Summary of Probability and Statistics, Spring 2015[1]

Sara van de Geer

Chapter 1 & 2

March 5, 2015

[1] Summary of *Wahrscheinlichkeitsrechnung und Statistik*, by H. Föllmer, H.-R. Künsch and additions by J. Teichmann, Feb. 2013

Dictionary for Probability and Statistics, Chapter 1 & 2

English $\to$ *German*

betting strategy: *Spielsystem*

conditional probability: *bedingte Wahrscheinlichkeit*

event: *Ereigniss*

expectation $\mathbb{E}$: *Erwartungswert* $\mathbb{E}$

matching problem: *Garderobenproblem*

outcome ω: *Ergebniss* ω

partition of Ω: *Zerlegung* von Ω

power set: *Potenzmenge*

probability measure $\mathbb{P}$: *Wahrscheinlichkeitsmass* $\mathbb{P}$

probability of success: *Erfolgswahrscheinlichkeit*

probability space $(\Omega, \mathcal{A}, \mathbb{P})$: *Wahrscheinlichkeitsraum* $(\Omega, \mathcal{A}, \mathbb{P})$

random variable X: *Zufallsvariabele* X

random walk: *Irrfahrt*

replacement: *Zurücklegen*

sample space Ω: *Grundraum* Ω

uniform distribution: *Gleichverteilung*

1 Chapter 1: Introduction

The *sample space* is denoted by Ω and subsets A of Ω are called *events*. In Chapter 2 we only consider countable Ω. In Chapter 3 we will introduce a collection $\mathcal{A}$ of "measurable" subsets of Ω. When Ω is countable on can take $\mathcal{A}$ as the collection of **all** subsets, the so-called *power set* of Ω. We need measure theory to deal with uncountable Ω.

A probability measure $\mathbb{P}$ is a mapping

$$\mathbb{P} : \ \mathcal{A} \to [0, 1]$$

which satisfies certain conditions: the axioms of Kolmogorov (see Chapter 3). For $A \in \mathcal{A}$ we say that $\mathbb{P}(A)$ is the probability of the event A.

There are several interpretations of probability. It can express one's belief in a certain event[2]. One can have a frequentist interpretation: the probability of an event is the frequency of occurrences of this event if we repeat the experiment infinitely often. One may want to define the probability of A as the number of outcomes where A occurs divided by the total number of outcomes[3] (this corresponds to the *uniform distribution* on all possible outcomes). One may also want to view probabilities (randomness) as complexity measures.

[2] For example: the probability that a nurse is a murderer is less that .00001 %.

[3] For example: the probability of life on a planet is equal to the number of planets with life divided by the total number of planets.

2 Chapter 2: Discrete probability space

2.1. Basics

Let Ω be countable and $\mathcal{A}$ be the power set of Ω.

Definition Consider a given mapping

$$p: \ \Omega \to [0,1]$$

with $\sum_\omega p(\omega) = 1$. We define

$$\mathbb{P}(A) := \sum_{\omega \in A} p(\omega), \ A \in \mathcal{A}.$$

We call $(\Omega, \mathcal{A}, \mathbb{P})$ a discrete *probability space*.

Two important discrete distributions

Geometric distribution $\Omega := \{1, 2, \ldots\}$, $p(\omega) := (1-p)^{\omega-1}p$ with $0 < p < 1$ a parameter.

Poisson distribution $\Omega = \{0, 1, 2, \ldots\}$, $p(\omega) := \mathrm{e}^{-\lambda}\lambda^k/k!$ with $\lambda > 0$ a parameter. We call this the Poisson(λ)-distribution.

