Advanced Probability, MA 587/687

(course guide)

Nikolai Chernov

The main part of the course is based on the electronic book

M. Finan, *A Probability Course for Actuaries; A Preparation for Exam P/1*

It is a large book (more than 500 pages). The purpose of this guide is to describe what needs to be covered from this book.

This guide can be used by the instructor, as it gives a suggested content of each lecture.

This guide can also be used by students. It emphasizes what they need to know from the book (and what can be ignored).

The present course builds upon Probability Theory, MA 485/585, which all the students are supposed to have taken earlier. The guide focuses on the parts of the above electronic book that present the material specific to Advanced Probability, MA 587/687.

1 Basic Definitions

2 Set Operations

3 The Fundamental Principle of Counting

4 Permutations and Combinations

Most of the content of the first four sections has been covered in Probability Theory, MA 485/585. So they only need to be quickly reviewed in class.

The only novel (or relatively novel) topics here are

- Cardinality $n(A)$ of a set A (page 11)
- Power set $\mathcal{P}(A)$ and its cardinality (page 13)
- Countable unions and intersections (page 21)
- Inclusion-Exclusion Principle for the cardinalities (page 23)

- Cartesian Product and its cardinality (page 24)
- Circular permutations (page 39)

An important note: The cardinality of the power set, $n(\mathcal{P}(A)) = 2^{n(A)}$, is related to the problem of partitioning of A into two subsets. The number of such partitions is equal to the cardinality of the power set.

Suggested homework problems:

1.3, 1.11, 2.5, 2.10, 2.17, 3.10, 4.5, 4.17, 4.18

5 Permutations and Combinations with Indistinguishable Objects

The material of this section is new, it was not included in Probability Theory, MA 485/585. This section must be covered in full.

Most important topics:

- Theorem 5.1 (and Example 5.1)
- Theorem 5.2 (and Example 5.3)
- Theorem 5.3 (with the introductory example before it)
- Remark 5.4
- Theorem 5.4 (with optional Example 5.9)

Suggested homework problems:

5.1, 5.7, 5.9, 5.12a

6 Probability: Definition and Properties

7 Properties of Probability

Most of the content of sections 6 and 7 has been covered in Probability Theory, MA 485/585. But now they can be presented in a somewhat more formal manner.

Probability Theory begins with the following basic constructions:

Probability space (or **sample space**) is the collection of all possible outcomes

Events are certain subsets of the probability space

Probability is a numerical value (a number in the interval $[0,1]$) assigned to each event

In any application where a procedure, or an experiment, or a game, may end up in more than one way (has more than one possible outcome), the above three basic constructions must be made before one can apply Probability Theory. Thus, one needs to describe ALL possible outcomes, ALL events, and specify their probabilities (or define a rule by which the probabilities can be computed).

A standard notation for the above constructions is the *triple* $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω denotes the set of all possible outcomes, $\mathcal{F}$ denotes the collection of events, and $\mathbb{P}$ denotes the probability.

In mathematical terms, Ω is just a set, whose elements are commonly denoted by ω, i.e., $\omega \in \Omega$. (The book uses symbol S instead of Ω.)

Next, $\mathcal{F}$ is a set whose elements are *subsets* of Ω, i.e., for each $A \in \mathcal{F}$ we can write $A \subset \Omega$. Each event is a subset of Ω, but not necessarily vice versa, i.e., not all subsets of Ω may qualify as events, see below.

Lastly, $\mathbb{P}$ is a function on $\mathcal{F}$ with values in $[0,1]$, i.e., $\mathbb{P}\colon \mathcal{F} \to [0,1]$. Its value on $A \in \mathcal{F}$ is denoted by $\mathbb{P}(A)$, this is what we call the *probability* of the event A.

The probability function $\mathbb{P}$ must satisfy certain requirements, stated as Axioms 1, 2, 3 in the book.

The set $\mathcal{F}$ of all the events also must satisfy certain requirements (not mentioned in the book). They are as follows:

(a) $\mathcal{F}$ must contain Ω, i.e., Ω must be an event (the largest event!)

(b) for any event $A \in \mathcal{F}$ its complement must be an event, too, i.e., $A^c \in \mathcal{F}$

(c) for any sequence of events $\{A_n\}_{n\geq 1}$, their union must be an event, i.e., $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$

It then follows that for any events $A_1, A_2, \dots$ their union, intersection, complements, differences, symmetric differences are events. In other words, one should be able to perform set operations on events. A collection $\mathcal{F}$ of subsets of Ω that satisfies these requirements is called **σ-algebra** (or **σ-field**).

There are no requirements placed on the set Ω itself.

In mathematics, the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called *measure space*, and $\mathbb{P}$ is called *measure*, or more specifically *probability measure.* Thus Probability Theory is mathematically a part of Measure Theory. The latter is covered fully in Real Analysis, MA 645/646. In this course we will try to avoid the abstract concepts of Real Analysis and focus on practical aspects.

It should be emphasized how young the mathematical theory of Probability is: Kolmogorov's axioms were introduced only in 1933. Until then Probability was mostly an empirical science (based on the "experimental" interpretation of probability described right before Kolmogorov's axioms).

Countable Subadditivity should be mentioned: for any sequence of events $\{E_n\}_{n\geq 1}$ (not necessarily mutually exclusive) we have

$$P\big(\cup_{n=1}^{\infty} E_n\big) \leq \sum_{n=1}^{\infty} P(E_n)$$

Two most common types of probability spaces can be described as follows:

> **Discrete spaces**: Ω is a finite or countable set (such as $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$). In such spaces, every subset is usually an event, i.e., we usually assume that $\mathcal{F} = \mathcal{P}(\Omega)$. Then the probability is fully determined by its values on one-point sets, because for every event $A = \{\omega_1, \omega_2, \ldots\}$ we have
>
> $$\mathbb{P}(A) = \mathbb{P}(\{\omega_1\}) + \mathbb{P}(\{\omega_2\}) + \cdots$$
>
> **Continuous spaces**: Ω is a the real line $\mathbb{R}$ (or an interval in it, such as $(a, b) \subset \mathbb{R}$), or a plane $\mathbb{R}^2$, or a space $\mathbb{R}^3$, etc., or any well-defined figure (domain) in $\mathbb{R}^k$, such as a rectangle, box, disk, ball, surface, etc. Events are "relatively good" subsets whose size (length, area, volume) can be determined (see a striking example below). In such spaces the probability of every one-point set is usually set to zero, i.e., $\mathbb{P}(\{\omega\}) = 0$ for every $\omega \in \Omega$. Therefore the probability $\mathbb{P}$ of large events $A \in \mathcal{F}$ must be defined differently. This is usually done with the help of certain functions (distribution functions, density functions, etc.).

Note that the length/area/volume cannot be determined for any subset of $\mathbb{R}$, or $\mathbb{R}^2$, or $\mathbb{R}^3$, respectively. The most spectacular example is known as

Banach–Tarski paradox: Given a solid ball in $\mathbb{R}^3$, there exists a decomposition of the ball into a finite number of non-overlapping pieces (i.e., subsets), which can then be put back together in a different way to yield two identical copies of the original ball. The reassembly process involves only moving the pieces around and rotating them, without changing their shape or size.

Figure 1: Banach–Tarski paradox illustrated

(This is often stated, colloquially, as "a pea can be chopped up and reassembled into the Sun".)

Thus if those pieces of the original ball had volume, we would observe an impossible situation: the volume doubles. This paradox demonstrates that those pieces of the ball cannot have volume, they are too "ugly".

Continuity of Probabilities should be mentioned: for any increasing sequence of events

$$E_1 \subset E_2 \subset \cdots \subset E_n \subset \cdots$$

we have

$$P\big(\cup_{n=1}^{\infty} E_n\big) = \lim_{n\to\infty} P(E_n)$$

and for any decreasing sequence of events

$$E_1 \supset E_2 \supset \cdots \supset E_n \supset \cdots$$

we have

$$P\big(\cap_{n=1}^{\infty} E_n\big) = \lim_{n\to\infty} P(E_n)$$

Draw the corresponding diagrams.

Example 6.7 was discussed in Probability Theory, MA 485/585. Generalize it as follows: suppose n objects are selected from a pool of N objects, with replacement. What is the probability that all the selected objects are distinct (no repetitions)? Answer:

$$P(\text{all distinct}) = \frac{N}{N} \cdot \frac{N-1}{N} \cdots \frac{N-n+1}{N}$$

For large N's we can approximate this probability as follows:

$$\begin{aligned} P(\text{all distinct}) &= e^{\ln\frac{N}{N}+\ln\frac{N-1}{N}+\cdots+\ln\frac{N-n+1}{N}} \\ &\approx e^{-\frac{1}{N}-\cdots-\frac{n-1}{N}} \\ &= e^{-\frac{n(n-1)}{2N}} \end{aligned}$$

We can make the following general conclusions:

- For $n \ll \sqrt{N}$ this probability is almost 1, so all the selected objects will be almost certainly distinct.
- For $n \gg \sqrt{N}$ the above probability is almost zero, so repetitions are almost unavoidable.
- The transition occurs when n is of order $\sqrt{N}$, then the above probability is neither close to zero nor close to one.

Relate this to random polls of large populations (and the popular in statistics "5% guideline" for sampling from large populations)

The inclusion-exclusion formula given in Theorem 7.1 for two events and in Theorem 7.2 for three events can be generalized to any number of events:

$$P(A_1 \cup \cdots \cup A_n) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots$$

Apply it to the following "matching problem":

Suppose n couples attend a party where at some point men are randomly paired with women for a dance. What is the probability that at least one husband dances with his own wife?

Solution: let A_i denote the event that the ith man dances with his wife. Then the probability that at least one husband dances with his wife is

$$\begin{aligned} P(A_1 \cup \cdots \cup A_n) &= \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots \\ &= n \cdot \tfrac{1}{n} - \tfrac{n(n-1)}{2} \cdot \tfrac{1}{n(n-1)} + \tfrac{n(n-1)(n-2)}{3!} \cdot \tfrac{1}{n(n-1)(n-2)} - \cdots \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots \end{aligned}$$

This is a finite sum (it has n terms), but for large n it converges to the infinite series

$$1 - \frac{1}{2!} + \frac{1}{3!} - \cdots = 1 - \frac{1}{e} \approx 0.6321$$

so that above probability is approximately $1 - \frac{1}{e} \approx 0.6321$.

Suggested homework problems:

6.11, 7.9, 7.11, 7.16

8 Probability and Counting Techniques

This section gives a nice and efficient method for solving many practical problems, so it deserves a review. Discuss Example 8.5.

Practice problems in the end of the section are either quite simple (8.1 to 8.10) or nearly impossible (8.11 to 8.13). The last three problems (all nearly impossible) are taken from actuarial exams, so it would be good to assign them. However they involve the material to be covered much later in the course, so it is just too early to assign them now. Problem 8.11 can be assigned in section 20.3, for example.

Suggested homework problem: 8.7.

9 Conditional Probability

10 Posterior Probabilities: Bayes' Formula

11 Independent Events

Most of the content of these three sections has been covered in Probability Theory, MA 485/585. So they only need be quickly reviewed in class.

Some useful notes:

Theorem 9.1 should be related to the material of Section 8.

For every fixed event A with $\mathbb{P}(A) > 0$ the function $\mathbb{P}_A \colon \mathcal{F} \to [0,1]$ defined by $\mathbb{P}_A(B) = \mathbb{P}(B|A)$ is a probability measure (different from the original $\mathbb{P}$). So the occurrence of an event A changes the probabilities of all the other events. Practical example: weather forecast for the next 10 days (in particular, the chances of rain, snow, etc.) is updated every day, as the current weather conditions are taken into account (i.e., the current events change the probabilities of the future events).

The definition of independent events is stated in the book as $\mathbb{P}(A|B) = \mathbb{P}(A)$ (given at the beginning of Section 11). It is technically incomplete, as it does not cover the important case $\mathbb{P}(B) = 0$. The standard official definition of independent events is $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ (in the book, it is stated in Theorem 11.1).

