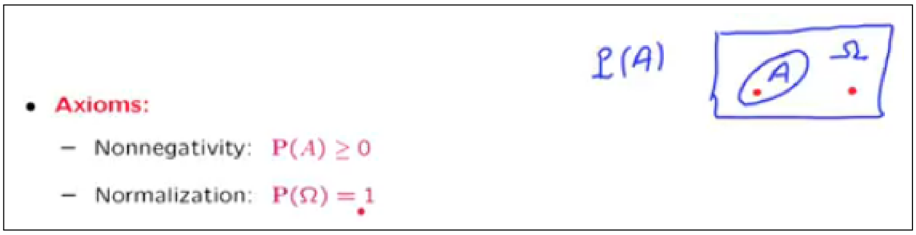
First, by convention, probabilities are always given in the range between 0 and 1. Intuitively:

- 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**.

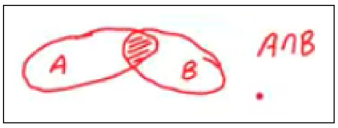
Our first axiom is a **nonnegativity axiom**, $P(A) \geq 0$. Namely, probabilities will always be non-negative numbers. It's a reasonable rule.

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: $P(\Omega) = 1$.

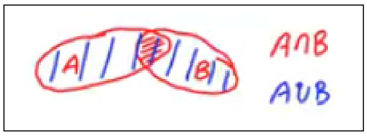


These two axioms are pretty simple and very intuitive. The more interesting axiom is the next one that says something a little more complicated.

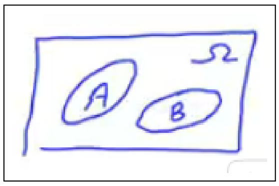
Before we discuss that particular axiom, a quick reminder about set theoretic notation. If we have two sets, let's say a set A, and another set, another set B, we use this particular notation, $A \cap B$, which we read as "A intersection B"to refer to the collection of elements that belong to both A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

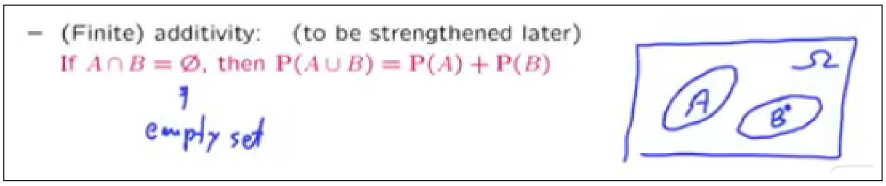


After this reminder about set theoretic notation, now let us look at the form of the third axiom. What does it say? If we have two sets, *two events*, *two subsets* of the sample space, which are *disjoint*. So here's our sample space. And here are the two sets that are disjoint:



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*:

- **Axioms:**
  - Nonnegativity: $\mathbf{P}(A) \geq 0$
  - Normalization: $\mathbf{P}(\Omega) = 1$
  - (Finite) additivity: (to be strengthened later)
    If $A \cap B = \emptyset$, then $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$

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.

Axioms	Consequences
$\mathbf{P}(A) \geq 0$	$\mathbf{P}(A) \leq 1$
$\mathbf{P}(\Omega) = 1$	$\mathbf{P}(\emptyset) = 0$

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 ($P(A) + P(A^c) = 1$):

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

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:

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

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

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

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:

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

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

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

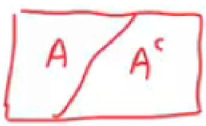
(a) $\mathbf{P}(A) \geq 0$

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

For disjoint events:

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

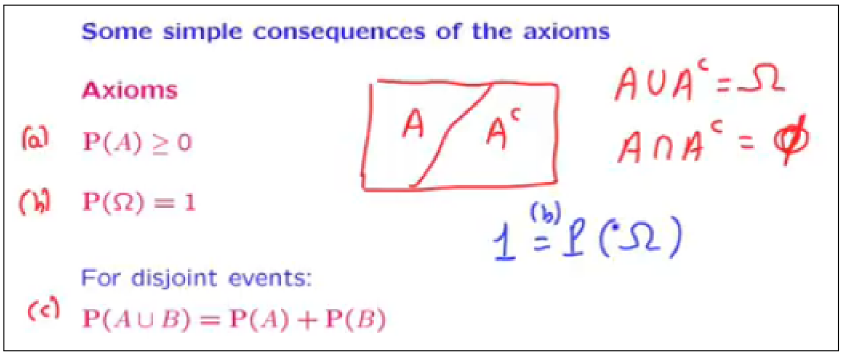
We start as follows. Let us look at the sample space and a subset of that sample space. Call it A. And consider the complement of that subset, A^c. The complement is the set of all elements that do not belong to the set A. So a set together with its complement make up everything, which is the entire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



$A \cup A^c = \Omega$

$A \cap A^c = \varnothing$

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:

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

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:

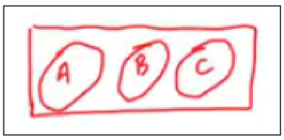
$$1 = P(\Omega) + P(\Omega^c)$$

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:

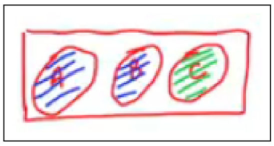
$$1 = P(\Omega) + P(\Omega^c)$$
$$1 = 1 + P(\emptyset) \Rightarrow P(\emptyset) = 0$$

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

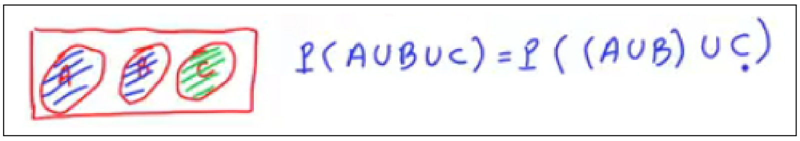
We finally come to the proof of the generalization of our additivity axiom from the case of two disjoint events to the case of three disjoint events. So we have our sample space. And within that sample space we have three events, three subsets. And these subsets are disjoint in the sense that any two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



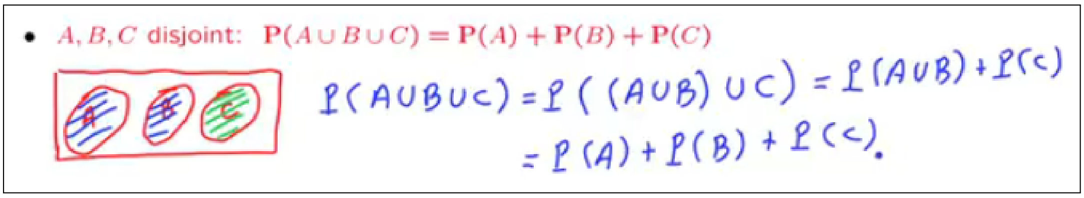
And we're interested in the probability of the union of A, B, and C, $P(A \cup B \cup C)$. How do we make progress? We have an additivity axiom in our hands, which applies to the case of the union of two disjoint sets. Here we have three of them. But we can do the following trick. We can think of the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

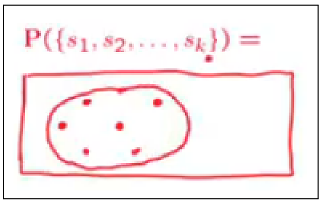


This is the proof for the case of three events. You should be able to follow this line of proof to write an argument for the case of four events and so on. And you might want to continue by induction. And eventually you should be able to prove that if the sets $A_1, \cdots A_k$ are disjoint, then the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

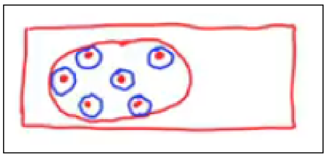
$$\text{If } A_1, \ldots, A_k \text{ disjoint} \Rightarrow P(A_1 \cup \cdots \cup A_k) = \sum_{i=1}^{k} P(A_i)$$

