


Essentials of the probability part of MASD

Jonas Peters
University of Copenhagen

November 15, 2018

These “essentials” should not be thought of as lecture notes. Instead, they contain a collection of the most important definitions and results that are presented in the lecture. All explanations, examples, proofs and remarks are missing. In this sense, reading these notes would not compensate for missing some lectures.

This symbol¹  denotes that there are some measure theoretic foundations that we skip. The interested students are welcome to look into such comments, but they are not important for this course, and can thus be ignored.

All information can be found in several textbooks. In particular, I recommend the one accompanying this course: Dimitri P. Bertsekas and John N. Tsitsiklis: “Introduction to Probability”, 2nd Edition, Athena Scientific.

Finally, there might be typos in these notes, so please tell me if you find some.

Jonas Peters

Copenhagen, October 2018

¹The picture is taken from <http://howtobike.info/images/CyclocrossBike.png>, 14.09.2016, 3:41pm UTC+01:00, with the kind permission from Matthew Schoolfield.

Chapter 1

Discrete Random Variables

Random Variables

Definition 1.1 X is called a *random variable* with distribution $\mathbb{P}(X \in \cdot)$ if all of the following hold:

- For any set¹ $A \subseteq \mathbb{R}$, we have $\mathbb{P}(X \in A) \geq 0$.
- $\mathbb{P}(X \in \mathbb{R}) = 1$.
- For any mutually disjoint A_1, A_2, \dots we have

$$\mathbb{P}(A_1 \cup A_2 \cup \dots) = \mathbb{P}(X \in A_1) + \mathbb{P}(X \in A_2) + \dots$$

Notation: $\mathbb{P}(X \in \{3\}) = \mathbb{P}(X = 3)$ mean the same thing. In this course, we denote random variables with capital letters and realizations, i.e., values in \mathbb{R} with small letters: e.g., $\mathbb{P}(X = x)$, denotes the probability that random variable X takes the value $x \in \mathbb{R}$.

Two close friends: pmf and cdf

Definition 1.2 Let X be a random variable. If there are $x_1, x_2, \dots \in \mathbb{R}$ such that


- for all $j \in \{1, 2, \dots\}$, we have $\mathbb{P}(X = x_j) > 0$, and
- $\sum_{j \geq 1} \mathbb{P}(X = x_j) = 1$,

we call X a *discrete random variable* with *support* x_1, x_2, \dots and the function p_X with

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

is called the *probability mass function* (pmf) of X . We say that it specifies the distribution of X .

If it is clear, which variable, we talk about, we sometimes write p instead of p_X .

¹ Strictly speaking, we only need to consider sets in the Borel σ -algebra on \mathbb{R} .

Remark 1.3 If X is a discrete RV with pmf p , then

$$\mathbb{P}(X \in A) = \sum_{x \in A} p(x).$$

Definition 1.4 We call a function $x \mapsto p(x)$ a *valid pmf* if

- for all x , we have $p(x) \geq 0$, and
- $\sum_x p(x) = 1$.

Proposition 1.5 If you draw r balls out of an urn with n balls, there are

	you draw with replacement	you draw without replacement
you care about the order the balls were drawn	n^r	$n(n-1)\cdots(n-r+1)$
you do not care about the order the balls were drawn	$\binom{n+r-1}{r}$	$\binom{n}{r}$

possible drawings. Here,

$$\binom{n}{r} := \frac{n!}{r!(n-r)!}$$

is called the binomial coefficient.

Definition 1.6 Let X be a random variable. Then,

$$F_X(x) := \mathbb{P}(X \leq x)$$

is called the *cumulative distribution function* (cdf) of X . We have

$$\lim_{x \rightarrow -\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x \rightarrow \infty} F_X(x) = 1.$$

Definition 1.7 Some important pmfs have names. They can be found in Table 1.1.