Note: the best way to illustrate independent events is to draw a rectangle (representing the probability space Ω), one horizontal strip (stretching across the rectangle, from left to right) and one vertical strip (stretching across the rectangle, from top to bottom). Show that the relative area of the intersection of the two strips is equal to the product of their relative areas.

A "philosophical" remark: *independence* in probability has some deep relation with *orthogonality* in geometry.

A historic remark: Kolmogorov's axioms of Probability Theory nicely place it in the context of Measure Theory in mathematics, i.e., Probability Theory appears to be a particular case of general Measure Theory. This point of view is quite common, but it is not exactly accurate. Kolmogorov himself said that these two disciplines evolve in parallel only until the notion of independence is introduced. After that they go separate ways. There is no such thing as "independence" in general Measure Theory, while in Probability it plays a central role (or *the* central role!).

Suggested homework problems:

9.3, 9.1, 10.3, 10.9, 11.5, 11.6, 11.8

12 Odds and Conditional Probability

This is really a tiny note, not a serious section. The part after Example 12.2 can be just ignored. No homework exercises.

13 Discrete Random Variables

14 Probability Mass Function and Cumulative Distribution Function

Most of the content of these two sections has been covered in Probability Theory, MA 485/585. But now they can be presented in a somewhat more formal manner.

A random variable can be defined as a function $X \colon \Omega \to \mathbb{R}$. Its domain is the probability space Ω, its values are real numbers. Each elementary outcome $\omega \in \Omega$ has a numerical value $X(\omega) \in \mathbb{R}$ assigned to it.

In practical situations, ω represents the actual outcome that occurs and $X(\omega)$ represents the value we *observe*. For example, ω denotes the current weather conditions and $X(\omega)$ the observed (measured) temperature of the outside air. Weather conditions determine the air temperature, but many different weather conditions may result in the same air temperature, i.e., we often have $X(\omega) = X(\omega')$ for distinct possible outcomes $\omega \neq \omega'$.

Quite commonly, the probability space Ω and the respective probability measure $\mathbb{P}$ are very difficult (or impossible) to describe, they remain unspecified, "behind the scene", while the values of X are observable, "visible". We will learn how to describe X in its own terms, without referring to Ω.

A random variable X is said to be **discrete** if its range is finite or countable, i.e., if all possible values of X can be numbered: $x_1, x_2, \ldots$. For example, they can be natural numbers, integers, rational numbers, etc.

The **probability function** (called **probability mass function** in the book) of a r.v. X is denoted by $p(x)$ and defined by $p(x) = \mathbb{P}(X = x)$. More formally,

$$p(x) = \mathbb{P}\big(\{\omega \in \Omega \colon X(\omega) = x\}\big) = \mathbb{P}\big(X^{-1}(x)\big)$$

(explain the meaning of X^{-1}, draw diagrams, give examples, such as if $f(x) = x^2$, then $f^{-1}(1) = \{1, -1\}$ and $f^{-1}(-1) = \emptyset$).

Of course, if X does not take value $x \in \mathbb{R}$, i.e., when x is *n*ot one of the $x_1, x_2, \ldots$, then the above set $X^{-1}(x)$ is empty, so its probability is zero, i.e., $p(x) = 0$. On the other hand, $p(x_1), p(x_2), \ldots$ are usually positive numbers (but some may be zero occasionally).

Note: $p(x_1) + p(x_2) + \cdots = 1$.

Given a random variable X, the p.m.f. $p(x)$ gives probabilities corresponding to individual numbers $x \in \mathbb{R}$. Likewise, we can define probabilities corresponding to intervals $(a, b) \subset \mathbb{R}$ as follows:

$$\mathbb{P}(a, b) = \mathbb{P}(a < X < b) = \mathbb{P}\big(\{\omega \in \Omega \colon X(\omega) \in (a, b)\}\big) = \mathbb{P}\big(X^{-1}(a, b)\big).$$

Quite obviously,

$$\mathbb{P}(a, b) = \sum_{x_i \in (a,b)} p(x_i)$$

Thus we now have a probability assigned to each interval (a, b). More generally, the probability of any subset $A \subset \mathbb{R}$ can be defined by

$$\mathbb{P}(A) = \mathbb{P}(X \in A) = \mathbb{P}\big(\{\omega \in \Omega \colon X(\omega) \in A\}\big) = \mathbb{P}\big(X^{-1}(A)\big)$$

and computed by

$$\mathbb{P}(A) = \sum_{x_i \in A} p(x_i)$$

Thus we obtain a probability measure on subsets of $\mathbb{R}$ (rather than on subsets of Ω were it was originally defined). This new probability measure is induced on $\mathbb{R}$ by the random variable X, and it often denoted by $\mathbb{P}_X$. It is called the **distribution** of the random variable X.

Note: $\mathbb{P}_X(A)$ can be computed in terms of values $x_1, x_2, \dots$ and their probabilities $p(x_1), p(x_2), \dots$ alone, without any reference to the original probability space Ω and the original probability measure $\mathbb{P}$ on it. Thus we can completely describe the random variable X in its own terms.

Quite commonly, the probability space Ω and the respective probability measure $\mathbb{P}$ remain unspecified, "behind the scene", and then the only "visible" part of the picture is the set of values $x_1, x_2, \dots$ of the random variable X and their probabilities $p(x_1), p(x_2), \dots$. They completely determine the distribution of X, which includes probabilities of intervals and other subsets of $\mathbb{R}$.

The **distribution function** (or **cumulative distribution function**) of a random variable X is defined by

$$F(a) = \mathbb{P}(X \leq a) = \mathbb{P}\big(\{\omega \in \Omega \colon X(\omega) \leq a\}\big) = \mathbb{P}\big(X^{-1}(-\infty, a]\big).$$

If the students have not seen distribution functions of discrete random variables in MA 485/585, discuss Examples 14.4 and 14.5 and assign Practice Problem 14.1.

Suggested homework problems: 13.7 and 13.11 (and possibly 14.1).

15 Expected Value of a Discrete Random Variable

16 Expected Value of a Function of a Discrete Random Variable

Most of the content of these two sections has been covered in Probability Theory, MA 485/585. But now they can be presented in a somewhat more formal manner.

Let $X \colon \Omega \to \mathbb{R}$ be a discrete random variable with values

$$x_1, x_2, \dots$$

and respective probabilities

$$p(x_1), p(x_2), \dots$$

The **mean value** (or **expected value**) of X is defined by

$$\mathbb{E}(X) = x_1 p(x_1) + x_2 p(x_2) + \cdots = \sum x_i p(x_i).$$

A more formal version of this formula is

$$\mathbb{E}(X) = \sum_{x \in X(\Omega)} x\, \mathbb{P}\bigl(X^{-1}(\{x\})\bigr)$$

A simple example: the indicator function I_A (Examples 14.3 and 15.3 in the book). It is a random variable, and we have $\mathbb{E}(I_A) = \mathbb{P}(A)$.

If the r.v. X takes infinitely many values, the computation of $\mathbb{E}(X)$ brings us to an infinite series. That infinite series has a well-defined sum if and only if it converges *absolutely*, i.e., if the sum

$$|x_1| p(x_1) + |x_2| p(x_2) + \cdots = \sum |x_i| p(x_i)$$

is finite. Thus the definition of the mean value $\mathbb{E}(X)$ <u>requires</u> the absolute convergence of the above series. If it does not converge absolutely, the mean value <u>does not exist</u>.

Example: suppose X takes values $0, -1, 2, -3, 4, \dots$ (i.e., $x_n = (-1)^n n$ for $n \geq 0$) with probabilities $p(x_n) = \frac{1}{(n+1)(n+2)}$, just like in Practice problem 13.7 in the book. Then the series

$$\sum_{n=0}^{\infty} x_n p(x_n) = 0 \cdot \frac{1}{1 \cdot 2} - 1 \cdot \frac{1}{2 \cdot 3} + 2 \cdot \frac{1}{3 \cdot 4} - \cdots$$

converges (its sum is approximately -0.0795). But does it give us the mean value of X? No, because the series of absolute values diverges:

$$\sum_{n=0}^{\infty} |x_n| p(x_n) = \sum_{n=0}^{\infty} \frac{n}{(n+1)(n+2)} = \infty$$

(this is an analogue of the classical harmonic series, $\sum_{n=1}^{\infty} \frac{1}{n} = \infty$). The random variable X in this example does NOT have a mean value.

A function $Y = g(X)$ of a random variable X should be defined as the composition of two functions: $X\colon \Omega \to \mathbb{R}$ and $g\colon \mathbb{R} \to \mathbb{R}$. Draw a diagram.

Now the formula for the mean value of Y should make better sense: if we denote $y_n = g(x_n)$ for $n \geq 1$ then

$$\mathbb{E}(Y) = \sum_{y \in Y(\Omega)} y\, \mathbb{P}\big(Y^{-1}(\{y\})\big) = \sum_{x \in X(\Omega)} g(x)\, \mathbb{P}\big(X^{-1}(\{x\})\big)$$

as it should be clear from the diagram that $y = g(x)$ and $X^{-1}(\{x\}) = Y^{-1}(\{y\})$.

Suggested homework problems: 16.6 and 16.11.

17 Variance and Standard Deviation

Most of the content of this section has been covered in Probability Theory, MA 485/585. So it only need be quickly reviewed in class.

Remind the students of the "official" definition of the variance

$$\mathsf{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}(X))^2\big]$$

and the "shortcut" formula

$$\mathsf{Var}(X) = \mathbb{E}(X^2) - \big[\mathbb{E}(X)\big]^2$$

Suggested homework problem: 17.2.

18 Binomial and Multinomial Random Variables

Binomial r.v.'s have been extensively covered in Probability Theory, MA 485/585, so they only need be quickly reviewed in class.

Multinomial random variables have not been mentioned in MA 485/585, so they should be introduced here. They constitute a small part, though.

Suggested homework problems: 18.6 and 18.19.

19 Poisson Random Variable

Poisson r.v.'s have been extensively covered in Probability Theory, MA 485/585, so they only need be quickly reviewed in class.

Theorem 19.1 (Poisson approximation to binomials) should be stated more formally:

Theorem. Let $\lambda > 0$ be a positive real number and $k \geq 0$ a non-negative integer. Let X_i, $i \geq 1$, be a sequence of binomial random variables with parameters n_i and p_i such that

$$\lim_{i\to\infty} n_i = \infty, \quad \lim_{i\to\infty} p_i = 0, \quad \lim_{i\to\infty} n_i p_i = \lambda,$$

Then the binomial probabilities

$$\mathbb{P}(X_i = k) = \binom{n_i}{k} p_i^k (1 - p_i)^{n_i - k}$$

converge to the Poisson probability

$$\lim_{i\to\infty} \mathbb{P}(X_i = k) = \frac{\lambda^k}{k!}\, e^{-\lambda}$$

Suggested homework problems: 19.9 and 19.14.

20 Other Discrete Random Variables

This is a long section consisting of three parts. The first is devoted to geometric random variables fully covered in MA 485/585. Just remind the students of the main formulas, including the mean and the variance.

The second covers Negative Binomials. Give the definition (both versions, with x and y on the first page of section 20.2). Give formulas for the mean and variance. Relate them to the corresponding formulas for the geometric r.v. and describe how they can be easily derived.

Indeed, the negative binomial r.v. is just the sum of r independent copies of the corresponding geometric r.v. (with the same probability od success), so we just need to multiply the mean and variance of the latter by r. This is based on the rules

$$\mathbb{E}(X_1 + \cdots + X_r) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_r)$$

and

$$\mathsf{Var}(X_1 + \cdots + X_r) = \mathsf{Var}(X_1) + \cdots + \mathsf{Var}(X_r)$$

that the students have learned in MA 485/585 (the second rule requires independence of $X_1, \dots, X_r$).

The third (and last) part, 20.3, is devoted to Hypergeometric Random Variable. Give the definition. The following "contingency table" helps to clarify the procedure:

Type	Drawn	Not drawn	Total
A	k	$n-k$	n
B	$r-k$	$N-n-(r-k)$	$N-n$
Total	r	$N-r$	N

Give the main formulas, including those for the mean and variance. Relate them to those of the binomial r.v., binomial(r,p), with r trials and probability of success $p = \frac{n}{N}$. The relation goes as follows.