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

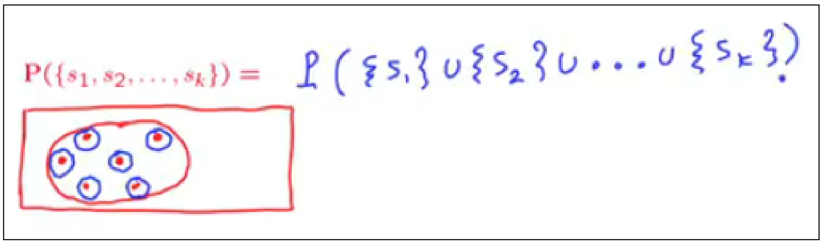
A very useful application of this comes in the case where we want to calculate the *probability of a finite set*. So here we have a sample space, and within that sample space we have some particular elements S_1, S_2, up to S_k, k of them. And these elements together form a finite set:



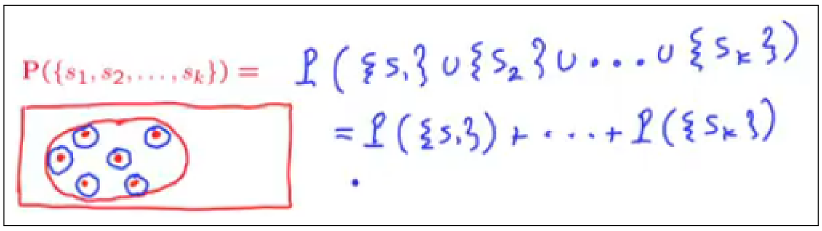
What can we say about the probability of this finite set? The idea is to take this finite set that consists of k elements and think of it as the union of several little sets that contain one element each:



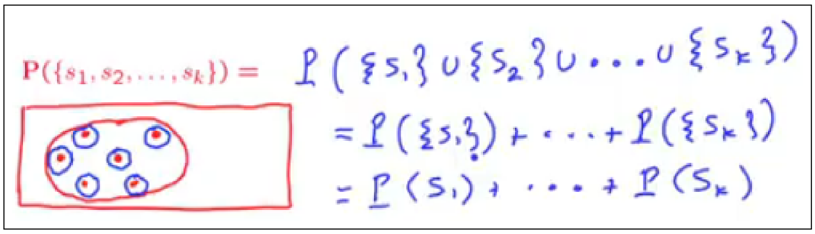
So set theoretically what we're doing is that we're taking this set with k elements and we write it as the union of a set that contains just S_1, a set that contains just the second element S_2, and so on, up to the k-th element:



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

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

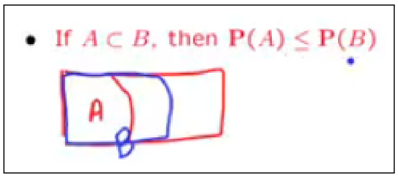
◯ True
◯ False

3.1.11 More properties of probabilities

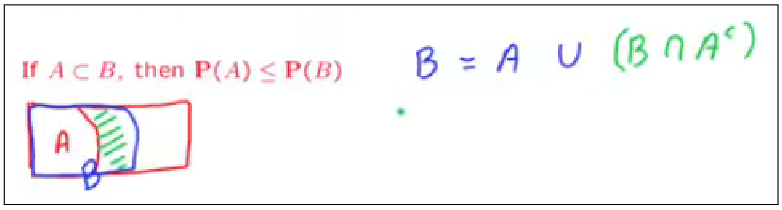
Video: More properties of probabilities (transcripts)

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

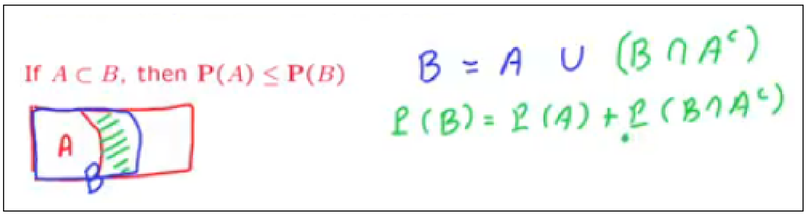
The first property is the following. If we have two sets and one set is inside the other, $A \subset B$, or $B \supset A$ — so we have a picture as follows. We have our sample space. And we have a certain set, A. And then we have a certain set, B, which is even bigger. So the set B is the bigger blue 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:



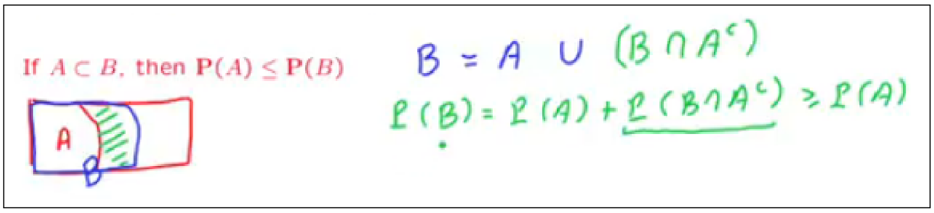
How do we prove this formally? The set B can be expressed as a union of two pieces. One piece is the set A itself. The second piece is whatever elements of B there are, that do not belong in A. What are these elements? They are elements that belong to B. And they do not belong to A, which means that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



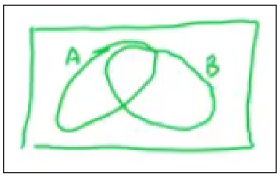
And since probabilities are non-negative, this expression here is at least as large as the probability of A. And this concludes the proof of the property that we wanted to show. Indeed, the probability of A is less than or equal to the probability of B:



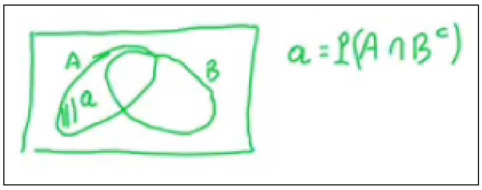
The next property we will show is the following. It allows us to write the probability of the union of two sets for the case now, where the two sets are not necessarily disjoint:

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

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

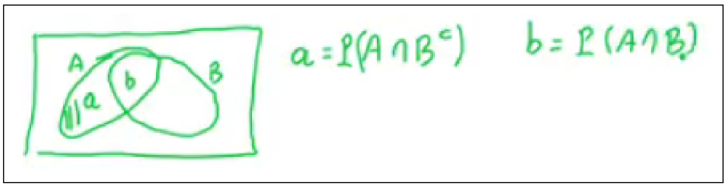


And we want to say something about the probability of the union of A and B. Now the union of A and B consists of three pieces. One piece is this one here. And that piece consists of those elements of A that do not belong to B: $A \cap B^c$. So they belong to B complement. This set has a certain probability,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

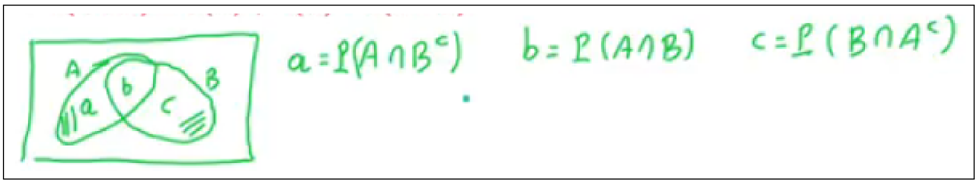


So "a" is the probability of this piece.

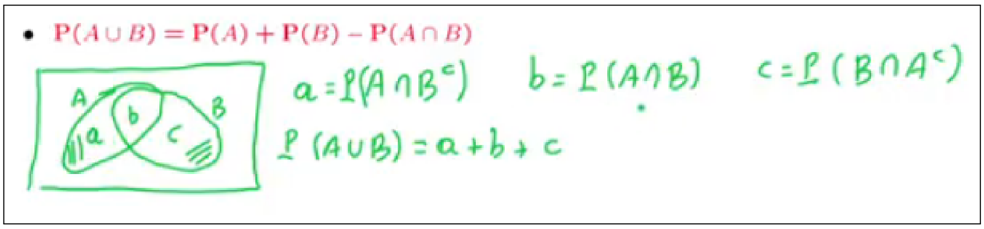
Another piece is this one here, which is the intersection of A and B, $A \cap B$. It has a certain probability that we denote by little "b". This is the probability of A intersection B, $P(A \cap B)$.