Expectation and Variance

Definition 1.8 Let X be a discrete random variable with pmf p . Then,

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

is called the *expectation* of X if the expression above is finite.

name	X takes values in	pmf	shorthand	interpretation
uniform	$\{1, 2, \dots, m\}$	$p(x) = \frac{1}{m}$	$X \sim \mathcal{U}(\{1, \dots, m\})$	throwing a die
Bernoulli	$\{0, 1\}$	$p(0) = 1 - \theta, p(1) = \theta$	$X \sim \mathcal{Ber}(\theta)$	one trial with success probability θ
binomial	$\{0, 1, 2, \dots, n\}$	$p(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$	$X \sim \mathcal{Bin}(n, \theta)$	# successes of n trials with succ. prob. θ
geometric	$\{1, 2, \dots\}$	$p(x) = \theta(1 - \theta)^{x-1}$	$X \sim \mathcal{Geo}(\theta)$	# trials with succ. prob. θ until 1st succ.
neg. bin.	$\{k, k + 1, \dots\}$	$p(x) = \binom{x-1}{k-1} \theta^k (1 - \theta)^{x-k}$	$X \sim \mathcal{NegBin}(\theta, k)$	# trials with succ. prob. θ until k succ.
Poisson	$\{0, 1, 2, \dots\}$	$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$	$X \sim \mathcal{Poi}(\lambda)$	# of occurrences of (ind.) events at rate λ

57

Table 1.1: Some important pmfs (probability mass functions) with names.

Remark 1.9 It is easy to see that for any discrete random variables X , for all functions² g and for all $\alpha \in \mathbb{R}$

$$\begin{aligned}\mathbb{E}g(X) &= \sum_x g(x)p(x), \\ \mathbb{E}\alpha X &= \alpha\mathbb{E}X,\end{aligned}$$

Definition 1.10 Let X be a discrete random variable with pmf p . Then,

$$\text{var}X := \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \sum_x x^2 p(x) - \left(\sum_x x p(x) \right)^2$$

is called the *variance* of X if the expression above is finite. In that case,

$$\text{sd}(X) := \sqrt{\text{var}(X)}$$

is called the *standard deviation* of X .

Random Vectors and their Joint Distribution

Definition 1.11 We call (X_1, X_2, \dots, X_n) , for some $n \in \mathbb{N}$, a random vector with distribution $\mathbb{P}((X_1, X_2, \dots, X_n) \in \cdot)$ if the following three conditions hold:


- For any set³ $A \subseteq \mathbb{R}^n$, we have $\mathbb{P}((X_1, \dots, X_n) \in A) \geq 0$.
- $\mathbb{P}((X_1, \dots, X_n) \in \mathbb{R}) = 1$.
- For any mutually disjoint A_1, A_2, \dots we have


$$\mathbb{P}(A_1 \cup A_2 \cup \dots) = \mathbb{P}((X_1, \dots, X_n) \in A_1) + \mathbb{P}((X_1, \dots, X_n) \in A_2) + \dots$$

Notation:

$$\begin{aligned}\mathbb{P}((X_1, \dots, X_n) \in \{(x_1, \dots, x_n)\}) &= \mathbb{P}((X_1, \dots, X_n) = (x_1, \dots, x_n)) \\ &= \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)\end{aligned}$$

mean the same thing. Hint: read the comma in the last expression as “and”. The last expression is the most common.

² Strictly speaking, we require g to be measurable in order to guarantee that $g(X)$ is a discrete random variable.

³ Strictly speaking, we only need to consider sets in the Borel σ -algebra on \mathbb{R}^n .

Definition 1.12 Let (X_1, \dots, X_n) be a random vector. If $\exists w_1, w_2, \dots$, elements of \mathbb{R}^n , such that

$$\sum_j \mathbb{P}((X_1, \dots, X_n) = w_j) = 1,$$

we call (X_1, \dots, X_n) a discrete random vector and the function

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) := \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$$

is called the *joint probability mass function* (joint pmf).

Remark 1.13 Let X and Y be discrete random variables with joint pmf $p_{X,Y}$. Then, we can reconstruct the pmf p_X of X :

$$p_X(x) = \sum_y p_{X,Y}(x, y).$$

It does not work the other way around, i.e., knowing p_X and p_Y is not sufficient in order to reconstruct $p_{X,Y}$. When we have a random vector (X, Y) , p_X is sometimes called a *marginal* pmf, as opposed to the joint pmf $p_{X,Y}$.