Random variables and expectation

Definition A *random variable* X is a mapping

$$X: \ \Omega \to \mathbb{R}.$$

We write

$$\mathbb{P}(X = x) := \mathbb{P}(\{\omega: \ X(\omega) = x\}).$$

Definition The *expectation* of a random variable X is

$$\mathbb{E}X := \sum_x x\mathbb{P}(X = x).$$

Lemma *Suppose* $X \in \{0, 1, 2, \ldots\}$. *Then*

$$\mathbb{E}X = \sum_{k=0}^{\infty} \mathbb{P}(X > k).$$

Linearity of the expectation Let X and Y be random variables and a and b be constants. Then

$$\mathbb{E}\Big(aX + bY\Big) = a\mathbb{E}X + b\mathbb{E}Y.$$

2.2. Urn models

Consider an urn with k white balls and $N-k$ red balls. Define $p := k/N$. We sample at random n balls from the urn.
1) Sampling with replacement gives a *binomial distribution*:

$$\mathbb{P}(x \text{ white balls}) = \binom{n}{k} p^x (1-p)^{N-x},\ x \in \{0, 1, \dots, n\}.$$

2) Sampling without replacement gives a *hypergeometric distribution*:

$$\mathbb{P}(x \text{ white balls}) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, x \in \{0, 1, \dots, n\} \cap [n+K-N, K].$$

Special case of binomial distribution: $p = 1/2$, $n := 2n$:

$$\mathbb{P}(X = x) = \binom{2n}{x} 2^{-2n},\ x \in \{0, 1, \dots, 2n\}.$$

So

$$\mathbb{P}(X = x) = \binom{2n}{n} 2^{-2n} \sim \frac{1}{\sqrt{n\pi}},$$

where the last result follows from Stirling's formula[4] .

2.3 Random walk

2.3.1. Definition of the random walk

Let $\Omega := \{\omega = (x_1, \dots, x_N) :\ x_i \in \{\pm 1\}\ \forall\ i\}$ and let $\mathbb{P}$ be the *uniform distribution*:

$$\mathbb{P}(A) := \frac{|A|}{|\Omega|},\ A \in \mathcal{A}.$$

Definition 2.1 Consider the random variables $X_i(\omega) :=$ i-th component of $\omega \in \Omega$, $i = 1, \dots, N$. Let $S_0 := 0$ and for $n = 1, \dots, N$, $S_n := \sum_{i=1}^n X_i$. Then $\{S_n\}_{n=0}^N$ is called a *random walk* (starting at zero).

Theorem 2.1 *We have*

$$\mathbb{P}(S_n = 2k - n) = \binom{n}{k} 2^{-n},\ k = 0, 1, \dots, n.$$

Corollary *It holds that*
$\mathbb{P}(S_{2n} = 0) = \binom{2n}{n} 2^{-2n} \sim 1/\sqrt{n\pi}$,
$\mathbb{P}(S_{2n-1} = 1) = \mathbb{P}(S_{2n} = 0)$.

[4]The notation $a \sim b$ means $a/b \to 1$ $(n \to \infty)$.

2.3.2. First visit at level $a \neq o$ and first return to zero

Let $a \in \mathbb{Z}$ and

$$T_a := \min\{n \geq 1 : \ S_n = a\}.$$

If no such n exists we define $T_a := \infty$.

Result
$\mathbb{P}(T_a > n) \to 0$ as $N \geq n \to \infty$,
$\mathbb{E}T_a \to \infty$ as $N \geq n \to \infty$

To prove this result we first prove
$\mathbb{P}(T_a > n) = \mathbb{P}(S_n \in (-a, a])$, $a \neq 0$,
$\mathbb{P}(T_0 > 2n) = \mathbb{P}(S_{2n} = 0)$.

Here in turn, we apply the *reflection principle*.

2.3.3. The arcsin law for the last visit at zero

Let $N := 2N$ and

$$L = \max\{0 \leq n \leq 2N : \ S_n = 0\}.$$

Theorem 2.4 *We have*

$$\mathbb{P}(L = 2n) = \binom{2n}{n}\binom{2(N-n)}{N-n}2^{-2N}, \ n = 0, 1, \ldots, N$$

Approximation For $N \to \infty$ and $n/N \to x \in [0, 1]$

$$\mathbb{P}(L = 2n) \sim \frac{1}{N}\frac{1}{\pi\sqrt{x(1-x)}}.$$

This is called the arcsin law because

$$\int_0^x \frac{1}{\pi\sqrt{u(1-u)}}du = 2\text{arcsin}(\sqrt{x}), \ 0 < x \leq 1.$$