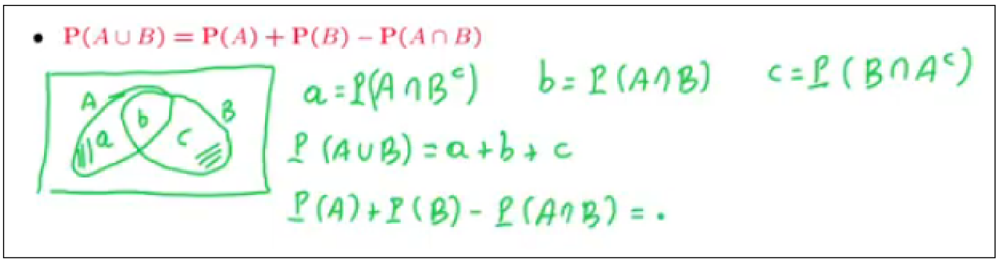
And finally, there's another piece, which is out here. And that piece has a certain probability "c". It is the probability of that set. And what is that set? That set is the following. It's that part of B that consists of elements that do not belong in A. So it's B intersection with the complement of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



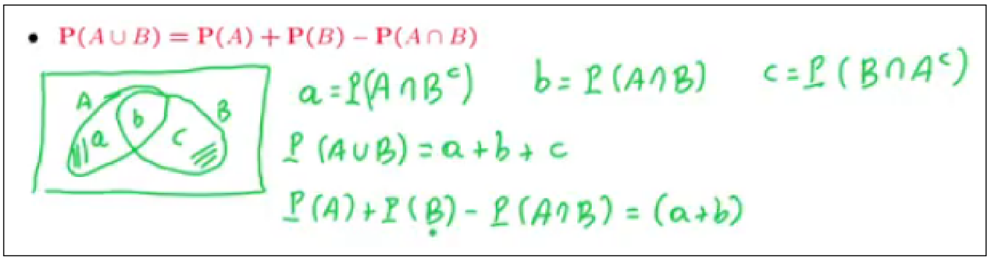
Now let's express the two sides of this equality here in terms of little "a", little "b", and little "c", and see whether we get the same thing. So the probability of A union B. A union B consists of these three pieces that have probabilities little "a", little "b", and little "c", respectively. And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



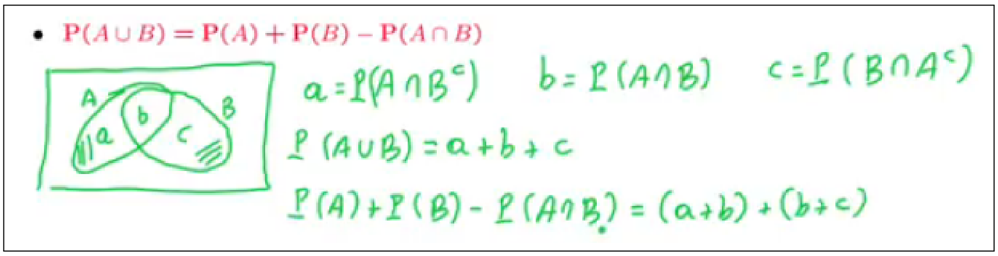
Let's look now at the right hand side of that equation and see whether we get the same thing. The probability of A plus the probability of B, minus the probability of A intersection B is equal to the following:



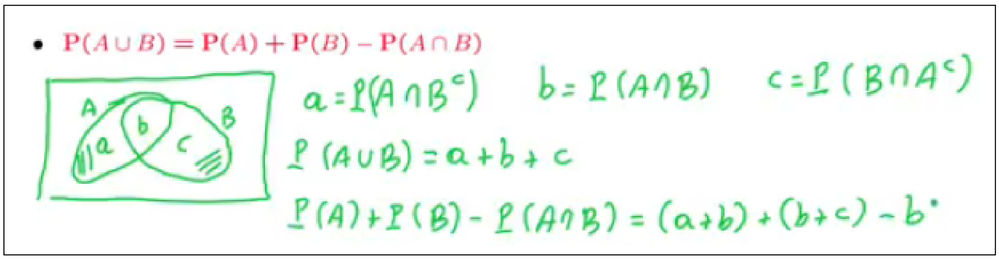
A consists of two pieces that have probabilities little "a" and little "b":



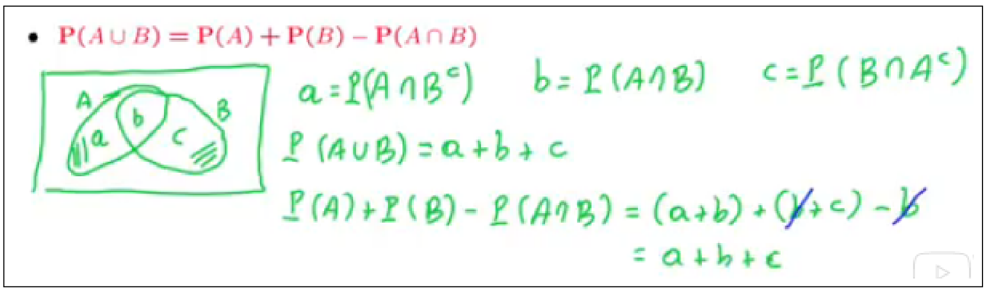
The set B consists of two pieces that have probabilities little "b" and little "c":



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,

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \overbrace{\mathbf{P}(A \cap B)}^{\geq 0}$$

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

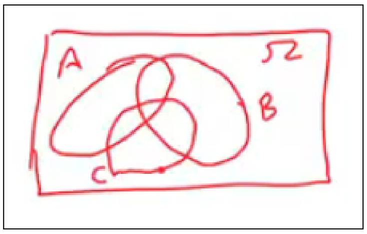
- $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$

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**.

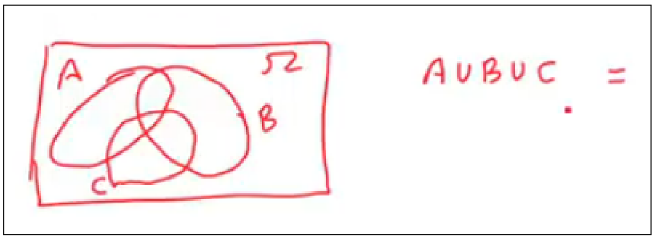
We finally consider one last consequence of our axioms. And namely, we are going to derive an expression, a way of calculating the probability of the union of three sets, not necessarily disjoint:

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

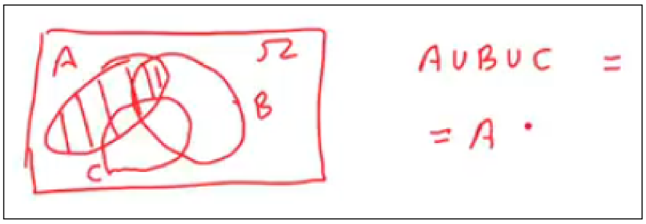
So we have our sample space. And within the sample space there are three sets — set A, set B, and set C:



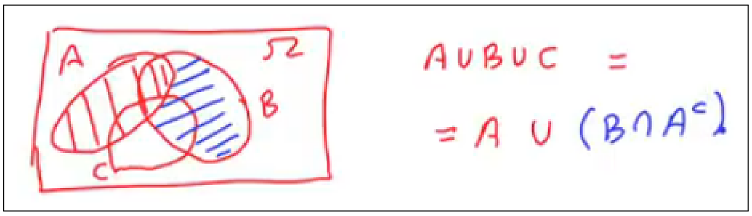
We are going to use a set theoretic relation. We are going to express the union of these three sets as the union of three disjoint pieces:



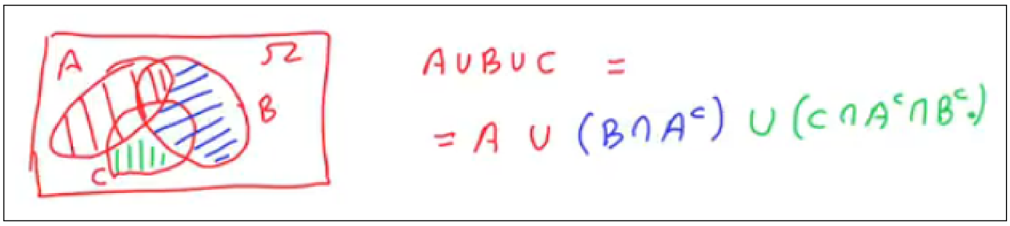
What are these disjoint pieces? One piece is the set A itself:



