Lecture 1: Probability

Zhentao Shi

August 6, 2018

Probability

Human beings are awed by uncertainty in daily life. In the old days, Egyptians consulted oracles, Hebrews inquired prophets, and Chinese counted on diviners to interpret tortoise shell or bone cracks. Even in today's Hong Kong fortunetellers are abundant.

Probability theory is a philosophy about uncertainty. Over centuries, mathematicians strived to contribute to the understanding of randomness. As measure theory matured in the early 20th century, Russian mathematician Andrey Kolmogorov (1903-1987) laid the foundation of modern probability theory in his book published in 1933. The formal mathematical language is a system that allows rigorous explorations that have made fruitful advancements, and is now widely accepted as scientific standard in academic and industrial research.

With the advent of big data, computer scientists have come up with a plethora of new algorithms that are aimed at revealing patterns from seemingly random data. Machine learning and artificial intelligence (AI) become buzz words. They defeat best human Go players, automate manufacturers, power self-driving vehicles, recognize human faces, and recommend online purchases. Behind their industrial success, statistics sheds light on the behavior of these algorithms. While statistical theory is built on modern probability theory, the latter is so far the most promising paradigm to rationalize existing algorithms and engineer new ones.

Economics has been an empirical social science since Adam Smith (1723-1790). Many numerical anecdotes appear in his *Wealth of Nations* published in 1776. Ragnar Frisch

(1895-1973) and Jan Tinbergen (1903-1994), two pioneers econometricians, were awarded in 1969 the first Nobel Prize in economics. Econometrics provides quantitative insights about economic data. It flourishes in real-world management practices, from households and firms up to governance at the global level. Today, the AI revolution is pumping fresh energy into research and exercise of econometric methods, while its very foundation is again modern probability theory.

In this preparatory course, we will have a brief introduction of the axiomatic probability theory along with familiar results covered in undergraduate *probability and statistics*. The level of this lecture note is close to

- Casella and Berger (2002): Statistical Inference (second edition)

Interested readers may want to read this textbook for more examples.

Probability Space

A *sample space* Ω is a collection of all possible outcomes. It is a set of things.

An *event* A is a subset of Ω. It is something of interest on the sample space.

A *σ-field*, denoted by $\mathcal{F}$, is a collection of $(A_i \subseteq \Omega)_{i\in\mathbb{N}}$ events such that

(i) $\emptyset \in \mathcal{F}$;
(ii) if an event $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$;
(iii) if $A_i \in \mathcal{F}$ for $i \in \mathbb{N}$, then $\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{F}$.

It is easy to show that $\Omega \in \mathcal{F}$ and $\bigcap_{i\in\mathbb{N}} A_i \in \mathcal{F}$. The σ-field can be viewed as a well-organized structure built on the ground of the sample space. The pair $(\Omega, \mathcal{F})$ is called a *measure space*.

Let $\mathcal{G} = \{B_1, B_2, \ldots\}$ be an arbitrary collection of sets, not necessarily a σ-field. We say $\mathcal{F}$ is the smallest σ-field generated by $\mathcal{G}$ if $\mathcal{G} \subseteq \mathcal{F}$, and $\mathcal{F} \subseteq \tilde{\mathcal{F}}$ for any $\tilde{\mathcal{F}}$ such that $\mathcal{G} \subseteq \tilde{\mathcal{F}}$. A *Borel σ-field* $\mathcal{R}$ is the smallest σ-field generated by the open sets on the real line $\mathbb{R}$.

A function $\mu : (\Omega, \mathcal{F}) \mapsto [0, \infty]$ is called a *measure* if it satisfies

(i) (positiveness) $\mu(A) \geq 0$ for all $A \in \mathcal{F}$;

(ii) (countable additivity) if $A_i \in \mathcal{F}$, $i \in \mathbb{N}$, are mutually disjoint, then

$$\mu\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mu(A_i).$$

Measure can be understand as weight or length. In particular, we call μ a *probability measure* if $\mu(\Omega) = 1$. A probability measure is often denoted as P. The triple $(\Omega, \mathcal{F}, P)$ is called a *probability space.*

Example: (probability measure) Personal wealth management: asset allocation. A probability function is not necessarily about uncertainty. Two allocation rules: 1. Equal weight. 2. Optimal weight.

So far we have answered the question: “What is a well-defined probability?”, but we have not yet answered “How to assign the probability?”

There are two major schools of thinking on probability assignement. One is *frequentist*, who considers probability as the average chance of occurrence if a large number of experiments are carried out. The other is *Bayesian*, who deems probability as a subjective brief. The principles of these two schools are largely incompatible, while each school has peculiar merit under different context.

