Revision of Statistics I in English

Barbara Choroś
Institut für Statistik and Ökonometrie
CASE - Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin

Probability theory (Wahrscheinlichkeitstheorie) is concerned with the outcomes of random experiments. These can be either real world processes or experiment. In both cases

1. the experiment has to be infinitely repeatable and
2. there has to be a well-defined set of outcomes.

The set of all possible outcomes of an exeriment is called the sample space which we will denote by S.

Probability (Wahrscheinlichkeit) is a measure $P(\cdot)$ which quantifies the degree of (un)certainty associated with an event.
$P(\cdot)$ is a probability measure. It is a function which assigns a number $P(A)$ to each event A of the sample space S. Axioms:

1. $P(A)$ is real-valued with $P(A) \geq 0$.
2. $P(S) = 1$.
3. If two events A and B are mutually exclusive ($A \cap B = \emptyset$), then

$$P(A \cup B) = P(A) + P(B)$$

Sample space (Ereignisraum) is a set of all possible events of a random experiment. Each event is thus a subset of the sample space. The impossible event is empty set, the sure event is the complete sample space.

Random variable (Zufallsvariable) (rv) is a (real) number which is assigned to every outcome of random experiment.
The outcome of an experiment is not necessarily a number, for example, the outcome when a coin is tossed can be 'heads' or 'tails'. A random variable is a function that associates a unique numerical value with every outcome of an experiment (i.e. value of the corresponding random variable). The value of the random variable will vary from trial to trial as the experiment is repeated.
X: random variable,
x_i: ($i = 1, \ldots, n$) results of n experiments (i.e values of the random variable X).

A random variable is created by assigning a real number to each event E_j (an outcome of an experiment). The event E_j is an element of the set S of all possible outcomes of an experiment. The random variable is then defined by a function that maps the elements of the set S with numbers on the real line.

$$X : E_j \to X(E_j) = x_j$$

There are two types of random variable - discrete and continuous.
A random variable has either an associated probability distribution (discrete random variable) or probability density function (continuous random variable).

A random variable is **one-dimensional** if the experiment only considers one outcome.

A random variable is called **discrete** if the set of all possible outcomes x_1, x_2, $\ldots$ is finite or countable. A **continuous random variable** is one which takes an infinite number of possible values.

Probability distribution of a discrete random variable (Wahrscheinlichkeitsverteilung einer diskreten ZV) is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the probability mass function.
More formally, the probability distribution of a discrete random variable X is a function which gives the probability $p(xi)$ that the random variable equals x_i, for each value x_i:

$$p(x_i) = P(X = x_i)$$

It satisfies the following conditions:

1. $0 \leq p(x_i) \leq 1$,
2. $\sum p(x_i) = 1$.

All random variables (discrete and continuous) have a **cumulative distribution function** (cdf). It is a function giving the probability that the random variable X is less than or equal to x, for every value x.
Formally, the cumulative distribution function $F(x)$ is defined to be:

$$F(x) = P(X \leq x), \qquad -\infty \leq x \leq \infty$$

A probability distribution is called **discrete** if its cumulative distribution function only increases in jumps.

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i)$$

A probability distribution is called **continuous** if its cumulative distribution function is continuous, which means that it belongs to a random variable X for which $P(X = x) = 0$ for all x in $\Re$.

For continuous random variable the cdf is the integral of its probability density function

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

The **probability density function** of a continuous random variable is a function which can be integrated to obtain the probability that the random variable takes a value in a given interval (a, b)

$$\int_{a}^{b} f(t)dt = F(b) - F(a) = P(a < X < b).$$

The probability density function $f(x)$ of a continuous random variable X is the derivative of the cumulative distribution function $F(x)$

$$f(x) = \frac{d}{dx}F(x)$$

A probability density function is any function $f(x)$ that obeys two conditions:

1. the total probability for all possible values of the continuous random variable X is 1
$$\int_{-\infty}^{\infty} f(t)dt = 1,$$
2. $f(x) > 0$ for all x.

Expected Value and Variance (Erwartungswert und Varianz)

1. Let X be the discrete random variable with outcomes x_i and the corresponding probabilities $f(x_i)$

$$\begin{aligned} E(X) &= \sum_i x_i f(x_i) \\ Var(X) &= \sum_i \{x_i - E(X)\}^2 f(x_i) = \sum_i x_i^2 f(x_i) - \{E(X)\}^2. \end{aligned}$$

2. Let X be the continuous random variable with density $f(x)$

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} x f(x)dx, \\ Var(X) &= \int_{-\infty}^{\infty} \{x - E(x)\}^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \{E(X)\}^2. \end{aligned}$$

Standard deviation (Standardabweichung) (std) denotes the square root of the variance, which summarizes the spread of the distribution. Large values of the standard deviation mean that the random variable X is likely to vary in a large neighbourhood around the expected value. Smaller values of the standard deviation indicate that the values of X will be concentrated around the expected value.

Covariance (Kovarianz) is the measure of how much two random variables vary together. If two variables tend to vary together (that is, when one of them is above its expected value, then the other variable tends to be above its expected value too), then the covariance between the two variables will be positive. If one of them is above its expected value and the other variable tends to be below its expected value, then the covariance between the two variables will be negative.
The covariance between two random variables X and Y, with expected values $E(X)$ and $E(Y)$,

$$Cov(X,Y) = E\{(X - EX)(Y - EY)\}.$$

Properities of Expected Value (Eigenschaften des Erwartungswerts)

1. $E(c) = c$, $c = const$,
2. $E(X + c) = E(X) + c$,
3. $E(X + Y) = E(X) + E(Y)$,
4. $E(cX) = cE(X)$,
5. if X and Y are independent, $E(XY) = E(X)E(Y)$.

Properities of Variance

1. $Var(d) = 0$, $d = const$,
2. $Var(dX) = d^2 Var(X)$,
3. $Var(dX + c) = d^2 Var(X)$,
4. $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y)$,
5. if X and Y are uncorrelated, $Var(X + Y) = Var(X) + Var(Y)$.