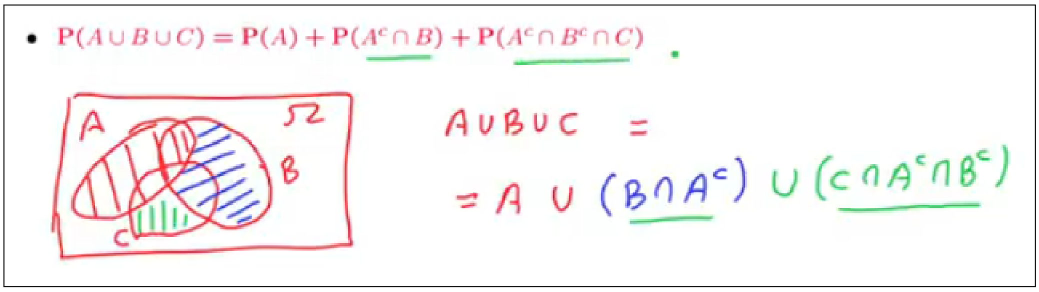
The second piece is going to be that part of B which is outside A. So this is the intersection of B with the complement of A:



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

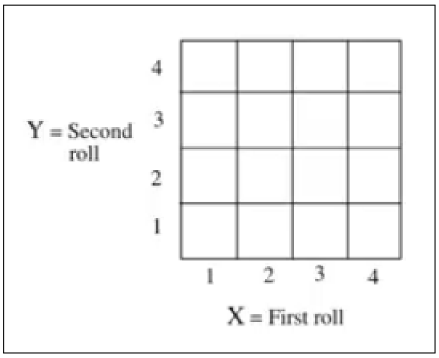
◯ True
◯ False

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

◯ True
◯ False

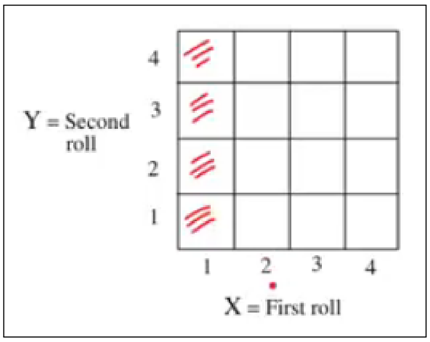
3.1.13 A discrete example

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:

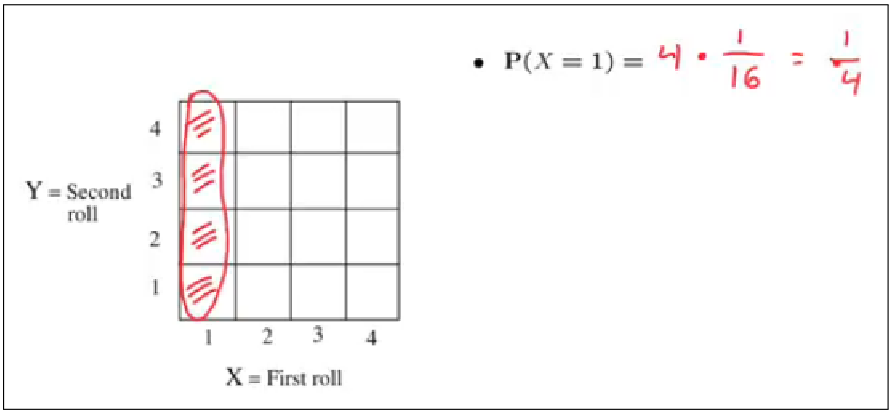


To continue, now we *need to specify a probability law*, some kind of probability assignment. To keep things simple, we're going to make the assumption that the 16 possible outcomes are all equally likely. And each outcome has a probability of 1 over 16 ($1/16$). Given this assumption, we will now proceed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

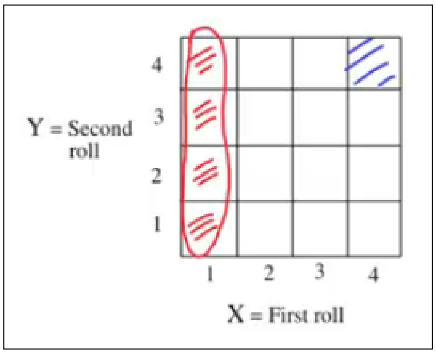


Let now Z stand for the smaller of the two numbers that came up in our two rolls. So for example, if $X = 2$ and $Y = 3$, then $Z = 2$, which is the smaller of the two:

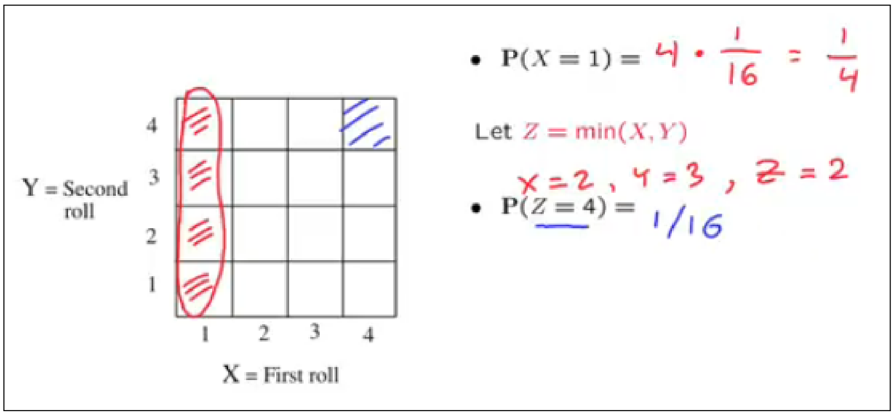
Let $Z = \min(X, Y)$

$x = 2,\ y = 3,\ z = 2$

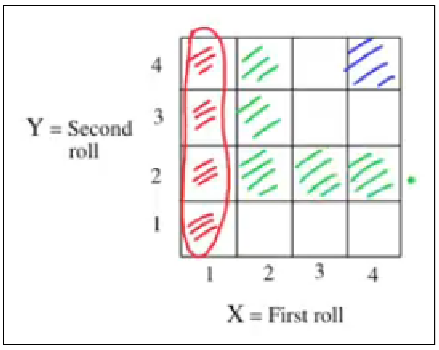
Let us try to calculate the probability that the smaller of the two outcomes is equal to 4, $Z = 4$. Now for the smaller of the two outcomes to be equal to 4, we must have that both X and Y are equal to 4. So this outcome here (in blue) is the only way that this particular event can happen:



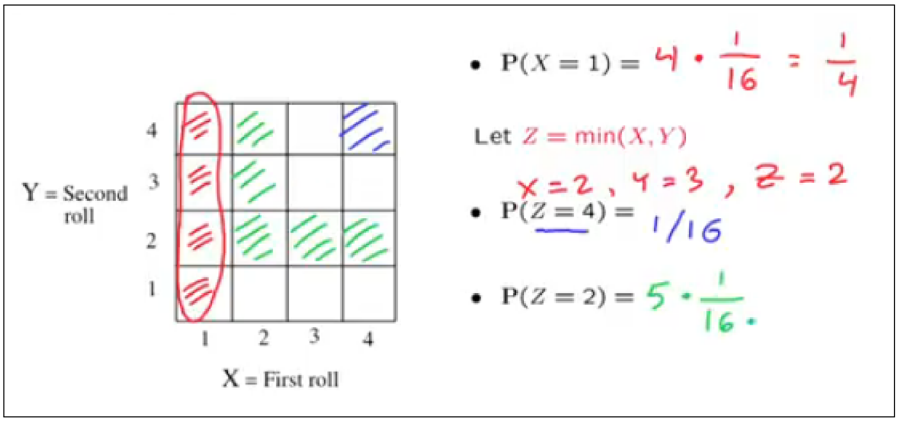
Since there's only one outcome that makes the event happen, the probability of this event is the probability of that outcome and is equal to 1/16.