2.3.4. The impossibility of a winning betting strategy

Definition 2.2 An event $A \subset \Omega$ is called *observable* at time n ($0 \leq n \leq N$) if its indicator function 1_A can be written as

$$1_A(\omega) = \phi_n(X_1(\omega), \ldots, X_n(\omega)), \ \forall\, \omega \in \Omega,$$

where $\phi_n : \{\pm 1\}^n \to \{0, 1\}$ is a given function. The collection $\mathcal{A}_n$ is defined as all events A that are observable at time n.

Definition 2.3 The mapping

$$T : \ \Omega \to \{0, 1, \ldots, N\}$$

is called a *stopping time* if $\{T = n\} \in \mathcal{A}_n$, $n \in \{0, \ldots, N\}$.

We now consider random variables $\{V_k\}_{k=1}^N$.

Definition A random variable V_k is called *observable* at time $k-1$ if

$$V_k(\omega) = \phi_{k-1}(X_1(\omega), \ldots, X_{k-1}(\omega)), \ \forall\, \omega \in \Omega,$$

where $\phi_{k-1} : \{\pm 1\}^{k-1} \to \mathbb{R}$ is a given function[5].

Definition A *betting strategy* is $\{(V \cdot S)_n := \sum_{k=1}^n V_k X_k : \ 1 \leq n \leq N\}$.

Impossibility of a winning betting strategy: For any stopping time T

$$\mathbb{E}(V \cdot S)_T = 0.$$

This result can be proved by writing

$$\tilde{V}_k := 1\{T \geq k\} \in \mathcal{A}_{k-1}$$

i.e. $\tilde{V}_k$ it is observable at time $k-1$ $(k = 1, \ldots, N)$.

2.4. Conditional probability

Definition Let $\mathbb{P}(B) > 0$. The *conditional probability* of A *given* B is defined as

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

Definition A *partition* of Ω is a collection of mutually disjoint events $\{B_i\}_{i \in I}$ such that $\cup_{i \in I} B_i = \Omega$.

Theorem 2.7 *(Law of total probability). Let $\{B_i\}_{i \in I}$ be a partition of Ω such that $\mathbb{P}(B_i) > 0$ for all i. Then*

$$\mathbb{P}(A) = \sum_{i \in I} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

Bayes' rule: When both $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$:

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

Corollary

$$\underbrace{\frac{\mathbb{P}(B|A)}{\mathbb{P}(B^c|A)}}_{\textit{posterior odds}} = \underbrace{\frac{\mathbb{P}(A|B)}{\mathbb{P}(A|B^c)}}_{\textit{likelihood ratio}} \times \underbrace{\frac{\mathbb{P}(B)}{\mathbb{P}(B^c)}}_{\textit{prior odds}}.$$

[5]These are not the same functions as used in Definition 2.2.

Theorem 2.9 *Let $\{B_i\}_{i\in I}$ be a partition of Ω such that $\mathbb{P}(B_i) > 0$ for all i. Then for $\mathbb{P}(A) > 0$*

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_{j\in I}\mathbb{P}(A|B_j)\mathbb{P}(B_j)}.$$

2.5. Conditional expectation for discrete random variables

Let X and Y be two discrete random variables. We define the conditional expectation of X given $Y = y$ as[6]

$$\mathbb{E}(X|Y=y) := \sum_x x\mathbb{P}(X=x|Y=y).$$

Note that $\mathbb{E}(X|Y=y)$ is a function of y. Let us write this as

$$\mathbb{E}(X|y) = h(y).$$

The conditional expectation of X given Y is

$$\mathbb{E}(X|Y) := h(Y).$$

Observe that $\mathbb{E}(X|Y)$ is a random variable (in this case a discrete one).