Some Illustrative Exercises

The following part gives some naive examples and simple propositions to help familiarize with the concepts.

For a fixed measure space $(\Omega, \mathcal{F})$, we can have different probability measure P on it.

Example Let $\Omega = \{a, b\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ where $\mathcal{P}(\Omega)$ denote the power set of Ω, i.e. the collection of all subsets of Ω. Define P_1, $P_2 : (\Omega, \mathcal{F}) \to [0, \infty)$ by

(i) $P_1(\Omega) = 1,\ P_1(\emptyset) = 0,\ P_1(\{a\}) = \frac{1}{2},\ P_1(\{b\}) = \frac{1}{2}$

(ii) $P_2(\Omega) = 1,\ P_2(\emptyset) = 0,\ P_2(\{a\}) = 1,\ P_2(\{b\}) = 0$

Both P_1 and P_2 are probability measures on $(\Omega, \mathcal{F})$. If we simply take this abstract example as tossing a coin where a and b represent two possible symbols, say a number and a flower. P_1 means we are tossing a fair coin with a number and a flower on each side respectively, and P_2 means we are tossing a coin with a flower on both sides.

A sequence of events $\{A_i\}_{i\in\mathbb{N}}$ is increasing (resp. decreasing) if $A_1 \subseteq A_2 \subseteq \cdots$ (resp. $A_1 \supseteq A_2 \supseteq \cdots$). The following is an easy implication of the *monotonic convergence theorem* in real analysis.

Proposition If $\{A_i\}_{i\in\mathbb{N}}$ is increasing (resp. decreasing), then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i\to\infty} P(A_i)$ (resp. $P\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{i\to\infty} P(A_i)$).

Example Consider the probability space $(\Omega, \mathcal{F}, P)$ where $\Omega = (0,1)$, $\mathcal{F}$ is the σ-field generated by $\mathcal{F}_0 = \{(a,b) : 0 < a < b < 1\}$, and P is a probability measure on $(\Omega, \mathcal{F})$ satisfying $P\{(a,b)\} = b - a$.

Note that $\left\{\frac{1}{2}\right\} \in \mathcal{F}$ since $\left\{\frac{1}{2}\right\} = \bigcap_{n=3}^{\infty}\left(\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n}\right) \in \mathcal{F}$. By the above proposition, we can easily get

$$P\left(\left\{\frac{1}{2}\right\}\right) = P\left(\bigcap_{n=3}^{\infty}\left(\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n}\right)\right) = \lim_{n\to\infty} P\left(\left(\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n}\right)\right) = \lim_{n\to\infty} \frac{2}{n} = 0$$

since $\left\{\left(\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n}\right)\right\}_{n=1}^{\infty}$ is decreasing. This example shows that some sets in the collection of events can be 0-measure.

Random Variable

The terminology *random variable* somewhat belies its formal definition of a deterministic mapping. It is a link between two measure spaces such that any event in the σ-field installed on the range can be traced back to an event in the σ-field installed on the domain.

Formally, a function $X : \Omega \mapsto \mathbb{R}$ is $(\Omega, \mathcal{F}) \setminus (\mathbb{R}, \mathcal{R})$ *measurable* if

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$$

for any $B \in \mathcal{R}$. *Random variable* is an alternative, and somewhat romantic, name for a measurable function. We say a measurable is a *discrete random variable* if the set $\{X(\omega) : \omega \in \Omega\}$ is finite or countable. We say it is a *continuous random variable* if the set $\{X(\omega) : \omega \in \Omega\}$ is uncountable.

A measurable function connects two measurable spaces. No probability is involved in its definition yet. While if a probability measure P is installed on $(\Omega, \mathcal{F})$, the measurable function X will induce a probability measure on $(\mathbb{R}, \mathcal{R})$. It is easy to verify that $P_X : (\mathbb{R}, \mathcal{R}) \mapsto [0, 1]$ is also a probability measure if defined as

$$P_X(B) = P\left(X^{-1}(B)\right)$$

for any $B \in \mathcal{R}$. (If $B_1, B_2 \in \mathcal{R}$ are disjoint, then $X^{-1}(B_1), X^{-1}(B_2) \in \mathcal{F}$ are also disjoint.) This P_X is called the probability measure *induced* by the measurable function X. The induced probability measure P_X is an offspring of the parent probability measure P though the channel of X.

Example: Asset and return.

Distribution Function

We go back to some terms that we have learned in a undergraduate probability course. A *(cumulative) distribution function* $F : \mathbb{R} \mapsto [0, 1]$ is defined as

$$F(x) = P(X \leq x) = P(\{X \leq x\}) = P(\{\omega \in \Omega : X(\omega) \leq x\}).$$

It is often abbreviated as CDF, and it has the following properties.