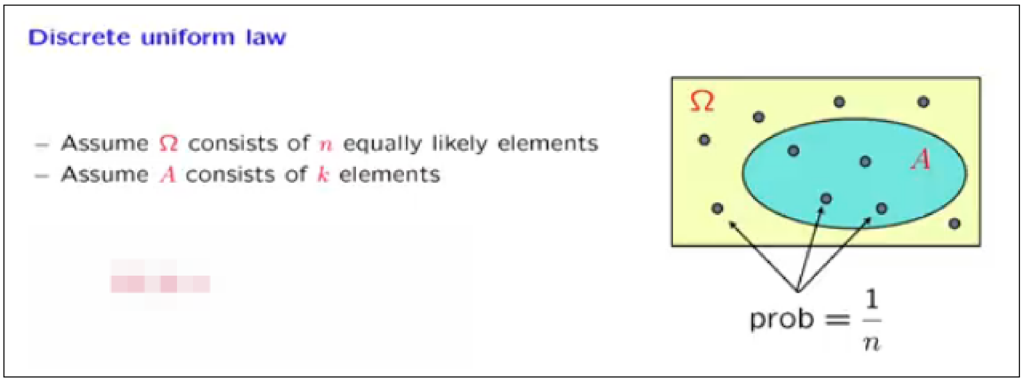
For another example, let's calculate the probability that the minimum is equal to 2. What does it mean that the minimum is equal to 2? It means that one of the dice resulted in a 2, and the other die resulted in a number that's 2 or larger. So we could have both equal to 2. We could have X equal to 2, 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:

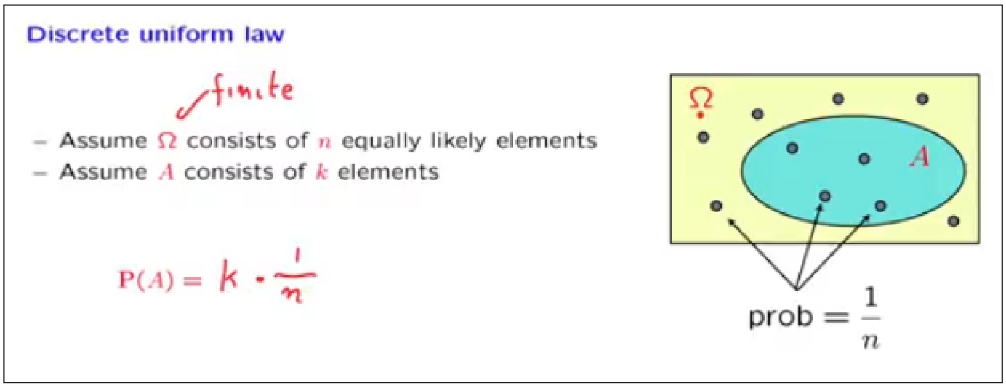


This particular example that we saw here is a special case of what is called a **discrete uniform law**. In a discrete uniform law, we have a *finite sample space*. And it has n elements. And we assume that these n elements are *equally likely*:



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:

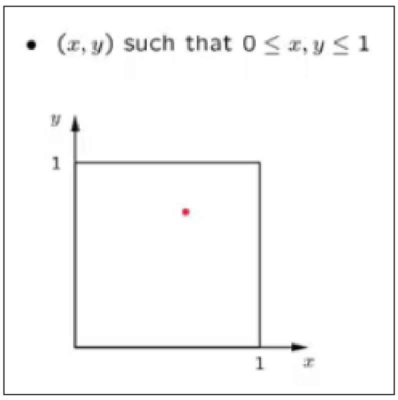
a) The value in the first roll is strictly larger than the value in the second roll.

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

3.1.15 A continuous example

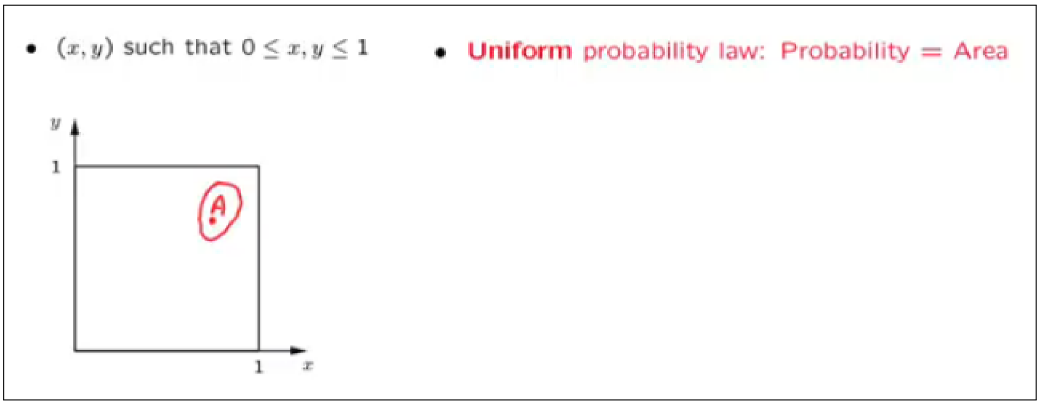
Video: A continuous example (transcripts)

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:

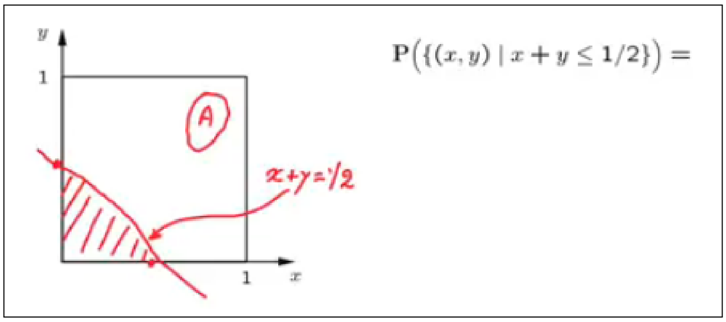


Once more, this is an arbitrary choice of a probability law. There's nothing in our assumptions so far that would force us to make this particular choice. And we just use it for the purposes of this example.

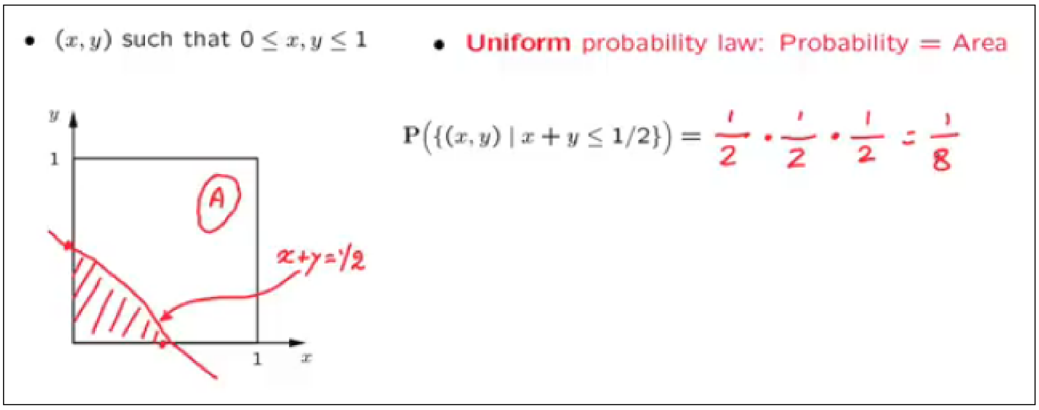
So now let us calculate some probabilities. Let us look at this event. This is the event that the sum of the two numbers that we get in our experiment is less than or equal to 1/2:

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

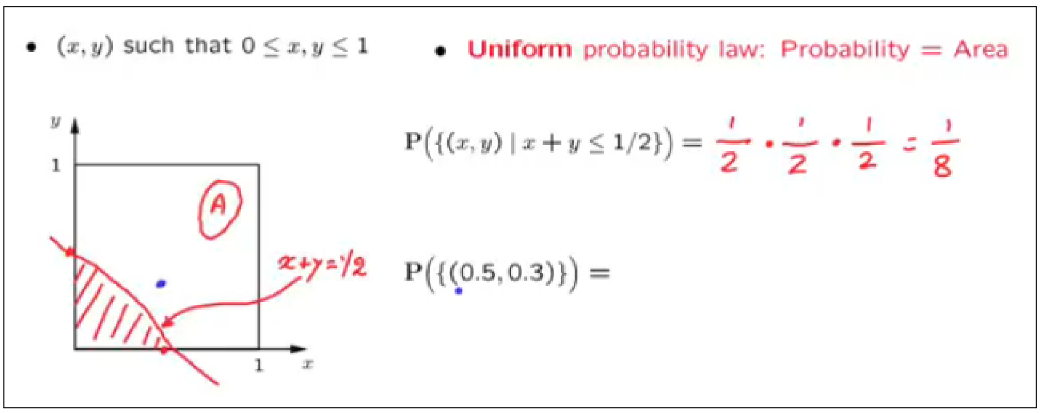
It is always useful to work in terms of a picture and to depict that event in a picture of the sample space. So in terms of that sample space, the points that make this event to be true are just a triangle that lies below the line, where this is the line, that's x plus y equals 1/2:



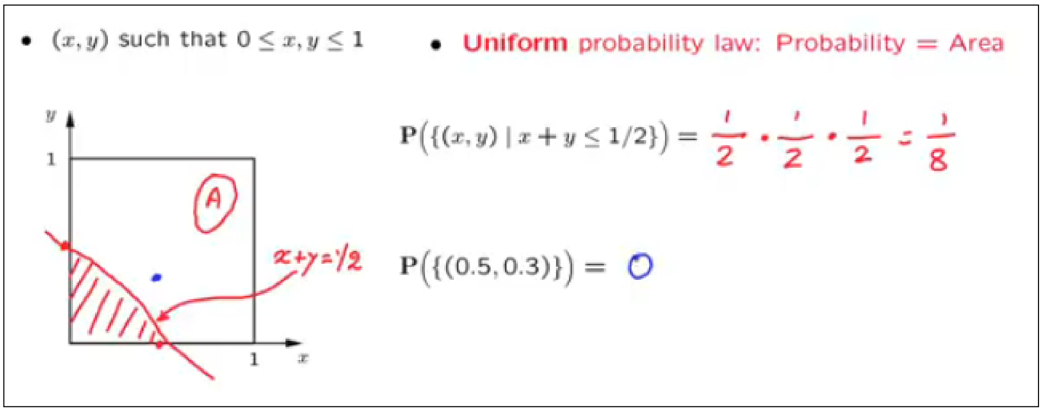
Anything below that line, these are the outcomes that make this event happen. So we're trying to find the probability of this red event. We have assumed that probability is equal to area. Therefore, the probability we're trying to calculate is the area of a triangle. And the area of a triangle is *1/2* 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.

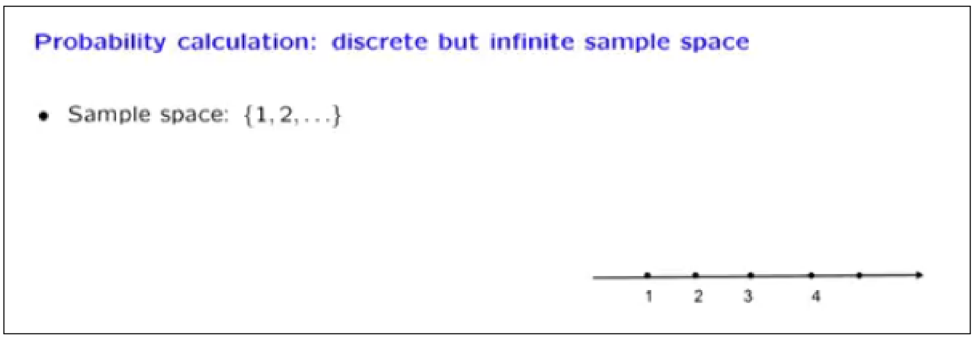
c) The value of x^2 is larger than or equal to the value of y.

3.1.17 Countable additivity

Video: Countable additivity (transcripts)

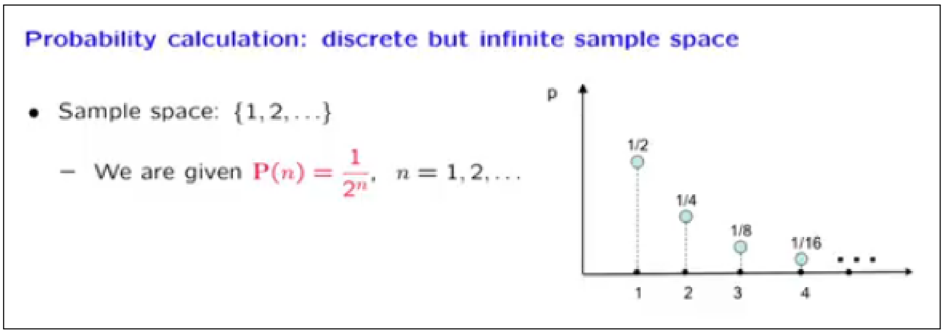
We have seen so far an example of a probability law on a *discrete and finite sample space* as well as an example with an *infinite and continuous* sample space.

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:

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

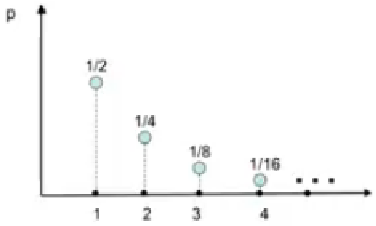
So indeed, it appears that we have the basic elements of what it would take to have a legitimate probability law.

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:

Probability calculation: discrete but infinite sample space

- Sample space: $\{1, 2, \ldots\}$
  - We are given $P(n) = \frac{1}{2^n}$, $n = 1, 2, \ldots$

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



- P(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:

P(outcome is even) $= P(\{2, 4, 6, \ldots\})$

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:

- P(outcome is even) $= P(\{2, 4, 6, \ldots\})$
$= P(\{2\} \cup \{4\} \cup \{6\} \cup \ldots)$.

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:

- P(outcome is even) $= P(\{2, 4, 6, \ldots\})$
$= P(\{2\} \cup \{4\} \cup \{6\} \cup \ldots) = P(2) + P(4) + P(6) + \ldots$

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:

- P(outcome is even) $= \mathbf{P}(\{2,4,6,\ldots\})$

$$= \mathbf{P}(\{2\}\cup\{4\}\cup\{6\}\cup\ldots) = \mathbf{P}(2)+\mathbf{P}(4)+\mathbf{P}(6)+\cdots$$

$$= \frac{1}{2^2}+\frac{1}{2^4}+\frac{1}{2^6}+\cdots = $$

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:

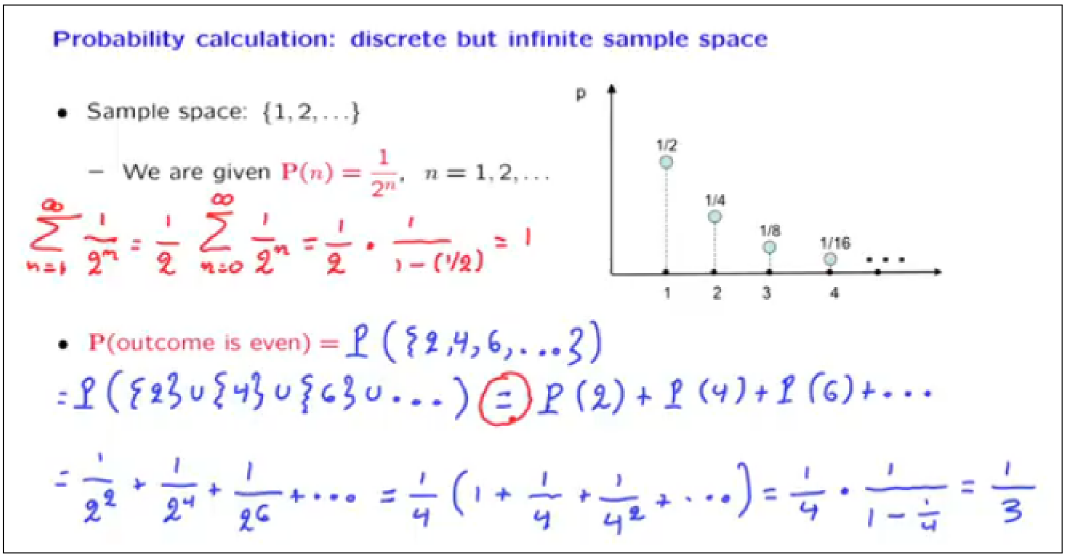
- P(outcome is even) $= \mathbf{P}(\{2,4,6,\ldots\})$

$$= \mathbf{P}(\{2\}\cup\{4\}\cup\{6\}\cup\ldots) = \mathbf{P}(2)+\mathbf{P}(4)+\mathbf{P}(6)+\cdots$$