Definition 1.14 Let (X, Y) be a discrete random vector.

i) For any B with $\mathbb{P}(X \in B) > 0$, we call

$$\mathbb{P}(Y \in A | X \in B) := \frac{\mathbb{P}(Y \in A, X \in B)}{\mathbb{P}(X \in B)}$$

the conditional distribution of Y , given $X \in B$.

ii) If (X, Y) has joint pmf $p_{X,Y}$, we define, for all x with $p_X(x) > 0$,

$$p(y|x) := \mathbb{P}(Y = y | X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

as the conditional pmf of Y given $X = x$.

Definition 1.15 Let X, Y, Z be random variables. They are called *mutually independent* if for all⁴ $A, B, C \subseteq \mathbb{R}$ we have

$$\mathbb{P}(X \in A, Y \in B, Z \in C) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)\mathbb{P}(Z \in C).$$

If (X, Y, Z) is a random vector with joint pmf p , this is equivalent to


$$p(x, y, z) = p(x)p(y)p(z) \quad \text{for all } x, y, z.$$


In the case of two variables this is equivalent to

$$p(y|x) = p(y) \quad \text{for all } x, y \text{ with } p(x) > 0.$$

Although we have written down the two above definitions for three random variables, they generalize to an arbitrary number of random variables.

Remark 1.16 Let X and Y be discrete random variables with joint pmf p . Then, for any⁵

⁴ Strictly speaking, this has to hold for all A, B and C from the Borel σ -algebra on \mathbb{R} .

⁵ Strictly speaking, we have to add “measurable”.

function g we have

$$\begin{aligned}\mathbb{E}g(X, Y, Z) &= \sum_{x, y, z} g(x, y, z)p(x, y, z), \\ \mathbb{E}(X + Y) &= \mathbb{E}X + \mathbb{E}Y, \\ \mathbb{E}(\alpha X) &= \alpha \mathbb{E}X, \\ \text{var}(\alpha X) &= \alpha^2 \text{var}(X).\end{aligned}$$

Definition 1.17 Let X, Y be a random vector. Then,

$$\text{cov}(X, Y) := \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y$$

is called the *covariance* between X and Y if the expression is finite. In this case, the expression

$$\rho(X, Y) := \frac{\text{cov}(X, Y)}{\text{sd}(X)\text{sd}(Y)}$$

is called the *correlation* between X and Y .

Proposition 1.18 Let (X, Y) be a random vector. If $\text{cov}(X, Y)$ exists (i.e., it is finite), we have

i)

$$-1 \leq \rho(X, Y) \leq 1;$$

ii)

$$|\rho(X, Y)| = 1 \quad \text{if and only if} \quad \text{there is } \alpha, \beta \text{ s.t. } Y = \alpha X + \beta$$

Proposition 1.19 Let (X, Y, Z) be a random vector and $\alpha \in \mathbb{R}$. If all terms below are finite, we have

$$i) \text{ cov}(X, Y) = \text{cov}(Y, X)$$

$$ii) \text{ cov}(X, X) = \text{var}(X)$$

$$iii) \text{ cov}(\alpha Z + X, Y) = \alpha \text{cov}(Z, Y) + \text{cov}(X, Y)$$

These properties mean that the covariance is a symmetric bilinear form.

Remark 1.20 i) If X and Y are independent, we have

$$\mathbb{E}XY = \mathbb{E}X\mathbb{E}Y$$

and

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$$

and

$$\text{cov}(X, Y) = \rho(X, Y) = 0.$$

Definition 1.21 Let X_1, X_2, \dots be discrete random variables that are mutually independent and that all have the same pmf. We then call X_1, X_2, \dots *independent and identically distributed* (i.i.d.) and write

$$X_1, X_2, \dots \stackrel{\text{iid}}{\sim} \text{Bin}(n, \theta),$$

for example.

Sampling

Definition 1.22 Given some i.i.d. variables

$$X_1, X_2, \dots, X_n,$$

we sometimes call this an *i.i.d. sample of size n* . In particular, when we consider realizations in a computer (e.g. using the function `np.random.binomial()`). Some authors call the realizations (as opposed to the random variables) the “sample”.