Essentials of the probability and statistics part of MAD

Jonas Peters
University of Copenhagen

November 30, 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 symbol[1]  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, November 2018

[1] The 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 2

Continuous Random Variables

Definition 2.1 Let X be a random variable. If there exists an integrable function $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that[1]

$$\forall A \subseteq \mathbb{R} \qquad \mathbb{P}(X \in A) = \int_A f(x)\, dx,$$

we call X a *continuous random variable* and f its *probability density function (pdf)*.

Remark 2.2 If X is a continuous random variable with pdf f, the cdf F is continuous and satisfies

$$F' = f.$$

Remark 2.3 You can repeat all definitions from Chapter 1 (see MASD) by replacing sums with integrals.

Definition 2.4 Let X be a random variable. We define the *median* m of X to be a value that satisfies

$$\mathbb{P}(X \leq m) \geq \frac{1}{2} \quad \text{and} \quad \mathbb{P}(X \geq m) \geq \frac{1}{2}.$$

It is not necessarily unique. If X is a continuous random variable, we can take any value m with $F(m) = 1/2$. If F is strictly monotonically increasing, m is unique.

Definition 2.5 Some important pdfs have names. They can be found in Table 2.1.

Lemma 2.6 *Let $X \sim \mathcal{N}(\mu, \sigma^2)$, i.e., X has the pdf*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Then,

$$\mathbb{E}X = \mu \quad \textit{and} \quad \operatorname{var}(X) = \sigma^2.$$

[1] Correct would be "for all Borel-measurable sets A"

name	X takes values in	pdf	shorthand
uniform	$[a, b]$	$f(x) = \frac{1}{b-a}$ if $a \leq x \leq b$, zero otherwise	$X \sim \mathcal{U}([a, b])$
Gaussian	$\mathbb{R}$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$	$X \sim \mathcal{N}(\mu, \sigma)$
exponential	$\mathbb{R}_{\geq 0}$	$f(x) = \lambda \exp(-\lambda x)$ if $x \geq 0$, zero otherwise	$X \sim \mathcal{E}xp(\lambda)$
student t	$\mathbb{R}$	please check	$X \sim t_n$
chi squared	please check		
beta	please check		
Pareto	please check		

Table 2.1: Some important pdfs (probability density functions) with names.

Lemma 2.7 *[NOT DISCUSSED IN CLASS] Let* $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ *and* $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ *be independent. Then, for all* $\alpha, \beta \in \mathbb{R}$*,*

$$\begin{aligned} X + Y &\sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \\ \alpha X &\sim \mathcal{N}(\alpha\mu_1, \alpha^2\sigma_1^2) \\ X + \beta &\sim \mathcal{N}(\mu_1 + \beta, \sigma_1^2). \end{aligned}$$

Chapter 3

Statistics

Estimators

Definition 3.1 Let $X_1, \ldots, X_n$ be i.i.d. with distribution P_{θ_0} for some $\theta_0 \in \mathbb{R}$. Then an *estimator* for θ is a function

$$\hat{\theta}_n : \mathbb{R}^n \to \mathbb{R}.$$

Definition 3.2 Let $X_1, \ldots, X_n$ be i.i.d. with distribution P_{θ_0} for some $\theta_0 \in \mathbb{R}$ and let $\hat{\theta}_n$ be an estimator for θ. Then, $\hat{\theta}_n(X_1, \ldots, X_n)$ is a random variable (sometimes, we simply write $\hat{\theta}_n$ instead of $\hat{\theta}_n(X_1, \ldots, X_n)$). We define the bias, variance and mean squared error as follows.

$$\begin{aligned}
\mathrm{BIAS}(\hat{\theta}_n) &:= \mathbb{E}\hat{\theta}_n(X_1, \ldots, X_n) - \theta_0 \\
\mathrm{Var}(\hat{\theta}_n) &:= \mathrm{var}\left(\hat{\theta}_n(X_1, \ldots, X_n)\right) \\
\mathrm{MSE}(\hat{\theta}_n) &:= \mathbb{E}\left(\hat{\theta}_n(X_1, \ldots, X_n) - \theta_0\right)^2
\end{aligned}$$

The smaller, the better!

Definition 3.3 Let $X_1, \ldots, X_n$ be an i.i.d. sequence of continuous random variables with pdf f_θ. Then, the joint pdf

$$f_\theta^{\text{joint}}(x_1, \ldots, x_n) = f_\theta(x_1) \cdot \ldots \cdot f_\theta(x_n)$$

is called the *likelihood*. The estimator

$$\hat{\theta}_n^{\mathrm{ML}}(x_1, \ldots, x_n) \ := \ \underset{\theta}{\operatorname{argmax}} \ f_\theta^{\text{joint}}(x_1, \ldots, x_n) \ = \ \underset{\theta}{\operatorname{argmax}} \ f_\theta(x_1) \cdot \ldots \cdot f_\theta(x_n)$$

is called the *maximum likelihood estimator (MLE)* for θ. This definition works analogously for discrete random variables if you replace the pdfs f_θ with pmfs p_θ.