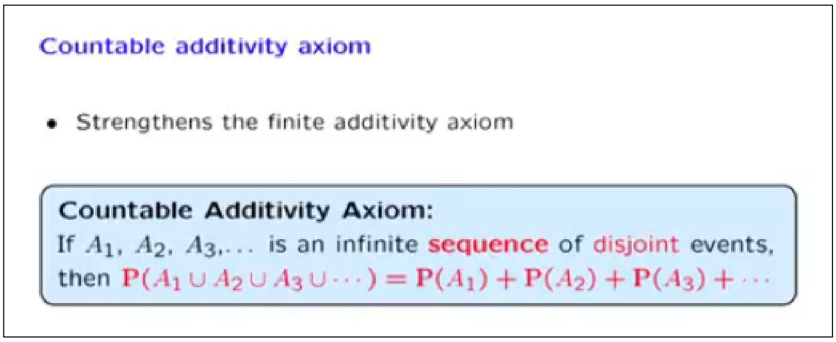
$$= \frac{1}{2^2}+\frac{1}{2^4}+\frac{1}{2^6}+\cdots = \frac{1}{4}\left(1+\frac{1}{4}+\frac{1}{4^2}+\cdots\right) = \frac{1}{4}\cdot\frac{1}{1-\frac{1}{4}} = \frac{1}{3}$$

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*.

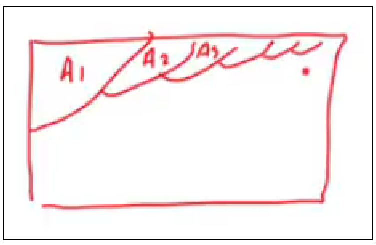
Our initial axiom talked about a disjoint union of two subsets and then later on we established a similar property for a *disjoint union of finitely many subsets*. But here we're talking about the *union of infinitely many subsets*. So this step here is not really allowed by what we have in our hands:



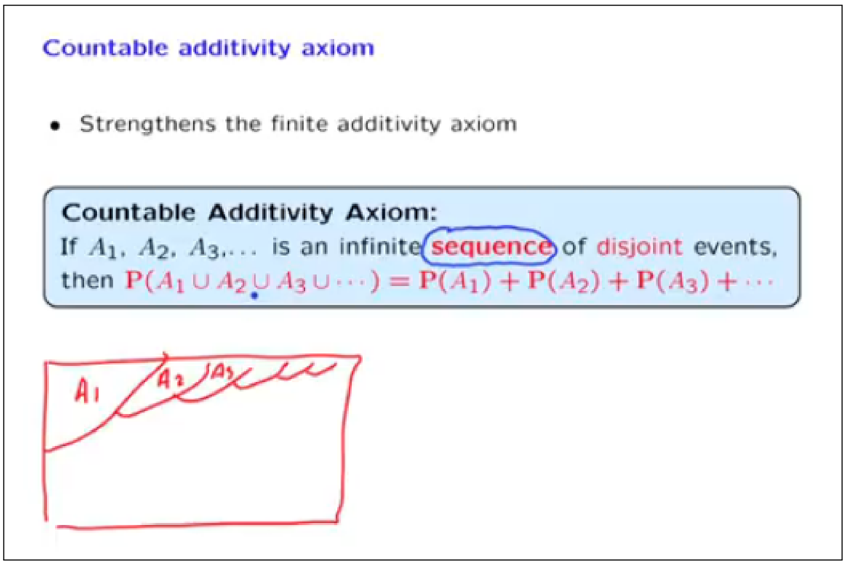
On the other hand, we would like our theory to allow this kind of calculation. The way out of this dilemma is to introduce an additional axiom that will indeed allow this kind of calculation. The axiom that we introduce is the following:



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:

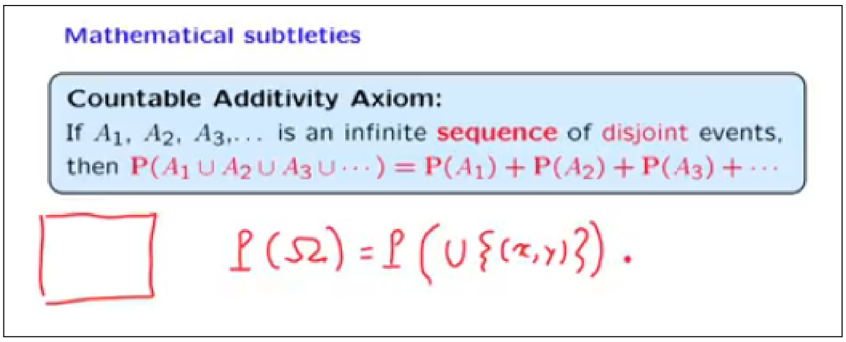


Then *the probability of the union of these events, of these infinitely many events, is* *the sum of their individual probabilities*. The key word here is the word **sequence**. Namely, these events, these sets that we're dealing with, can be arranged so that we can talk about the first event, A1, the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

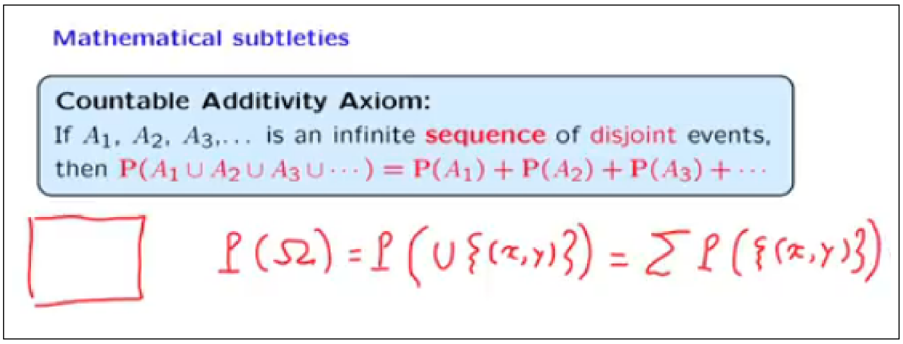


To appreciate the issue that arises here and to see why the word sequence is so important, let us consider the following calculation. Our sample space is the unit square. And we consider a model where the probability of a set is its area, as in the examples that we considered earlier.

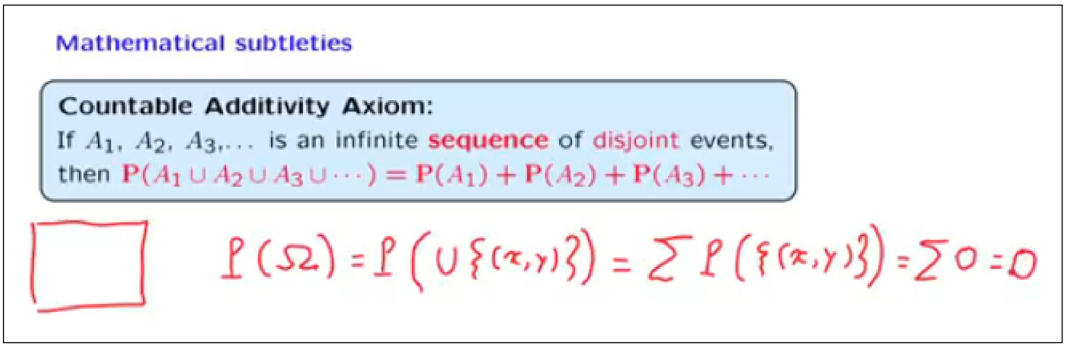
Let us now look at the probability of the overall sample space. Our sample space is the unit square and the unit square can be thought of as the union of various sets that consist of single points. So it's the union of subsets with one element each. And it's a union taken over all the points in the unit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