We can represent the hypergeometric r.v. X as $X = X_1 + \cdots + X_r$, where X_i takes two values: 1 (if the object drawn at step i is of type A) and 0 (otherwise, i.e., if the object is of type B). Then

$$\mathbb{E}(X) = \mathbb{E}(X_1 + \cdots + X_r) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_r)$$

It is easy to convince ourselves that each X_i is just a Bernoulli r.v. with $\mathbb{P}(X_i) = p = \frac{n}{N}$. Thus the mean value of the sum is $rp = \frac{rn}{N}$, i.e., the hypergeometric r.v. and the binomial(r,p) have exactly the same mean value.

For the variance, the formula given right before Example 20.12 is too complicated and should be avoided. The one given in the part (c) of the solution is easier to use and interpret; it can be written as follows:

$$\mathsf{Var}(X) = r \cdot \frac{n}{N} \cdot \frac{N-n}{N} \cdot \frac{N-r}{N-1}$$

If we ignore the last factor, $\frac{N-r}{N-1}$, we would just get the variance of the binomial(r,p), i.e., $rp(1-p)$. So where does that last factor come from?

It is because $X_1, \dots, X_n$ are *not* independent (unlike in the binomial model), so the formula for the variance must include covariance terms:

$$\begin{aligned}\mathsf{Var}(X_1 + \cdots + X_r) &= \mathsf{Var}(X_1) + \cdots + \mathsf{Var}(X_r)\\ &\quad + 2\mathsf{Cov}(X_1, X_2) + \cdots + 2\mathsf{Cov}(X_{r-1}, X_r)\end{aligned}$$

Since the objects are drawn from a finite population, drawing an object of type A at the ith step reduces the likelihood that an object of type A will be drawn at other steps. So the variables $X_1, \dots, X_n$ are dependent; more precisely they are *negatively* correlated, and this negative correlation reduces the total variance. The reduction is represented by the factor $\frac{N-r}{N-1}$, which is less than one. Note that if $r = 1$, i.e., if only *one* object is drawn, then there are no correlations, and respectively that extra factor is $\frac{N-1}{N-1} = 1$.

More precisely, the contribution of the covariance terms can be computed as follows. First, one can easily convince oneself that the covariance $\mathsf{Cov}(X_i, X_j)$ is the same for each pair i, j. Then compute the covariance $\mathsf{Cov}(X_1, X_2)$ (this should be a routine exercise), then multiply that covariance by the doubled number of pairs i, j, i.e., by $r(r-1)$. One will get exactly the above formula for the variance $\mathsf{Var}(X)$. There is not need to bring the details in class, just outline the approach.

Last note: there is a clear connection of Negative Binomials with geometric r.v.'s and Hypergeometric with binomials. Thus perhaps their names can be regarded as misleading and should be switched...:-)

Suggested homework problems: 20.11, 20.17, 20.26, 20.34, and 8.11 (bonus)

21 Properties of the Cumulative Distribution Function

Most of the content of this section has been covered in Probability Theory, MA 485/585. So it needs to be just reviewed in class, with emphasis on the basic properties and formulas:

Properties: The distribution function $F(t)$ of any random variable has three basic properties:

- $0 \le F(t) \le 1$
- $F(t)$ is increasing (not strictly): $F(t_1) \le F(t_2)$ for any $t_1 < t_2$
- $\lim_{t\to\infty} F(t) = 1$ and $\lim_{t\to-\infty} F(t) = 0$
- $F(t)$ is right-continuous: $F(t) = \lim_{n\to\infty} F(t + \frac{1}{n})$

The right continuity is based on the *continuity of probabilities* (Section 7), explain.

In fact, the above three properties are *characteristic* properties, i.e., any function $F(t)$ with these properties is a distribution function for some random variable. This gives an idea of what distribution functions are.

Distribution functions need not be continuous, we have seen examples of discontinuous distribution functions in Section 14.

The left-sided limit of a distribution function has the following meaning:

$$\lim_{n\to\infty} F(t - \tfrac{1}{n}) = \mathbb{P}(X < t)$$

This relation is also based on the *continuity of probabilities* (Section 7), explain. We will briefly denote the left-sided limit by

$$F(t-) = \lim_{n\to\infty} F(t - \tfrac{1}{n})$$

Basic formulas relating the distribution function to probabilities:

$$\begin{aligned}
\mathbb{P}(X \le t) &= F(t) \\
\mathbb{P}(X < t) &= F(t-) \\
\mathbb{P}(X \ge t) &= 1 - F(t-) \\
\mathbb{P}(X > t) &= 1 - F(t) \\
\mathbb{P}(t < X < s) &= F(s-) - F(t) \\
\mathbb{P}(t < X \le s) &= F(s) - F(t) \\
\mathbb{P}(t \le X < s) &= F(s-) - F(t-) \\
\mathbb{P}(t \le X \le s) &= F(s) - F(t-) \\
\mathbb{P}(X = t) &= F(t) - F(t-)
\end{aligned}$$

Do Example 21.5 in class.

An important fact: $F(t)$ is discontinuous at $t \in \mathbb{R}$ if and only if $\mathbb{P}(X = t) > 0$, which is clear from the last formula above. Such real numbers (points) $t \in \mathbb{R}$ are called *atoms* of the random variable X (they carry a positive probability).

If X is discrete, then all its values are atoms, and the entire probability distribution is concentrated on atoms. This is one “extreme”.

On the other hand, continuous random variables are characterized by the fact that $\mathbb{P}(X = t) = 0$ for *all* $t \in \mathbb{R}$, i.e., they have no atoms. Equivalently, their distribution function is continuous everywhere. This is the other “extreme”.

In the studies of Probability Theory, we usually discuss discrete random variables and continuous random variables separately, as two main classes, we do not “mix” them together. In actuarial practice, however, one naturally encounters “mixed” random variables, which have some atoms and some continuous components.

For example, suppose the damage to a house is modeled by an exponential random variable X with parameter $\lambda > 0$, i.e., $\mathbb{P}(X \le t) = 1 - e^{-\lambda t}$ for $t > 0$. This is a continuous random variable. Now suppose the insurance policy has a franchise clause (not to be confused with deductible) of \$500, i.e., the claims below \$500 are not honored, but the claims above \$500 are paid in full. Then the amount of payment becomes discontinuous – all the claims below \$500 are rejected, so they “lump” into a zero payment: $t = 0$ becomes an atom. Payments above \$500 remain modeled by the continuous exponential distribution. Draw the respective graph of the distribution function.

A similar situation (a mixed distribution) arises when the insurance policy has a maximum amount of payment.

Example 21.5 is a mixed random variable.

Suggested homework problems: 21.3, 21.5, 21.9

22 Continuous random variables

Most of the content of this section has been covered in Probability Theory, MA 485/585, but now it should be presented more formally.

A random variable X is said to be continuous if $\mathbb{P}(X = t) = 0$ for any $t \in \mathbb{R}$. Equivalently, its distribution function is continuous. This is the official definition of continuous random variables.

In most cases, continuous random variables have a *probability density function*, i.e. a function $f(x)$ such that

$$\mathbb{P}(X \in B) = \int_B f(x)\,dx$$

for any set $B \subset \mathbb{R}$ of real numbers (see the book).

However, there are continuous random variables that *do not* have a density function. Therefore it is not exactly correct to define continuous random variables in terms of a density function (as one may not exist).

Such sloppy definitions are commonly given in books (and in the MA 485/585 Probability course), though. The reason is that continuous random variables that do not have a density function are merely a mathematically exotic phenomenon, they are never encountered in practice. So for all practical purposes we can just forget about them and think that all continuous random variables have a density function.

The density function $f(x)$ has the following two properties:

$$f(x) \geq 0 \qquad \text{for all} \;\; x \in \mathbb{R}$$

and

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

The second property allows us to determine an unknown constant in the formula for $f(x)$. For example, in homework problem 22.11 one needs to determine k before computing the desired probability.

The relation between the distribution function $F(x)$ and the density function $f(x)$ is as follows:

$$f(x) = F'(x)$$

so $F(x)$ is an antiderivative of $f(x)$. However not every antiderivative of $f(x)$ is $F(x)$. One has to use the following rule:

$$F(t) = \int_{-\infty}^{t} f(x)\, dx$$

It should be also emphasized that $f(x)$ need not be specified at every single point $x \in \mathbb{R}$. For example, the uniform random variable X on the interval $(0,1)$ has the following 'rectangular' density function:

$$f(x) = \begin{cases} 1 & \text{for} \;\; 0 < x < 1 \\ 0 & \text{for} \;\; x < 0 \;\; \text{or} \;\; x > 1 \end{cases}$$

This relation specifies $f(x)$ everywhere except $x = 0$ and $x = 1$. The values $f(0)$ and $f(1)$ do not matter, they can be set to 0 or to 1 (or to any real number, for that matter). It is also common to ignore points where $f(x)$ happens to be discontinuous and specify $f(x)$ only on its continuity intervals.

Last minor note: Remark 22.2 in the book is incorrect. There are random variables for which the density function $f(x)$ *does not* converge to zero as $x \to \infty$ or $x \to -\infty$.

Suggested homework problem: 22.11

23 Expectation, Variance and Standard Deviation

Most of this was covered in MA 485/585.

It is worth reviewing useful formulas in Theorems 23.1 and 23.2.

Stress that for mixed random variables one needs to combine summation and integration for computing mean (and variance).

Do Examples 23.4 and 23.6 (without routine details).

Example 23.10 actually introduces Pareto distribution, it is worth discussing in detail (without computing mean and variance).

Remind the students of percentiles and the median.

Do Example 23.13.

Suggested homework problems: 23.11, 23.12, 23.13 (for extra credit). Note: in Problem 23.13 the random variable is mixed!

24 The Uniform Distribution Function

Nothing new, just skip it.

25 Normal Random Variables

The integration of the density (involving polar coordinates) is worth doing in class.

Say that μ is a *location parameter* and σ is a *scale parameter* for the family of normal distributions. Explain the meaning, with graphs. Define the Z-score: $Z = \frac{X-\mu}{\sigma}$ (see the proof of Theorem 25.2). Describe its practical importance as a measure of standing (in a population or a sample).

Emphasize the importance of percentiles in practical applications. Do Example 25.3, part (c). Note: in Probability, percentiles are denoted by π_p, which means $\mathbb{P}(X \leq \pi_p) = p$, so the subscript p corresponds to the 'left' (lower) part of the distribution. In statistics, percentiles for the standard normal distribution are denoted by z_α, which means $\mathbb{P}(X > z_\alpha) = \alpha$, so the subscript α corresponds to the 'right' (upper) part of the distribution. In other words, $z_\alpha = \pi_{1-\alpha}$. Draw a picture.

Why is such a different in notation? Because in statistics α is usually small, so the right tail represents *rare, unusual* values of the random variable, while the bulk of the distribution (below z_α) represents its *typical, common* values. According to a general philosophy of statistics, the most interesting conclusions are made when the random variable takes its unusual values. Those are most important, statisticians focus on tails of the distribution, rather than its bulk, so the notation z_α is convenient since α represents the size of the *tail*.

Discuss the diagram in Example 25.4. Add the interval $\mu \pm 3\sigma$ to it. The corresponding probability is 99.7%.

Normal approximation to binomials: explain that Theorem 25.3 is not enough to approximate individual binomial probabilities, one also needs its "local version" (no need to state it, though).

Suggested homework problems: 25.8, 25.13, 25.17. (Give a hint for 25.8)

26 Exponential Random Variables

Most of it was covered in MA 485/585, except the uniqueness property (Theorem 26.1). This should be presented in class (perhaps the use of logarithms will make the proof easier to follow).

If Poisson process has not been discussed in MA 485/585 (it is the last topic of the course, often left out), then it is a good time to cover it now.

Suggested homework problems: 26.11, 26.14.