(i) $\lim_{x\to-\infty} F(x) = 0$,

(ii) $\lim_{x\to\infty} F(x) = 1$,

(iii) non-decreasing,

(iv) right-continuous $\lim_{y\to x^+} F(y) = F(x)$.

For continuous distribution, if there exists a function f such that for all x,

$$F(x) = \int_{-\infty}^{x} f(y)\, dy,$$

then f is called the *probability density function* of X, often abbreviated as PDF. It is easy to show that $f(x) \geq 0$ and $\int_a^b f(x)\, dx = F(b) - F(a)$.

Example We have learned many parametric distributions like the binary distribution, the Poisson distribution, the uniform distribution, the normal distribution, χ^2, t, F and so on. They are parametric distributions, meaning that the CDF or PDF can be completely characterized by a few parameters.

Example `R` has a rich collection of distributions implemented in a unified rule: `d` for density, `p` for probability, `q` for quantile, and `r` for random variable generation. For instance, `dnorm`, `pnorm`, `qnorm`, and `rnorm` are the corresponding functions of the normal distribution, and the parameters μ and σ can be specified in the arguments of the functions.

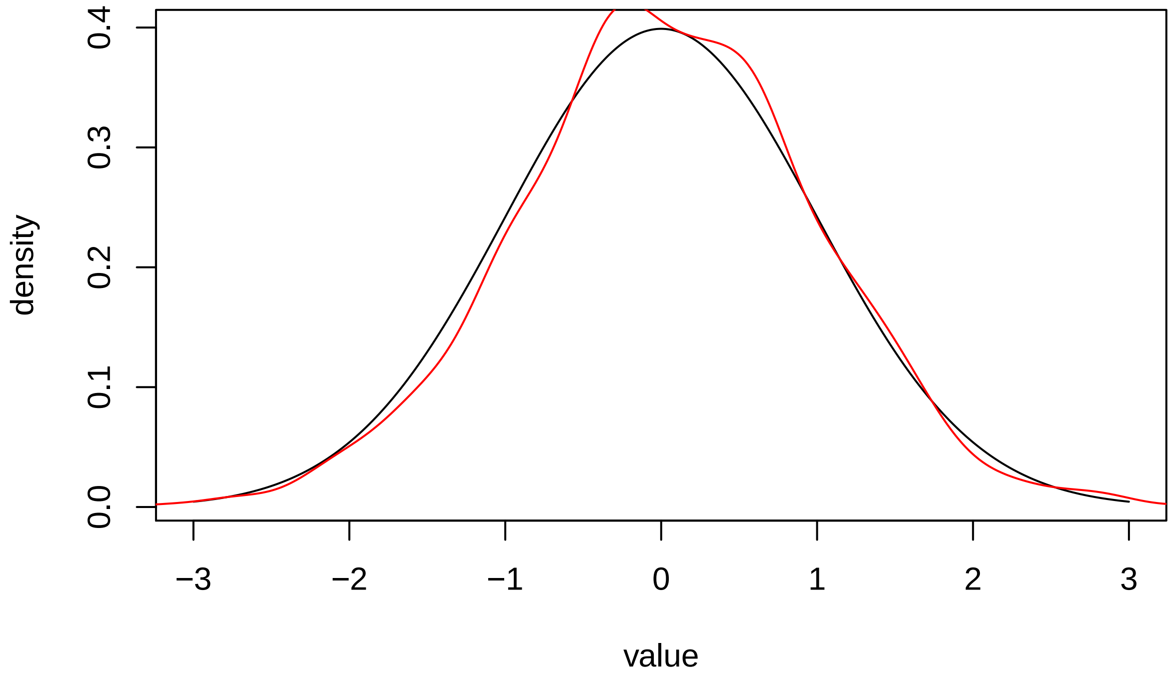
Below is a piece of `R` code for demonstration.

1. Plot the density of standard normal distribution over an equally spaced grid system `x_axis = seq(-3, 3, by = 0.01)` (black line).
2. Generate 1000 observations for $N(0, 1)$. Plot the kernel density, a nonparametric estimation of the density (red line).
3. Calculate the 95th quantile and the empirical probability of observing a value greater than the 95th quantile. In population, this value is 5%. What is the number coming out of this experiment?

(Since we do not fix the random seed in the computer, the outcome is slightly different each time we run the code.)

```
x_axis = seq(-3, 3, by = 0.01)


y = dnorm(x_axis)
plot(y = y, x=x_axis, type = "l", xlab = "value", ylab = "density")
z = rnorm(1000)
lines( density(z), col = "red")
```



```
crit = qnorm(.95)


mean( z > crit )
```

```
## [1] 0.04
```