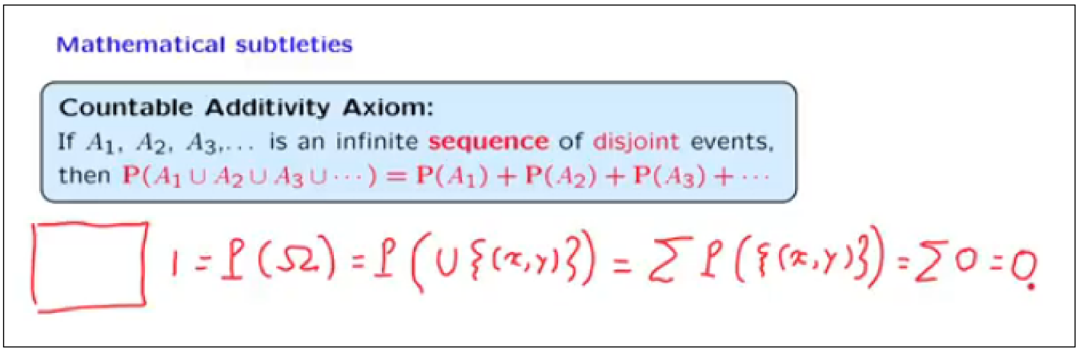
Then we think about additivity. We observe that these subsets are disjoint. If we're considering different points, then we get disjoint single element sets. And then an additivity property would tells us that the probability of this union is the sum of the probabilities of the different single element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



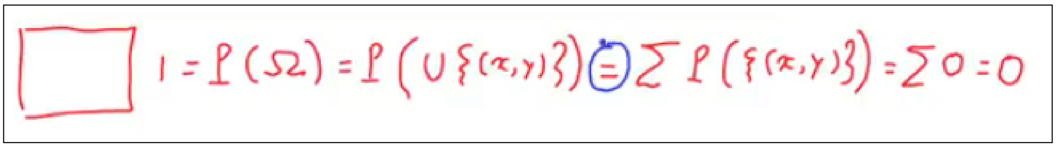
Now, as we discussed before, single element subsets have 0 probability. So we have a sum of lots of 0s and the sum of 0s should be equal to 0:



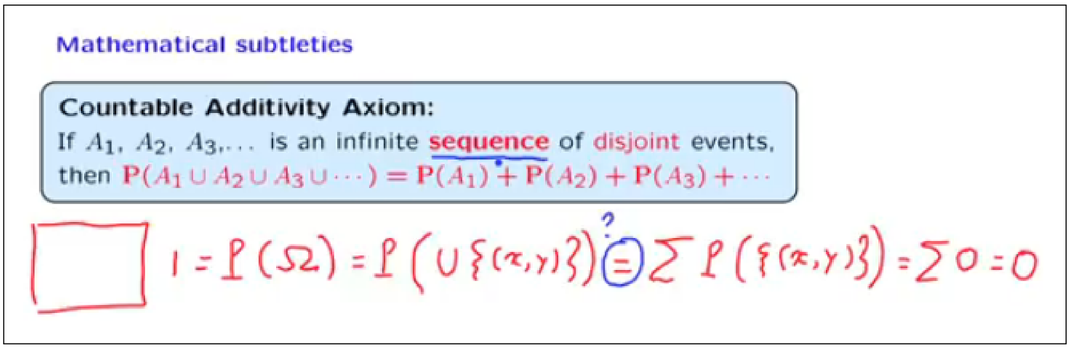
On the other hand, by the probability axioms, the probability of the entire sample space should be equal to 1. And so we have established that 1 is equal to 0:



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:

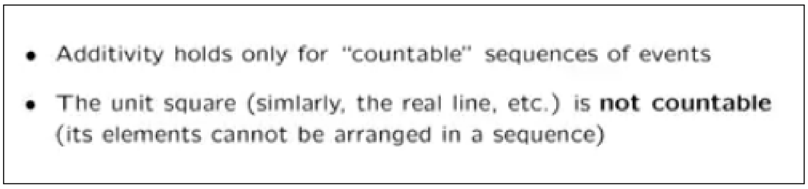


So this step here is questionable. You might argue that the unit square is the union of disjoint single element sets, which is the case that we have in additivity axioms. But the *additivity axiom only applies when we have a sequence of events*:



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**:



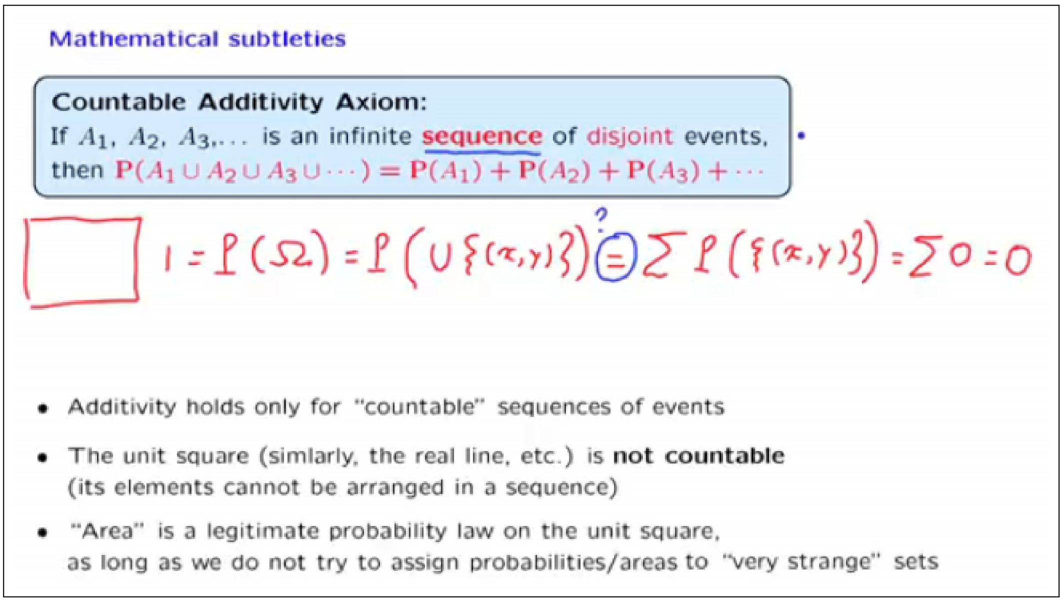
This is a deep and fundamental mathematical fact. What it essentially says is that *there are two kinds of infinite sets*:

- *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.)

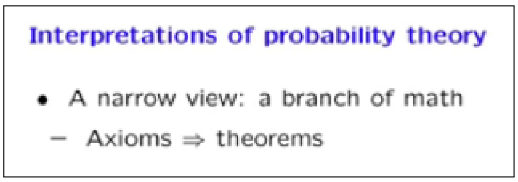
◯ Yes
◯ No

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

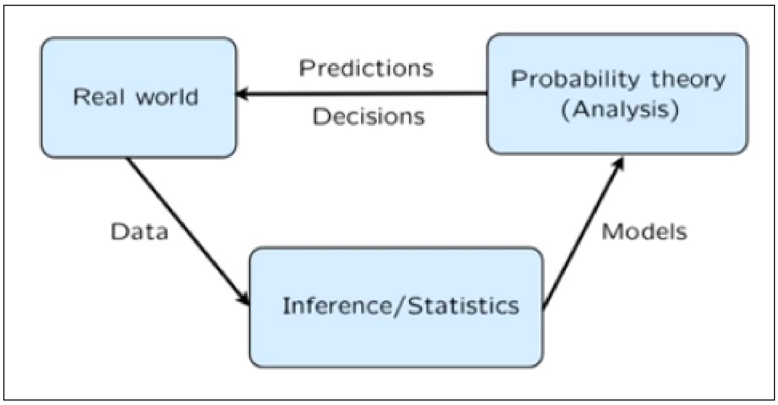
- A framework for analyzing phenomena with uncertain outcomes
  - Rules for consistent reasoning
  - Used for predictions and decisions

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.

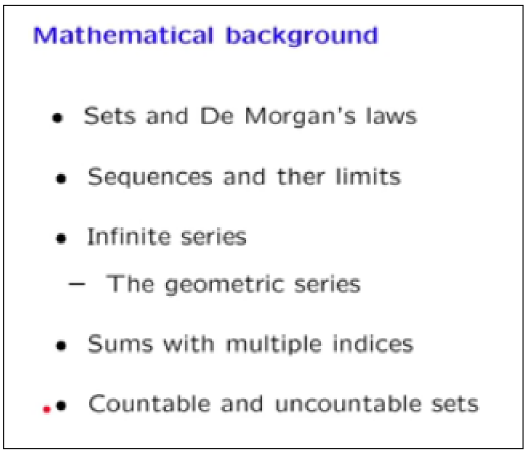
3.2 Mathematical background: Sets; sequences, limits, and series; (un)countable sets

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.