27 Gamma and Beta Distributions

This was never mentioned in MA 485/585. Needs to be covered fully.

Introduce the Gamma distribution via the exponential distribution: $X = \text{Gamma}(\lambda, n)$ is the sum of n independent random variables, $X = X_1 + \cdots + X_n$, where each X_i is exponential(λ). Gamma(λ, α) is just a generalization of this to non-integral values of the second parameter. Note: $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Motivate the Beta distribution – it describes proportions (example: proportion of students getting a passing grade in a calculus class). Explain the role of a and b: the fraction $\frac{a}{a+b}$ represents the average proportion, and the magnitude of a and b determines the spread (the higher a and b, the narrower the spread). Qualitatively interpret the formulas in Theorem 27.3: if we set $p = \frac{a}{a+b}$ and $q = 1 - p$, then $\mathbb{E}(X) = p$ and $\mathsf{Var}(X) = \frac{pq}{a+b+1}$. Thus when $a, b \to 0$, the distribution converges to the "extreme" case of a Bernoulli random variable which only takes two values, 0 and 1. When $a, b \to \infty$, then it converges to the other "extreme": $X = p$ is a constant (no variation). Note a symmetry of the beta function: $B(a, b) = B(b, a)$.

Suggested homework problems: 27.4 (note: waiting time to catch a fish is an exponential r.v.), 27.7 (bonus), 27.9 (bonus), 27.16(a), 27.19 (bonus).

28 Distribution of a Function of a Random Variable

Nothing new, just skip it.

29 Joint Distributions

Mostly a repetition of MA 485/585, but the joint distribution function F_{XY} should be discussed a bit more thoroughly. Derive a formula for the probability of a rectangle (page 291), similar to the inclusion-exclusion formula.

An optional theoretical topic: Remind of the distinction between continuous and absolutely continuous random variables. Note that it is harder to make such a distinction for *pairs* of random variables. In particular, it is no longer true that the following two properties are equivalent:

- For any $(x, y) \in \mathbb{R}^2$ we have $\mathbb{P}(X = x, Y = y) = 0$
- $F_{XY}(x, y)$ is a continuous function of two variables

Give an example, such as $X \equiv \text{const}$, Y is uniform$(0, 1)$.

For the reason above, there is only one definition of "jointly continuous" random variables, and it corresponds to *absolute continuity*, i.e., the existence of a density function.

Define *marginal distributions*, with formulas for p.m.f. and p.d.f. of marginal distributions.

Suggested homework problems: 29.9, 29.12, 29.14. Give hints for 29.9 and 29.12 (draw the domain of integration).

30 Independent Random Variables

Mostly covered in MA 485/585. Remind of the formulas characterizing the independence, in terms of p.m.f. and p.d.f. Discuss the criteria for independence (Theorem 30.2), do Examples 30.2 and 30.3 (the latter can be done fast without any calculations).

Suggested homework problems: 30.11, 30.13, 30.16 (give hints for all)

31 Sums of Two Independent Random Variables

This was not covered in MA 485/585. Derive the convolution formula first in the discrete case (easy), then in the continuous case (Theorem 31.1). No need to prove, just make analogy with the discrete case summation with integration.

Review Examples 31.2 and 31.7. Interpret the result of Example 31.2 (the sum of two independent Poissons is a Poisson). Emphasize the importance of determining the limits of integration in Example 31.7. Note and explain general facts: the sum of several independent geometric random variables is a negative binomial (this helps with homework problem 31.5); the sum of several independent exponential random variables is a gamma, and the sum of two gammas is a gamma (Example 31.9).

Suggested homework problems: 31.5, 31.18, 31.21

32 Conditional Distributions: Discrete Case

This was not covered in MA 485/585. Cover thoroughly. Motivate this section as follows: when two random variables X and Y are *not* independent, they affect each other. But how exactly? The answer is the conditional distribution. In practical terms, the known value of one random variable can be used to predict the other, or compute the average value of the other.

On the other hand, in many practical experiments one knows X and then the distribution of Y *depending on* X, so one knows p_X and $p_{X|Y}$. Then one can find the joint p.m.f. by the formula $p_{XY}(x,y) = p_Y(y) \cdot p_{X|Y}(x|y)$. Example from MA 485: roll a die, read its value Y and then toss a coin Y times, then X is the number of heads observed.

Do Example 32.2, also compute the conditional expectation $\mathbb{E}(X|Y = 2)$, because it is involved in Practice Problem 32.11.

Suggested homework problems: 32.3 and 32.11. Give a hint on 32.11 (conditional distribution is simply geometric).

33 Conditional Distributions: Continuous Case

This was not covered in MA 485/585. Cover thoroughly. Introduce continuous formulas by analogy with discrete ones (summation is replaced with integrations). Check that $\int f_{X|Y}\, dx = 1$. Do Example 33.2 (note that the joint density is unbounded! and interesting fact by itself). Relate the conditional C.D.F. (after Example 33.2) to probabilities.

Suggested homework problems: 33.9, 33.10 and 33.15. Give hints.

34 Joint Probability Distributions of Functions of Random Variables

This was not covered in MA 485/585, but its importance is quite limited. There are no actuarial-related practice problems. Suggestion: quickly go over the theoretical formulas with the Jacobian of the transformation (of which the students may have heard before). The most interesting example is the Box-Muller transformation (Problem 34.8), which can be done quickly on the board. (It was a graduate homework exercise in MA 585.)

No need to assign homework problems.

35 Properties of Expectation

Nothing really new, except the formula for the expectation of a function of two random variables (quickly do Example 35.1). Right after that give the formula for the expectation of a product of two independent random variables (Proposition 35.5).

A very nice Example 25.2 is worth discussing (relate it to the Matching Problem done in Section 7).

Sample mean is worth mentioning, explain its significance in statistical terms.

Suggested homework problems: 35.6, 35.9, 35.15. Give a hint for 35.15.

36 Covariance, Variance of Sums, and Correlation

Almost nothing new. Review the formula for the variance of the sum of n random variables. Present Theorem 36.2 (with a short proof), and as a corollary – a description of the two extreme cases $\rho_{XY} = \pm 1$.

Emphasize that two random variables may be (i) independent, (ii) uncorrelated, and (iii) correlated. The “uncorrelated” situation is intermediate, it can be regarded as “hidden dependence” (not visible in many practical cases). Illustrate by graphs (Figure 36.1).

Suggested homework problems: 36.11, 36.16, 36.27.

37 Conditional Expectation

It was actually introduced in Section 32, so just review it here. Give the "double expectation property" (Theorem 37.1). Relate it to the law of total probability.

Discuss the prediction problem. Start with a single random variable X and define the *best predictor* (or *estimator*) c to be the number that minimizes the mean squared error $\mathbb{E}(X-c)^2$. Then $c = \mathbb{E}(X)$ is the best predictor. Now in the case of two random variable X and Y suppose the value of X is observed and then Y needs to be predicted. Then the best predictor is $\mathbb{E}(Y|X)$, which is exactly the conclusion of Theorem 37.2 (state it without proof).

Suggested homework problem: 37.19 (for graduate students). Give a hint.

38 Moment Generating Function

Almost nothing new, just review quickly. Discuss the existence/convergence issue. Do the Practice Problem 38.6 in class (the M.G.F. of a Cauchy random variable does not exist).

Note: for exponential and Gamma random variables, $M_X(t)$ is defined for $t < \lambda$. Review the formula for the M.G.F. of a sum of independent random variables, compare it to the convolution formula...

Do Practice Problem 38.8 in class, a very nice exercise.

Suggested homework problems: 38.11 (give a hint, relate it to Gamma) and 38.28 (extra credit for graduate students)

39 The Law of Large Numbers

Almost nothing new in theory. Discuss (to some extent) the difference between Weak and Strong versions of the L.L.N. Mention "Amazing shrinking sliding rectangles"?

Suggested homework problem: 39.4

40 The Central Limit Theorem

If time permits, give a derivation using Moment Generating Functions and Taylor expansion (with the purpose of "clearing the mystery of normal distributions"). This goes as follows.

The Central Limit Theorem 40.2 states that

$$\frac{\sqrt{n}}{\sigma}\Big(\frac{X_1+\cdots+X_n}{n}-\mu\Big)$$

is approximately a standard normal random variable. Here $X_1,\dots,X_n$ are i.i.d. random variables with mean $\mathbb{E}(X_i)=\mu$ and variance $\mathsf{Var}(X_i)=\sigma^2$. The above formula can be written is a better way:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma}$$

Let us denote $Y_i=\frac{X_i-\mu}{\sigma}$. Then the above formula becomes

$$\frac{Y_1+\cdots+Y_n}{\sqrt{n}}$$

The new variables $Y_1,\dots,Y_n$ are also i.i.d. random variables, and they have simpler mean and variance:

$$\mathbb{E}(Y_i)=\frac{\mu-\mu}{\sigma}=0,\qquad \mathsf{Var}(Y_i)=\frac{\mathsf{Var}(X_i)}{\sigma^2}=1$$

Transforming X_i to Y_i is called *centering* and *norming* of the given random variables. The moment generating function of Y_i has the following properties:

$$M_{Y_i}(0)=1,\qquad M'_{Y_i}(0)=0,\qquad M''_{Y_i}(0)=1$$

We approximate this function by its Taylor polynomial of the second degree:

$$M_{Y_i}(t)\approx 1+0\cdot t+\tfrac{1}{2}\,t^2=1+\tfrac{1}{2}\,t^2$$

Now due to independence of $Y_1,\dots,Y_i$ we have

$$M_{Y_1+\cdots+Y_n}(t)=[M_{Y_i}(t)]^n\approx[1+\tfrac{1}{2}\,t^2]^n$$

Lastly we need to divide $Y_1+\cdots+Y_n$ by $\sqrt{n}$, which corresponds to dividing t by $\sqrt{n}$ in the formula for the moment generating function (see formulas after Example 38.5, and also in MA 485/585). This gives

$$M_{\frac{Y_1+\cdots+Y_n}{\sqrt{n}}}(t)\approx\Big[1+\tfrac{t^2}{2n}\Big]^n$$

Now taking the limit as $n \to \infty$ (a calculus exercise) gives

$$\lim_{n\to\infty}\left[1+\frac{t^2}{2n}\right]^n = e^{\frac{t^2}{2}}$$

which is exactly the moment generating function of the standard normal random variable!

Suggested homework problems: 40.12, 40.14, 40.16

41 Markov Chains

This part of the course is not mentioned in MA 485/585, it should be covered in full and at slow pace.

41.1 Introduction

Markov chains are very important in theoretical sciences (math physics, dynamical systems, etc.) and practical applications (marketing, banking and finance). Markov chains are required by Actuarial Exam MLC.

In basic Probability Theory, we mostly deal with independent events and independent random variables. While these cover the majority of practical applications, there are some where dependence between events and random variables is substantial. Markov chains are sequences of events and/or random variables that are *dependent*.

To introduce Markov chains, give the first example from the introductory book, pages 1–6 (Lower/Middle/Upper classes). By this example introduce basic elements of Markov chains: states, transition matrix, probability vector, multiplication formulas, stationary vector (equilibrium), regular matrices, representation by graphs, etc.

41.2 Two practical tasks

Explain how to do two practical tasks: (i) determine whether a given matrix is regular (refer to Note on page 7, which needs to be corrected as follows: "if two different powers, $\mathbf{P}^i$ and $\mathbf{P}^j$, of the transition matrix $\mathbf{P}$ have zeroes in the same positions, then the chain is not regular") and (ii) find a stationary vector. For this purpose, recommended exercises are 20, 21, 23, 25, 27, 28 from the introductory book.

41.3 Marketing model

Suppose there are four leading brand names of a certain product (such as toothpaste). Each has a certain market share. Customers tend to switch from one brand name to another when they buy the product next time. The probabilities of switching from brand i to brand j can be determined by market analysis, and they remain fairly stable in time. So the dynamics of the market can be described by a Markov chain. See the illustration.

This model allows forecasting future market distributions. It also explains the convergence back to equilibrium after occasional fluctuations (caused by sales at discount prices, aggressive advertisement, etc.)