Convergence of Estimators

Definition 3.4 Let $X_1, X_2, \ldots$ be a sequence of random variables and X another random variable. Let F_n and F denote the cdfs of X_n and X, respectively. We say

- X_n *converges to* X *in probability* and write $X_n \xrightarrow{\mathbb{P}} X$ if for all $\varepsilon > 0$,
$$\mathbb{P}(|X_n - X| > \varepsilon) \to 0 \quad \text{for } n \to \infty.$$
- X_n *converges to* X *in distribution* and write $X_n \xrightarrow{\mathcal{L}} X$ if
$$F_n(x) \to F(x) \quad \text{for } n \to \infty,$$
for all x, at which F is continuous.

Proposition 3.5 *We have*
$$X_n \xrightarrow{\mathbb{P}} X \;\Rightarrow\; X_n \xrightarrow{\mathcal{L}} X.$$
In general, the converse does not hold. For any constant $c \in \mathbb{R}$, however, we have
$$X_n \xrightarrow{\mathcal{L}} c \;\Rightarrow\; X_n \xrightarrow{\mathbb{P}} c.$$

Definition 3.6 Let $\hat{\theta}_n$ be an estimator for θ_0. We call $\hat{\theta}_n$ *consistent* if
$$\hat{\theta}_n(X_1, \ldots, X_n) \xrightarrow{\mathbb{P}} \theta_0.$$

Lemma 3.7 *[Chebyshef Inequality] Let X be a random variable with finite mean and variance. Then*
$$\mathbb{P}\left(|X - \mathbb{E}X| > k\sqrt{\mathrm{var}(X)}\right) \;\leq\; \frac{1}{k^2}.$$
Equivalently, we have
$$\mathbb{P}\left(|X - \mathbb{E}X| > \varepsilon\right) \;\leq\; \frac{\mathrm{var}(X)}{\varepsilon^2}.$$

Theorem 3.8 *[Weak law of large numbers] Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with finite mean and finite variance. The sample mean is a consistent estimator for the true mean, that is*
$$\bar{X} := \bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \;\xrightarrow{\mathbb{P}}\; \mathbb{E}X_1.$$

Theorem 3.9 *[Central Limit Theorem] Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with finite mean and finite variance. Then,*
$$\sqrt{n}\,\frac{\frac{1}{n}\sum_{i=1}^{n} X_i \,-\, \mathbb{E}X_1}{\sqrt{\mathrm{var}(X_1)}} \;\xrightarrow{\mathcal{L}}\; Z,$$
where $Z \sim \mathcal{N}(0,1)$.

Confidence Intervals

Definition 3.10 Let $X_1, \ldots, X_n$ be i.i.d. with distribution that depends on θ. If $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ satisfy

$$\mathbb{P}(a \leq \theta \leq b) \geq 1 - \alpha,$$

the interval $[a, b]$ is called a $(1-\alpha)$-*confidence interval for* θ.

Definition 3.11 Let X be a random variable. The r-*quantile* of X is the number x, such that

$$\begin{aligned} &\mathbb{P}(X < x) \leq r \quad \text{and} \\ &\mathbb{P}(X > x) \leq 1 - r. \end{aligned}$$

This number x is sometimes denoted by ...

in general	$X \sim \mathcal{N}(0,1)$	$X \sim t_n$
q_r	z_r	$t_{n;r}$

.

Hypothesis Testing

Definition 3.12 Let $X_1, \ldots, X_n$ be i.i.d. random variables. Let H_0 be a hypothesis about their distribution and let $0 < \alpha < 1$. A function

$$d : \mathbb{R}^n \to \{H_0, H_1\}$$

is called a statistical test for H_0 if

$$\mathbb{P}_{H_0}(d = H_1) \leq \alpha.$$

There are two errors:

- *type I error*: H_0 is correct but $d = H_1$.
- *type II error*: H_0 is false but $d = H_0$.

That is, a statistical test bounds the probability of making a type I error. The value $\mathbb{P}_{H_0}(d = H_1)$ is called the *size* of the test, and α the *significance level*.

Often, the decision function has the form

$$d(x_1, \ldots, x_n) := \begin{cases} H_0 & \text{if } T(x_1, \ldots, x_n) \notin \mathcal{R} \\ H_1 & \text{if } T(x_1, \ldots, x_n) \in \mathcal{R} \end{cases}$$

for a so-called *test statistic* $T : \mathbb{R}^n \to \mathbb{R}$ and *rejection region* $\mathcal{R} \subseteq \mathbb{R}$.

Remark 3.13 Since statistical tests only bound the type I error, H_0 and H_1 should always be chosen such that the type I error is the “worse error”, that is the error which is more important to avoid.

Remark 3.14 Performing a statistical test contains the following six steps (the example shows a so-called "two-sided one-sample z-test"):

1. Write down a model for the data,
   e.g. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2 = 3)$.
2. Write down the hypotheses,
   e.g. $H_0 : \mu = 2$ and $H_1 : \mu \neq 2$.
3. Write down a test statistic T and its distribution under H_0,
   e.g. $T = \sqrt{n}(\bar{X} - 2)/\sqrt{3}$; under H_0 we have: $T \sim \mathcal{N}(0, 1)$.
4. Write down a significance level,
   e.g. $\alpha = 0.05$.
5. Compute the rejection region,
   e.g. $\mathcal{R} = (-\infty, z_{0.025}] \cup [z_{0.975}, \infty) = (-\infty, -1.96] \cup [1.96, \infty)$.
6. Compute the test statistic from the data and report the test result,
   e.g. $T = 0.42 \notin \mathcal{R}$, i.e. H_0 is not rejected.