Theorem *(Iterated expectations)*

$$\mathbb{E}\Big(\mathbb{E}(X|Y)\Big) = \mathbb{E}X.$$

Let X be a random variable which we want to predict using the random variable Y by some function of Y, say $g(Y)$. We then call $\mathbb{E}(X-g(Y))^2$ the (squared) *prediction error.*

Theorem 2.10 *The minimizer over all functions $g : \mathbb{R} \to \mathbb{R}$ of $\mathbb{E}(X-g(Y))^2$ is given by $g(Y) = \mathbb{E}(X|Y)$.*

2.6 Independence

Definition 2.6 The events A and B are called *independent* if

$$\mathbb{P}(A\cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

The events $\{A_j\}_{j\in J}$ are called *pairwise independent* if

$$\mathbb{P}(A_i\cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)\ \forall\ i\neq j.$$

[6]We consider only values of y with $\mathbb{P}(Y=y) > 0$.

They are called *independent* if for all $I \subset J$

$$\mathbb{P}(\cap_{i\in I} A_i) = \prod_{i\in I} \mathbb{P}(A_i).$$

The random variables $X_1, \ldots, X_n$ are called *independent* if

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) \ \forall \ (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Note: the events $\{A_i\}$ are independent iff their indicator functions $\{\mathrm{l}_{A_i}\}$ are independent.

Lemma 2.4 *Suppose $X_1, \ldots, X_n$ are independent. Then*

$$\mathbb{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} \mathbb{E} X_i.$$

2.6.2. The binomial distribution

Let $X_1, \ldots, X_n$ be independent with

$$\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = p, i = 1, \ldots, n,$$

where $0 < p < 1$ is a parameter. Define

$$S_n := \sum_{i=1}^{n} X_i.$$

Then

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, 1, \ldots, n.$$

In other words, S_n has the binomial distribution with parameters n and p (Bin(n, p)-distribution).

Approximation of the binomial distribution by the normal distribution

The standard normal distribution We call

$$\phi(x) := \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right], \ x \in \mathbb{R}$$

the *density* of the *standard normal distribution*. We call

$$\Phi(x) := \int_{-\infty}^{x} \phi(u) du, \ u \in \mathbb{R}$$

the *distribution function* of the *standard normal distribution*.

Theorem 2.11 *(de Moivre-Laplace). Let p be fixed and let $A > 0$ be a fixed constant (i.e. both not depending on n). Suppose k grows with n and satisfies $|k - np| \leq A\sqrt{n}$. Then for $n \to \infty$*

$$\mathbb{P}(S_n = k) \sim \frac{1}{\sigma}\phi\Big(\frac{k-\mu}{\sigma}\Big),$$

where $\mu := np$ and $\sigma^2 := np(1-p)$.

2.6.3. The Poisson distribution

Approximation of the binomial distribution by the Poisson distribution

Suppose X has the binomial distribution with parameters n and p where

$$p = \frac{\lambda}{n}$$

for some $\lambda > 0$ not depending on n. Then for $n \to \infty$ and k fixed

$$\mathbb{P}(X = k) \sim \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}.$$

In other words, X is then approximately Poisson distributed.

Some further properties of the Poisson distribution

Theorem 2.13 *Let X_1 and X_2 be independent and suppose that for all $k \in \{0, \ldots, n\}$ and all $n \in \{0, 1, 2, \ldots\}$*

$$\mathbb{P}(X_1 = k | X_1 + X_2 = n) = \binom{n}{k} 2^{-n}$$

(i.e., given the sum $X_1 + X_2 = n$, the random variable X_1 has a $\mathrm{bin}(n, \frac{1}{2})$-distribution). Then there is a $\lambda > 0$ such that both X_1 as well as X_2 have a Poisson distribution with parameter λ.

Theorem 2.14 *Let X_1 and X_2 be independent, and suppose*[7]

$$X_1 \sim^D \text{Poisson}(\lambda_1), \;\; X_2 \sim^D \text{Poisson}(\lambda_2).$$

Then

$$X_1 + X_2 \sim^D \text{Poisson}(\lambda_1 + \lambda_2).$$

[7] The notation $\sim^D$ means “has distribution”