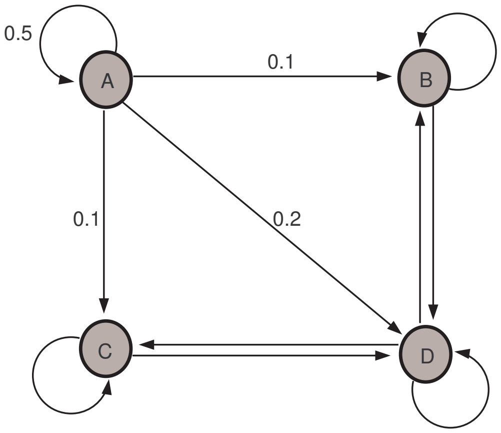

Figure 2: Marketing model. The states A, B, C, D represent brand names.

41.4 Banking model

Another interesting example comes from banking practice. A bank has current customers and former customers that are divided into categories depending on how long ago they left, i.e., according to their "recency". For each category there is a certain probability of coming back, otherwise they slide into the next category. See the illustration.

Describe the corresponding structure of the transition matrix (only two non-zero entries in each row). Indicate why it is regular despite having so many zeros.

41.5 Random walk

One can recall Random Walk covered in MA 485/585 as yet another example of a Markov chain. It has infinitely many states, though. But the restricted random walk (with upper and lower limit values) has finitely many states, so it is a perfect example. Its interesting feature is that it is not regular (unlike the previous examples that were all regular).

41.6 General notation

The formal notation are different in the introductory and advanced books. We need to fix some for the use in class. The following notation can be used:

- States are labeled by $1, \ldots, r$ (so r denotes the number of states)
- Transition probability from state i to state j is denoted by p_{ij}

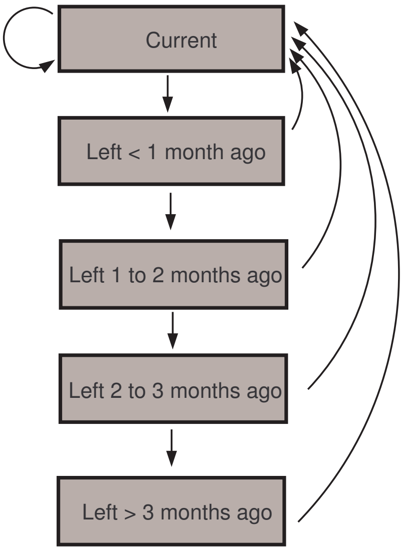

Figure 3: Banking model. The states represent current and recent customers.

- Transition matrix is $\mathbf{P} = (p_{ij})$; it is a square matrix of size $r \times r$
- Probability vectors are denoted by $\mathbf{v}$, their components by v_i
- Initial distribution is described by a probability vector $\mathbf{v}^{(0)}$
- Distribution at time n is described by a probability vector $\mathbf{v}^{(n)}$
- Stationary distribution is described by a probability vector $\mathbf{w}$

It is also convenient to introduce "unity" vector $\mathbf{u}$ with components $(1, 1, \ldots, 1)$. The matrix $\mathbf{P}$ is stochastic, which means $\mathbf{P}\mathbf{u} = \mathbf{u}$. Thus $\mathbf{u}$ is an eigenvector corresponding to the eigenvalue $\lambda = 1$. All the powers of $\mathbf{P}$ are also stochastic matrices because the relation $\mathbf{P}\mathbf{u} = \mathbf{u}$ easily implies $\mathbf{P}^n\mathbf{u} = \mathbf{u}$ for all $n \geq 1$. This can be used to show that $\mathbf{P}$ has no eigenvalues larger than one. Thus $\lambda = 1$ is the largest (maximal) eigenvalue, i.e., for all other eigenvalues λ we have $|\lambda| \leq 1$.

41.7 Transition formula and stationary vector

The basic transition formula is $\mathbf{v}^{(n+1)} = \mathbf{v}^{(n)}\mathbf{P}$. This easily implies $\mathbf{v}^{(n)} = \mathbf{v}^{(0)}\mathbf{P}^n$, hence $\mathbf{P}^n$ consists of n-step transition probabilities (from time 0 to time n). The stationary distribution $\mathbf{w}$ satisfies $\mathbf{w} = \mathbf{w}\mathbf{P}$. Taking transpose gives $\mathbf{w} = \mathbf{P}^T\mathbf{w}$ where $\mathbf{P}^T$ is the transpose of $\mathbf{P}$ (for vectors, we can adopt a "sloppy" convention: they are either row vectors or column vectors depending on their place in a particular formula).

The relation $\mathbf{w} = \mathbf{P}^T\mathbf{w}$ shows that $\mathbf{w}$ is an eigenvector of $\mathbf{P}^T$ corresponding to the eigenvalue $\lambda = 1$. Recall that a matrix and its transpose always have the same eigenvalues (but different eigenvectors). Thus the matrices $\mathbf{P}$ and $\mathbf{P}^T$ have the same eigenvalues but different eigenvectors. Since $\mathbf{P}$ has an eigenvalue $\lambda = 1$ (with the eigenvector $\mathbf{u}$), then so does $\mathbf{P}^T$, therefore there is always a stationary vector $\mathbf{w}$.

41.8 Simplex and a bit of topology

Another way to show the existence of a stationary vector is mathematically more sophisticated. The set of all probability vectors is a simplex (draw images for $r = 2$ and $r = 3$). A simplex is a compact connected set (explain). Now the formula $\mathbf{v} \mapsto \mathbf{v}\mathbf{P}$ defines a transformation of that simplex into itself. And a general theorem by Brouwer asserts that any continuous map of a compact convex set into itself has a fixed point. In our case this means that there is a vector $\mathbf{w}$ such that $\mathbf{w} = \mathbf{w}\mathbf{P}$.

41.9 (Non)uniqueness of stationary vector

It may happen that there is more than one stationary vector, i.e., there are $\mathbf{w}_1 = \mathbf{w}_1\mathbf{P}$ and $\mathbf{w}_2 = \mathbf{w}_2\mathbf{P}$, where $\mathbf{w}_1 \neq \mathbf{w}_2$ are some distinct vectors. In that case for any two constants c_1, c_2 we have $(c_1\mathbf{w}_1 + c_2\mathbf{w}_2) = (c_1\mathbf{w}_1 + c_2\mathbf{w}_2)\mathbf{P}$, so the whole 2D plane spanned by the vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ consists of stationary vectors. That plane intersects the simplex in a line (illustrate) that stretches across the simplex and hits its border. The border is made of vectors whose one component is zero (illustrate). Thus we can make a conclusion: if there is more than one stationary vector, then one of them has a zero component.

This allows us to verify that for regular Markov chains the stationary vector is unique (and all its components are positive). Indeed, suppose one of the components of $\mathbf{w}$ is zero, i.e., $w_i = 0$ for some i. Due to the stationarity we have $w_i = \sum_j p_{ji}^{(k)} w_j$ for every $k \geq 1$, where $p_{ji}^{(k)}$ denote the components of $\mathbf{P}^k$. For some k all $p_{ji}^{(k)} > 0$ are positive numbers, hence $w_i = 0$ can only happen if $w_j = 0$ for all j, which is impossible.

41.10 Convergence to stationary vector

A more elaborate argument shows that for any initial vector $\mathbf{v}^{(0)}$ the sequence $\mathbf{v}^{(n)} = \mathbf{v}^{(0)}\mathbf{P}^n$ converges to $\mathbf{w}$, as $n \to \infty$. It goes as follows. First, according to linear algebra

$$\mathbf{v}^{(0)} = (v_1^{(0)}, \dots, v_r^{(0)}) = v_1^{(0)}\mathbf{e}_1 + \dots v_r^{(0)}\mathbf{e}_r$$

where $\mathbf{e}_1, \dots, \mathbf{e}_r$ denote canonical basis vectors. Hence

$$\mathbf{v}^{(0)}\mathbf{P}^n = v_1^{(0)}\mathbf{e}_1\mathbf{P}^n + \dots v_r^{(0)}\mathbf{e}_r\mathbf{P}^n$$

Thus it is enough to show that $\mathbf{e}_i\mathbf{P}^n$ converges to $\mathbf{w}$, as $n \to \infty$, for every $i = 1, \dots, r$. The vector $\mathbf{e}_i\mathbf{P}^n$ is actually the ith row of the matrix $\mathbf{P}^n$. Thus our goal is to show that all the rows of $\mathbf{P}^n$ converge to the *same* row vector, as $n \to \infty$. This means that in every column of $\mathbf{P}^n$ we have almost equal numbers, and in the limit $n \to \infty$ every column will consist of the one number repeated r times. The ith column of $\mathbf{P}^n$ is actually $\mathbf{P}^n\mathbf{e}_i$, so we need to show that the components of the vector $\mathbf{P}^n\mathbf{e}_i$ get closer to each other, as $n \to \infty$.

Let us fix $i = 1, \dots, r$ and see how the vector $\mathbf{P}^n\mathbf{e}_i$ changes as n grows. Let M_n denote the largest component of $\mathbf{P}^n\mathbf{e}_i$ and m_n the smallest one. Our goal is to show that $M_n - m_n$ converges to zero, as $n \to \infty$. Let $k \geq 1$ be such that $\mathbf{P}^k$ consists of positive numbers, and $d > 0$ denote the smallest of those numbers. From the relation $\mathbf{P}^{n+k}\mathbf{e}_i = \mathbf{P}^k\mathbf{P}^n\mathbf{e}_i$ we can deduce that

$$M_{n+k} \leq dm_n + (1-d)M_n$$

and

$$m_{n+k} \geq (1-d)m_n + dM_n$$

Subtracting the second inequality from the first gives

$$M_{n+k} - m_{n+k} \leq (1-2d)(M_n - m_n)$$

Thus the difference $M_n - m_n$ decreases at least by a factor $1 - 2d < 1$ after every k steps. Hence it converges to zero, as desired.

41.11 Irregular chains: Erenfest model and maze

Our next goal is to discuss Markov chains that are not regular. Give examples of periodic chains. The simplest example is a cycle. A more interesting example is the Erenfest model (Example 11.8 on page 410 in the advanced book). Give the transition matrix for $N = 4$ balls. Draw the graph. Let the students guess why this chain is periodic. Use color chalk to mark the periodic classes of states.

An even more interesting example is a maze (Example 11.22 on page 440–441 in the advanced book). Give the transition matrix (perhaps only partially). Again, let the students guess why it is not regular (periodic).

41.12 Irreducible (ergodic) chains

Introduce the notion of irreducible (ergodic) Markov chain. Explain why the previous examples were irreducible (though not regular). Stress that regular chains are irreducible, but not vice versa.

For irreducible chains, the stationary state $\mathbf{w}$ is still unique, just as for regular chains. The argument is almost the same as it is for regular chains, except we have to find $w_j > 0$ first and then find $k \geq 1$ such that $p_{ji}^{(k)} > 0$.

However it is no longer true that for any initial vector $\mathbf{v}^{(0)}$ the sequence $\mathbf{v}^{(n)} = \mathbf{v}^{(0)}\mathbf{P}^n$ would converge to the stationary vector $\mathbf{w}$. Give examples for cycles and other periodic chains (the Erenfest model, the maze).

It is interesting to find the stationary vector for several examples: cycles, the Erenfest model, and the maze (see the corresponding examples in the advanced book). In all of these examples the stationary vector can be guessed intuitively and the reasons can be clearly explained.

It helps to visualize the evolution of probabilities as follows. Start with the maze example. One can think of a **population** of rats (rather than one wandering rat) which move around the maze chaotically, each rat following the rules of the maze. Then the fraction of that population in each room represents the probability for a single rat to be in that room. The movement of the whole population can be controlled and understood easier than that of a single rat. This picture can be used to explain the description of the stationary vector and the periodic character of the evolution of probabilities.

In more general Markov chains, one can think that water flows in pipes between reservoirs (the states are reservoirs and the arrows representing transitions between states are pipes). Then one can explain how the water flows between the states and eventually its amount (level) stabilizes in each reservoir for regular chains (or up to a period, in irreducible chains).

41.13 Cesaro convergence

As we said, it is no longer true that for any initial vector $\mathbf{v}^{(0)}$ the sequence $\mathbf{v}^{(n)} = \mathbf{v}^{(0)}\mathbf{P}^n$ would converge to the stationary vector $\mathbf{w}$. On the other hand, Cesaro averages of the vectors $\mathbf{v}^{(0)}, \dots, \mathbf{v}^{(n)}$ will converge to the stationary vector $\mathbf{w}$:

$$\frac{1}{n+1}\big(\mathbf{v}^{(0)} + \cdots + \mathbf{v}^{(n)}\big) \to \mathbf{w}$$

as $n \to \infty$. Explain this fact in practical terms (counting average number of "rats" in each room of the maze over an interval of time).

The above Cesaro convergence can be easily seen as follows. Replacing $\mathbf{v}^{(n)}$ with $\mathbf{v}^{(0)}\mathbf{P}^n$ gives

$$\frac{\mathbf{v}^{(0)}}{n+1}\big(\mathbf{I} + \mathbf{P} + \cdots + \mathbf{P}^n\big) \to \mathbf{w}$$

where $\mathbf{I}$ is the identity matrix. Now let us multiply both sides by $\mathbf{I} - \mathbf{P}$ on the right. After cancelations we get

$$\frac{\mathbf{v}^{(0)}}{n+1}\big(\mathbf{I} - \mathbf{P}^{n+1}\big) \to \mathbf{wI} - \mathbf{wP} = \mathbf{w} - \mathbf{w} = 0$$

But indeed, the left hand side converges to zero because the denominator grows (and $\mathbf{P}^{n+1}$ does not: remember that it is a transition matrix, so its components are ≤ 1).

On the subject of irreducible (ergodic) Markov chains, recommended homework exercises from the advanced book are: 3, 4, 6, 9, 25, 26, 27, 28, 31(bonus) from Section 11.3 (pages 442–447). These are appropriate for 600 level students, but some of them may be assigned to 500 level students as well.

41.14 Disconnected chains

Next we move to more general Markov chains that are not even irreducible. Show simple examples with isolated states and/or non-communicating groups of states; see Figure 6 on page 11 in the introductory book. (Another example: a maze with an isolated room or with a wall dividing the maze into disconnected "quarters".) Describe consequences, in particular non-uniqueness of stationary states. Show that there are stationary states with zero components (which is consistent with our previous analysis).

Markov chains with disconnected parts can easily be divided into disconnected groups of states, and then one can describe the "life" in each group separately.

41.15 One-way transitions

A more interesting type of non-irreducible chains are those with one-way transitions; see Figures 4 and 5 on pages 10–11 in the introductory book. Those are connected but not irreducible (i.e., not ergodic). Show examples. Describe ultimate consequences: mass leaks out of some states and they eventually dry out. So every stationary vector has zero components corresponding to those states.

We call states from which one-way transitions exist to other states "transient" or "non-essential" (to justify the latter term say that eventually those states dry out, nothing will be left there eventually).

This takes us to the last big topic in the theory of Markov chains: absorbing states. This topic perhaps requires more technical work from students than other topics.

41.16 Absorbing chains

Give simple examples on pages 10–11 of the introductory book. Also mention restricted random walks from MA 485, illustrated by "Drunkard's walk" on page 416 of the advanced book. Use it as a primary example. Its transition matrix $\mathbf{P}$ is given on page 416; see also below.

A more serious example (Ex. 6 in the introductory book) comes from life sciences, describe it (maybe just briefly).

Define the canonical form of $\mathbf{P}$ – page 13 in the introductory book and page 417 in the advanced book; note that they are arranged differently, choose and fix the latter one for the class use:

$$\mathbf{P} = \begin{bmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

41.17 Analysis of primary example

Give several first powers of $\mathbf{P}$ for the primary example:

$$\mathbf{P} = \left[\begin{array}{ccc|cc} 0 & 0.5 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0.5 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{c|c} \mathbf{Q} & \mathbf{R} \\ \hline \mathbf{0} & \mathbf{I} \end{array}\right]$$

second power:

$$\mathbf{P}^2 = \begin{bmatrix} 0.25 & 0 & 0.25 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.25 & 0.25 \\ 0.25 & 0 & 0.25 & 0 & 0.5 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

third power:

$$\mathbf{P}^3 = \begin{bmatrix} 0 & 0.25 & 0 & 0.625 & 0.125 \\ 0.25 & 0 & 0.25 & 0.25 & 0.25 \\ 0 & 0.25 & 0 & 0.125 & 0.625 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

fourth power:

$$\mathbf{P}^4 = \begin{bmatrix} 0.125 & 0 & 0.125 & 0.625 & 0.125 \\ 0 & 0.25 & 0 & 0.375 & 0.375 \\ 0.125 & 0 & 0.125 & 0.125 & 0.625 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

fifth power:

$$\mathbf{P}^4 = \begin{bmatrix} 0 & 0.125 & 0 & 0.6875 & 0.1875 \\ 0.125 & 0 & 0.125 & 0.375 & 0.375 \\ 0 & 0.125 & 0 & 0.1875 & 0.6875 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Explain in intuitive terms what is going on. Present the expressions for $\mathbf{P}^n$ in the canonical form:

$$\mathbf{P}^n = \begin{bmatrix} \mathbf{Q}^n & (\mathbf{I} + \mathbf{Q} + \cdots + \mathbf{Q}^{n-1})\mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

Next step is to find the limit of $\mathbf{P}^n$ as $n \to \infty$. The following fact takes place:

$$\mathbf{I} + \mathbf{Q} + \cdots + \mathbf{Q}^{n-1} \to (\mathbf{I} - \mathbf{Q})^{-1}$$

as $n \to \infty$. Explain why (multiply by $\mathbf{I} - \mathbf{Q}$). The matrix

$$\mathbf{F} = (\mathbf{I} - \mathbf{Q})^{-1}$$

is called the fundamental matrix. Thus we get

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{bmatrix} \mathbf{0} & \mathbf{FR} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

For the primary example, we have

$$\mathbf{Q} = \begin{bmatrix} 0 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0 \end{bmatrix}$$

then

$$\mathbf{I} - \mathbf{Q} = \begin{bmatrix} 1 & -0.5 & 0 \\ -0.5 & 1 & -0.5 \\ 0 & -0.5 & 1 \end{bmatrix}$$

and

$$\mathbf{F} = (\mathbf{I} - \mathbf{Q})^{-1} = \begin{bmatrix} 1.5 & 1 & 0.5 \\ 1 & 2 & 1 \\ 0.5 & 1 & 1.5 \end{bmatrix}$$

Now we can find the block **FR** in the limit matrix:

$$\mathbf{FR} = \begin{bmatrix} 0.75 & 0.25 \\ 0.5 & 0.5 \\ 0.25 & 0.75 \end{bmatrix}$$

Explain the meaning of the block **FR**: it gives “absorption probabilities”, i.e., probabilities to end up in one of the absorbing states if the system is originally in a non-absorbing state.

Next we compute the sum $\mathbf{I} + \mathbf{P} + \cdots + \mathbf{P}^n$. Its limit satisfies

$$\mathbf{I} + \mathbf{P} + \cdots + \mathbf{P}^n \to \begin{bmatrix} \mathbf{F} & * \\ \mathbf{0} & * \end{bmatrix}$$

as $n \to \infty$. Thus, the fundamental matrix **F** gives us the expected number of visits to each non-absorbing state before absorption occurs. Illustrate by the primary example.

Moreover, the vector **Fu** (where again **u** denotes the column vector all of whose components are 1) gives the average “time to absorption” from every non-absorbing state. For the primary example

$$\mathbf{Fu} = \begin{bmatrix} 1.5 & 1 & 0.5 \\ 1 & 2 & 1 \\ 0.5 & 1 & 1.5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 3 \end{bmatrix}$$

Ask a trick question in class: suppose there is only one absorbing state; what is the block **FR**? Give answer, explain.

On the subject of absorbing Markov chains, recommended homework exercises are: 36 and 37 from the introductory book (page 17) and 9, 10 from the advanced book (page 423). The latter two can be given to 600 level students only.

41.18 No memory property

To finish the subject of Markov chains, emphasize their characteristic property: probabilities of transitions to other states are fully determined by where you are now, i.e., by your current state. They do not depend on your "prehistory", i.e., on the states you have visited before. In other words, "the future only depends on the present and not on the past".

This modeling principle may not be always realistic. For the very first example, with Lower/Middle/Upper classes, the chances to go up or down the "social ladder" may depend on where the family has been for several generations, not just one (traditions, habits, "inertia" may play a role).

On the other hand, as studies show, for marketing models (mentioned earlier) Markov chains are very appropriate – they describe the market evolution quite accurately.

41.19 Chains with memory

If however longer memory needs to be taken into account, a more complex Markov chain can be always constructed accordingly.

For example, suppose a drunkard walks on the line, as in a random walk model, but his probabilities are as follows. After a left step he makes another left step with probability 75% and turns around to make a right step with probability 25%. Similarly, after a right step he makes another right step with probability 75% and turns around to make a left step with probability 25%. (The drunkard has "inertia".)

Thus if the drunkard is in state n, he will move either to $n+1$ or to $n-1$, but the probabilities of these moves depend on the previous move, i.e., on the "immediate history" of the drunkard's walk. Now we can define a more complex Markov chain as follows. The states are not just positions of the drunkard but pairs consisting of his position n and the direction (L or R) of the previous move. So the states are (n, L) and (n, R) for each whole number n. This new chain has twice as many states as the old one.

The transition probabilities for this new chain are, according to our description:

$$(n, L) \overset{0.75}{\to} (n-1, L) \qquad (n, L) \overset{0.25}{\to} (n+1, R)$$

and

$$(n, R) \overset{0.75}{\to} (n+1, R) \qquad (n, R) \overset{0.25}{\to} (n-1, L)$$

This is a more complex Markov chain, but still a Markov chain.

42 Multivariate Normal Distributions

This is another topic where dependence plays a crucial role.

42.1 Motivation

Many variables naturally have normal (or close to normal) distributions. Weight and height of a randomly selected person are two classical examples. Suppose the height X has normal distribution with mean 165 cm and standard deviation 15 cm. Let the weight Y have normal distribution with mean 150 lb and standard deviation 35 lb. It is not hard to see that these two normal random variables are NOT independent – taller people tend to be heavier and shorter people lighter. There is correlation between X and Y (the latter can be measured by correlation coefficient $\rho = \rho_{X,Y}$).

For example, suppose we know that somebody's height is 175 cm (which is above average). What can we say about his weight? Is it still true that his weight has a normal distribution with the same mean 150 lb and the same standard deviation 35 lb. No, his statistical average weight should be over 150 lb. So the mean (and perhaps the standard deviation) have to be recomputed. How? This is the subject of this last part of the course.

42.2 Review of the normal distribution $\mathcal{N}(\mu, \sigma^2)$

Recall that the density of a normal random variable $X = \mathcal{N}(\mu, \sigma^2)$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} = c\, e^{-\frac{1}{2}(\alpha x^2 + \beta x)}$$

where we expanded the square $(x-\mu)^2 = x^2 - 2\mu x + \mu^2$, so that

$$\alpha = \frac{1}{\sigma^2}, \qquad \beta = -\frac{2\mu}{\sigma^2}$$

and we "incorporated" the factor $e^{-\frac{\mu^2}{2\sigma^2}}$ into the coefficient c. Thus the density can be generally described as

$$f(x) = c\, e^{-\frac{1}{2}Q(x)}$$

where $Q(x) = \alpha x^2 + \beta x$ is a quadratic polynomial (without a free term) and c a normalizing coefficient (whose value is fully determined by the requirement $\int f(x)\, dx = 1$). In the polynomial Q, the first coefficient α is the reciprocal of the variance σ^2, i.e., $a = 1/\sigma^2$, and the second coefficient β determines the mean value μ.

42.3 Two independent normals

Now let $X = \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y = \mathcal{N}(\mu_Y, \sigma_Y^2)$ be two normal random variables. We begin with the simplest relation between them: X and Y are independent. Then their joint density is

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y}\, e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}} \\ &= c\, e^{-\frac{1}{2}(Ax^2+By^2+Dx+Ey)} \end{aligned}$$

where again we expand the squares

$$(x-\mu_X)^2 = x^2 - 2\mu_X x + \mu_X^2$$

and

$$(y-\mu_Y)^2 = y^2 - 2\mu_Y y + \mu_Y^2$$

so that

$$A = \frac{1}{\sigma_X^2}, \quad B = \frac{1}{\sigma_Y^2}, \quad D = -\frac{2\mu_X}{\sigma_X^2}, \quad E = -\frac{2\mu_Y}{\sigma_Y^2}$$

and we "incorporated" the factor $e^{-\frac{\mu_X^2}{2\sigma_X^2}-\frac{\mu_Y^2}{2\sigma_Y^2}}$ into the leading coefficient c.

Thus again the joint density can be generally described as

$$f(x,y) = c\, e^{-\frac{1}{2}Q(x,y)}$$

where $Q(x,y) = Ax^2 + By^2 + Dx + Ey$ is a quadratic polynomial (without a free term).

42.4 Two dependent normals

Note that the product xy is missing from our formula for Q. This is because we assumed that X and Y were independent, so that their joint density function is a product of $f_X(x)$ and $f_Y(y)$. When X and Y are dependent, then $Q(x,y) = Ax^2 + By^2 + Cxy + Dx + Ey$, so that

$$f(x,y) = c\, e^{-\frac{1}{2}Q(x,y)} = c\, e^{-\frac{1}{2}[Ax^2+By^2+Cxy+Dx+Ey]}$$

Now, because of the term Cxy, the function $f(x,y)$ cannot be represented as $f_X(x)f_Y(y)$.

42.5 Plots of the joint density

To visualize the density $f(x, y)$ one can draw its level curves $f(x, y) =$const. These are the curves where $Q(x, y) =$const, i.e.,

$$Ax^2 + By^2 + Cxy + Dx + Ey = \text{ const}$$

This is a quadratic equation, whose solution is a quadratic curve – ellipse, hyperbola, or parabola (they are collectively called "conic sections"). For the reason that the function $f(x, y)$ must have a finite integral (because $\int f(x, y)\, dx\, dy = 1$), the above curves can only be ellipses. More precisely, they are concentric ellipses with common directions of axes.

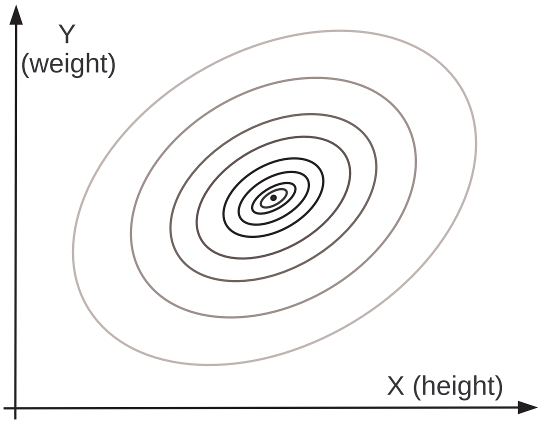

Figure 4: Level curves of $f(x, y)$.

The figure illustrates those ellipses for the variables X (height) and Y (weight). The function $f(x, y)$ takes larger values on smaller ellipses (closer to their common center) and smaller values on larger ellipses (farther from the center). The maximal value of f is taken right at the center. The graph of $f(x, y)$ looks like a "bell" (see page 310 in the electronic notes).

The figure clearly demonstrates dependence (positive correlation) between X and Y: taller people tend to be heavier and shorter people lighter. For this reason the common major axis of our ellipses has positive slope. If X and Y were independent, then we would have $C = 0$ (the term Cxy would be missing) and the ellipses would have horizontal and vertical axes (no slope).

42.6 Conversion to matrices

Let us figure out the meaning of the term Cxy. The quadratic (second order) part of the formula for Q can be presented in matrix form:

$$Ax^2 + By^2 + Cxy = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} A & C/2 \\ C/2 & B \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

The matrix $\begin{bmatrix} A & C/2 \\ C/2 & B \end{bmatrix}$ is related to the so-called **covariance matrix** of the pair X, Y:

$$\mathbf{V} = \begin{bmatrix} \mathsf{Cov}(X,X) & \mathsf{Cov}(X,Y) \\ \mathsf{Cov}(Y,X) & \mathsf{Cov}(Y,Y) \end{bmatrix} = \begin{bmatrix} \mathsf{Var}(X) & \mathsf{Cov}(X,Y) \\ \mathsf{Cov}(X,Y) & \mathsf{Var}(Y) \end{bmatrix}$$

(we used the facts $\mathsf{Cov}(X,X) = \mathsf{Var}(X)$, $\mathsf{Cov}(Y,Y) = \mathsf{Var}(Y)$, and $\mathsf{Cov}(X,Y) = \mathsf{Cov}(Y,X)$). Because the diagonal terms are actually variances of our random variables, this matrix is often called **variance-covariance matrix**. Note that this matrix is symmetric, just like $\begin{bmatrix} A & C/2 \\ C/2 & B \end{bmatrix}$.

Now what is the relation between these two matrices? When X and Y are independent, then, as we have seen before

$$\begin{bmatrix} A & C/2 \\ C/2 & B \end{bmatrix} = \begin{bmatrix} 1/\sigma_X^2 & 0 \\ 0 & 1/\sigma_Y^2 \end{bmatrix}$$

and

$$\mathbf{V} = \begin{bmatrix} \mathsf{Var}(X) & 0 \\ 0 & \mathsf{Var}(Y) \end{bmatrix} = \begin{bmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{bmatrix}$$

This indicates that these matrices are inverse of each other.

In the general case, the inverse of $\mathbf{V}$ is

$$\mathbf{V}^{-1} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1/\sigma_X^2 & -\rho/(\sigma_X\sigma_Y) \\ -\rho/(\sigma_X\sigma_Y) & 1/\sigma_Y^2 \end{bmatrix}$$

This is what $\begin{bmatrix} A & C/2 \\ C/2 & B \end{bmatrix}$ is. So we conclude that

$$A = \frac{1}{(1-\rho^2)\sigma_X^2}, \qquad B = \frac{1}{(1-\rho^2)\sigma_Y^2}, \qquad C = -\frac{2\rho}{(1-\rho^2)\sigma_X\sigma_Y}$$

42.7 Joint density (without matrices)

The overall formula for the joint density looks like this:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}q(x,y)} \tag{1}$$

where

$$q(x,y) = \frac{1}{1-\rho^2}\left[\Big(\frac{x-\mu_X}{\sigma_X}\Big)^2 + \Big(\frac{y-\mu_Y}{\sigma_Y}\Big)^2 - 2\rho\Big(\frac{x-\mu_X}{\sigma_X}\Big)\Big(\frac{y-\mu_Y}{\sigma_Y}\Big)\right] \tag{2}$$

(see also the formulas on top of page 308 in the hand-outs).

42.8 Joint density (with matrices)

Note that the denominator in the formula for $f(x,y)$ can be given as

$$f(x,y) = \frac{1}{2\pi\sqrt{\det \mathbf{V}}}\, e^{-\frac{1}{2}q(x,y)}$$

(the reason for this determinant will be made clear shortly), and the expression for $q(x,y)$ can be given in matrix form:

$$q(x,y) = \begin{bmatrix} x-\mu_X & y-\mu_Y \end{bmatrix} \mathbf{V}^{-1} \begin{bmatrix} x-\mu_X \\ y-\mu_Y \end{bmatrix}$$

To the left and right of $\mathbf{V}^{-1}$ we have the same vector, it is just positioned as a row-vector on the left and as a column-vector on the right (to make the multiplication possible).

We can fully convert the formula to a vector-matrix form if we treat $\mathbf{X} = [x,y]$ and $\boldsymbol{\mu} = [\mu_X, \mu_Y]$ as vectors. Then

$$f(x,y) = \frac{1}{2\pi\sqrt{\det \mathbf{V}}}\, e^{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})\mathbf{V}^{-1}(\mathbf{X}-\boldsymbol{\mu})}$$

42.9 Conditional mean and standard deviation

Let us go back to our motivating question: knowing a person's height, how do we recompute his/her average (expected) weight? Remember that we knew the overall means and standard deviations for the height and weight of a randomly selected person, i.e., we knew μ_X, σ_X and μ_Y, σ_Y. The previous formulas indicate that we need also the correlation coefficient ρ between the height and the weight.

Now what we need is *conditional* mean (and *conditional* standard deviation) for the weight Y, given that the height X takes a specific value, $X = x$. The following formulas give us the conditional mean:

$$\mathbb{E}(Y|x) = \mu_{Y|x} = \mu_Y + \rho \, \frac{\sigma_Y}{\sigma_X} \, (x - \mu_X) \tag{3}$$

and conditional standard deviation:

$$\sigma_{Y|x} = \sigma_Y \sqrt{1 - \rho^2} \tag{4}$$

(see the formulas on page 307 in the electronic notes). Note that the conditional mean $\mathbb{E}(Y|x)$ depends on x (as it should!), but rather strangely the conditional standard deviation $\sigma_{Y|x}$ does not depend on x (still, it *has* to be recomputed, as it is not the same as the given standard deviation σ_Y).

42.10 Practical use

Now we can compute conditional probabilities regarding the unknown weight, given the known height, by standard formulas for normal distributions:

$$\mathbb{P}(a < Y < b | X = x) = \Phi\left(\frac{b - \mu_{Y|x}}{\sigma_{Y|x}}\right) - \Phi\left(\frac{a - \mu_{Y|x}}{\sigma_{Y|x}}\right)$$

The formulas for the conditional mean, standard deviation, and probabilities of X, given a specific value $Y = y$, are the same, except we need to switch X and Y (and use the same ρ).

Note that if X and Y are independent, then $\rho = 0$, and nothing has to be recomputed (the conditional mean and standard deviation are the same as the given ones). Only when there is a correlation $\rho \neq 0$, things have to be adjusted.

Recommended homework exercise is 5.6-1 on page 311 in the electronic notes.

42.11 Representation by standard normals

Next, recall that a general normal random variable $X = \mathcal{N}(\mu, \sigma)$ can be related to a standard normal $Z = \mathcal{N}(0, 1)$ by a formula $X = \mu + \sigma Z$. There is a similar representation for a pair of normals.

Let Z_1, Z_2 be a "standard pair" so that both Z_1 and Z_2 have standard normal distribution $\mathcal{N}(0, 1)$ and they are independent. Let

$$\begin{aligned} X &= \mu_X + \sigma_X Z_1 \\ Y &= \mu_Y + \rho \sigma_Y Z_1 + \sqrt{1 - \rho^2} \sigma_Y Z_2 \end{aligned} \tag{5}$$

Then it is not hard to see, by standard rules of probability, that

$$\begin{aligned}
\mathbb{E}(X) &= \mu_X + \sigma_X \mathbb{E}(Z_1) = \mu_X \\
\mathsf{Var}(X) &= \sigma_X^2 \mathsf{Var}(Z_1) = \sigma_X^2 \\
\mathbb{E}(Y) &= \mu_Y + \rho\sigma_Y \mathbb{E}(Z_1) + \sqrt{1-\rho^2}\sigma_Y \mathbb{E}(Z_2) = \mu_Y \\
\mathsf{Var}(Y) &= \rho^2\sigma_Y^2 \mathsf{Var}(Z_1) + (1-\rho^2)\sigma_Y^2 \mathsf{Var}(Z_2) = \sigma_Y^2 \\
\mathsf{Cov}(X,Y) &= \rho\sigma_X\sigma_Y \mathsf{Var}(Z_1) + \rho\sqrt{1-\rho^2}\sigma_X\sigma_Y \mathsf{Cov}(Z_1,Z_2) = \rho\sigma_X\sigma_Y
\end{aligned}$$

which is exactly what we need.

Now the joint density of Z_1, Z_2 is

$$f(z_1,z_2) = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{z_1^2}{2}} \times \frac{1}{\sqrt{2\pi}}\,e^{-\frac{z_2^2}{2}} = \frac{1}{2\pi}\,e^{-\frac{z_1^2+z_2^2}{2}}$$

To get the joint density of X, Y we can apply the rules of Section 34 of Finan's book. This involve

(a) Changing variables from z_1, z_2 to x, y, which leads to the replacement of the sum $z_1^2 + z_2^2$ in the exponent with $q(x,y)$ given by (2)

(b) dividing by the absolute value of the Jacobian

The Jacobian is

$$\det\begin{bmatrix}\sigma_X & 0 \\ \rho\sigma_Y & \sqrt{1-\rho^2}\sigma_Y\end{bmatrix} = \sigma_X\sigma_Y\sqrt{1-\rho^2}$$

This is why we get the factors $\sigma_X\sigma_Y\sqrt{1-\rho^2}$ in the denominator of (1).

42.12 Conditional mean and standard deviation (revisited)

By the way, the formulas (5) can help us understand (3) and (4). Indeed, if $X = x$ is a specific value of X, then $Z_1 = \frac{x-\mu_X}{\sigma_X}$, and

$$Y = \mu_Y + \rho\,\frac{\sigma_Y}{\sigma_X}\,(x-\mu_X) + \sqrt{1-\rho^2}\sigma_Y Z_2$$

Now (3) and (4) follow by standard rules of probability, because $\mathbb{E}(Z_2) = 0$ and $\mathsf{Var}(Z_2) = 1$.

42.13 Linear transformation for a pair of r.v.'s (without matrices)

Now consider a general linear transformation of two random variables X_1, X_2 into two other random variables Y_1, Y_2 given by

$$\begin{aligned} Y_1 &= aX_1 + bX_2 \\ Y_2 &= cX_1 + dX_2 \end{aligned}$$

Then the new mean values are

$$\begin{aligned} \mu_{Y_1} &= a\mu_{X_1} + b\mu_{X_2} \\ \mu_{Y_2} &= c\mu_{X_1} + d\mu_{X_2} \end{aligned}$$

The new variances and covariance are

$$\begin{aligned} \mathsf{Var}(Y_1) &= a^2\mathsf{Var}(X_1) + b^2\mathsf{Var}(X_2) + 2ab\,\mathsf{Cov}(X_1, X_2) \\ \mathsf{Var}(Y_2) &= c^2\mathsf{Var}(X_1) + d^2\mathsf{Var}(X_2) + 2cd\,\mathsf{Cov}(X_1, X_2) \\ \mathsf{Cov}(Y_1, Y_2) &= ac\,\mathsf{Var}(X_1) + bd\,\mathsf{Var}(X_2) + (ad + bc)\,\mathsf{Cov}(X_1, X_2) \end{aligned}$$

42.14 Linear transformation for a pair of r.v.'s (with matrices)

The above formulas look complicated, but again matrix notation comes to the rescue. Let us express everything in terms of "random vectors"

$$\mathbf{X} = (X_1, X_2) \qquad \text{and} \qquad \mathbf{Y} = (Y_1, Y_2)$$

Their mean values are also vectors

$$\boldsymbol{\mu}_{\mathbf{X}} = (\mu_{X_1}, \mu_{X_2}) \qquad \text{and} \qquad \boldsymbol{\mu}_{\mathbf{Y}} = (\mu_{Y_1}, \mu_{Y_2})$$

and their covariance matrices will be denoted by

$$\mathbf{V}_{\mathbf{X}} = \begin{bmatrix} \mathsf{Var}(X_1) & \mathsf{Cov}(X_1, X_2) \\ \mathsf{Cov}(X_1, X_2) & \mathsf{Var}(X_1) \end{bmatrix}$$

and

$$\mathbf{V}_{\mathbf{Y}} = \begin{bmatrix} \mathsf{Var}(Y_1) & \mathsf{Cov}(Y_1, Y_2) \\ \mathsf{Cov}(Y_1, Y_2) & \mathsf{Var}(Y_1) \end{bmatrix}$$

Now the transformation of $\mathbf{X}$ into $\mathbf{Y}$ can be written in matrix form as

$$\mathbf{Y} = \mathbf{A}\mathbf{X}, \qquad \mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

then their mean values are related by

$$\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{A}\boldsymbol{\mu}_{\mathbf{X}}$$

and their covariance matrices by

$$\mathbf{V}_{\mathbf{Y}} = \mathbf{A}\mathbf{V}_{\mathbf{X}}\mathbf{A}^T$$

(where $\mathbf{A}^T$ denotes the transpose of the matrix $\mathbf{A}$).

42.15 Reduction to an independent pair

If X_1, X_2 are two normal random variables, then Y_1, Y_2 are also two normal random variables, and one can always find a matrix $\mathbf{A}$ such that Y_1, Y_2 will be independent. Indeed, all we need is to make the covariance matrix $\mathbf{V}_{\mathbf{Y}}$ diagonal (to exclude the cross product term from the corresponding quadratic polynomial).

It is known in linear algebra that if $\mathbf{A}$ is a rotation matrix (such matrices are called orthogonal matrices), then $\mathbf{A}^T$ coincides with the inverse $\mathbf{A}^{-1}$, and then the matrix $\mathbf{A}\mathbf{V}_{\mathbf{X}}\mathbf{A}^{-1}$ corresponds to the rotation of the coordinate system.

Again, it is known in linear algebra that every symmetric matrix (and $\mathbf{V}_{\mathbf{X}}$ is symmetric!) can be made a diagonal matrix by a rotation of the coordinate frame. After that $\mathbf{V}_{\mathbf{Y}}$ will be diagonal, so Y_1, Y_2 will be independent.

Geometrically, the rotation transforms the ellipses corresponding to X_1, X_2 into the ellipses corresponding to Y_1, Y_2. After the rotation, the axes of the ellipses will be aligned with the coordinate axes.

42.16 Bonus exercise

A bonus homework exercise (consisting of three parts):

(a) Let X_1, X_2 be two normal random variables with mean $\boldsymbol{\mu}_{\mathbf{X}} = (\mu_{X_1}, \mu_{X_2})$ and covariance matrix $\mathbf{V}_{\mathbf{X}}$. Show that

$$\mathbf{Y} = \mathbf{A}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})$$

gives us two independent standard normal random variables $Y_1, Y_2 = \mathcal{N}(0,1)$ if the matrix $\mathbf{A}$ is symmetric and satisfies $\mathbf{A} = \mathbf{B}^{-1}$ and $\mathbf{B}^2 = \mathbf{V}_{\mathbf{X}}$. (The matrix $\mathbf{B}$ is called the *square root* of $\mathbf{V}_{\mathbf{X}}$.)

(b) Does the matrix $\mathbf{A}$ always exist?

(c) How can you find $\mathbf{A}$ if you are given $\mathbf{V}_{\mathbf{X}}$?

42.17 Three or more normals

Let us extend our motivating example. Suppose that, along with the height and weight of a randomly selected person, we record his/her blood pressure (or some other physical parameter). Now we have three random variables, X, Y, Z, that are all quite dependent on each other. Correlations between them may be positive or negative.

This brings us to a model involving several normal random variables. We denote them by $X_1, X_2, \dots, X_n$. Now the formulas may get very complicated, but vectors and matrices again come to the rescue.

42.18 Joint density for n normals

Let $X_1, X_2, \dots, X_n$ be n normal random variables. Their mean values can be represented as a vector

$$\boldsymbol{\mu} = (\mu_{X_1}, \mu_{X_2}, \dots, \mu_{X_n})$$

Their covariance matrix has size $n \times n$ and can be represented by

$$\mathbf{V} = \begin{bmatrix} \mathsf{Cov}(X_1, X_1) & \cdots & \mathsf{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathsf{Cov}(X_n, X_1) & \cdots & \mathsf{Cov}(X_n, X_n) \end{bmatrix}$$

Now the formula for the joint density of our normal random variables is

$$f(x_1, \dots, x_n) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{V}}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

where $\mathbf{x} = (x_1, x_2, \dots, x_n)$. This formula is almost identical to the one we had for just two normal random variables. Vectors and matrices can handle an arbitrary number of variables easily!

The level surfaces of the joint density function $f(x_1, \dots, x_n)$ are ellipsoids in $\mathbb{R}^n$. They all have a common center and common directions of axes.

42.19 Independence criterion

Normal random variables $X_1, X_2, \dots, X_n$ are independent if and only if all their covariances are zero:

$$\mathsf{Cov}(X_i, X_j) = 0 \qquad \text{for all } i \neq j$$

In this case the covariance matrix $\mathbf{V}$ will be diagonal, so its inverse $\mathbf{V}^{-1}$ will be diagonal, too, and the joint density $f(x_1, \dots, x_n)$ will clearly factor into a product of individual densities. This factorization is a characteristic property of independent random variables.

42.20 Linear transformation for n normals

Suppose n normal random variables $X_1, X_2, \ldots, X_n$ are transformed into other random variables $Y_1, Y_2, \ldots, Y_n$ linearly, i.e., by

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2n}X_n \\ &\cdots \\ Y_n &= a_{n1}X_1 + a_{n2}X_2 + \cdots + a_{nn}X_n \end{aligned}$$

Again we use vector notation for our random variables

$$\mathbf{X} = (X_1, \ldots, X_n) \qquad \text{and} \qquad \mathbf{Y} = (Y_1, \ldots, Y_n)$$

and their mean values

$$\boldsymbol{\mu}_{\mathbf{X}} = (\mu_{X_1}, \ldots, \mu_{X_n}) \qquad \text{and} \qquad \boldsymbol{\mu}_{\mathbf{Y}} = (\mu_{Y_1}, \ldots, \mu_{Y_n})$$

Their covariance matrices will be denoted by

$$\mathbf{V}_{\mathbf{X}} = \begin{bmatrix} \mathsf{Cov}(X_1, X_1) & \cdots & \mathsf{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathsf{Cov}(X_n, X_1) & \cdots & \mathsf{Cov}(X_n, X_n) \end{bmatrix}$$

and

$$\mathbf{V}_{\mathbf{Y}} = \begin{bmatrix} \mathsf{Cov}(Y_1, Y_1) & \cdots & \mathsf{Cov}(Y_1, Y_n) \\ \vdots & \ddots & \vdots \\ \mathsf{Cov}(Y_n, Y_1) & \cdots & \mathsf{Cov}(Y_n, Y_n) \end{bmatrix}$$

Now the transformation of $\mathbf{X}$ into $\mathbf{Y}$ can be written in matrix form as

$$\mathbf{Y} = \mathbf{A}\mathbf{X}, \qquad \mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

Then their mean values are related by

$$\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{A}\boldsymbol{\mu}_{\mathbf{X}}$$

and their covariance matrices by

$$\mathbf{V}_{\mathbf{Y}} = \mathbf{A}\mathbf{V}_{\mathbf{X}}\mathbf{A}^T$$

(where again $\mathbf{A}^T$ denotes the transpose of the matrix $\mathbf{A}$).

42.21 Reduction to an independent set

If $X_1, \dots, X_n$ are n normal random variables, then $Y_1, \dots, Y_n$ are also n normal random variables, and one can always find a matrix $\mathbf{A}$ such that $Y_1, \dots, Y_n$ will be independent. All we need is to make the covariance matrix $\mathbf{V_Y}$ diagonal.

It is known in linear algebra that if $\mathbf{A}$ is an orthogonal matrix, then $\mathbf{A}^T$ coincides with the inverse $\mathbf{A}^{-1}$, and then the matrix $\mathbf{A}\mathbf{V_X}\mathbf{A}^{-1}$ corresponds to the transformation of the coordinate system.

Again, it is known in linear algebra that every symmetric matrix (and $\mathbf{V_X}$ is symmetric!) can be made a diagonal matrix by an orthogonal transformation of the coordinate frame. After that $\mathbf{V_Y}$ will be diagonal, so $Y_1, \dots, Y_n$ will be independent.