N.I. Lobachevsky State University
of Nizhni Novgorod

Probability theory and mathematical statistics:

Law of Large Numbers
(Statistical stability revisited)
Random variables

Associate Professor
A.V. Zorine

In the previous lecture we studied Bernoulli trials. For $k < p(n+1)$ the probability of at most k successes

$$Q(k;n,p) = \sum_{j=0}^{k} b(j;n,p)$$

was dominated by

$$b(k;n,p)\frac{(n-k+1)p}{(n+1)p-k}.$$

Recall that the number of occurrences of an event A in n independent trials was denoted by $\mu(A,n)$. A random experiment is said to be *statistically stable* if relative frequency

$$\frac{\mu(A;n)}{n}$$

is oscillating around certain number $\mathrm{P}^*(A)$.

Now we are able to make some estimates.

$$1 > b(0;n,p) + \ldots + \underbrace{b(k;n,p) + b(k+1;n,p) \ldots + b([(n+1)p];n,p)}_{[(n+1)p]+1-k \text{ terms}}$$

Now we are able to make some estimates.

$$1 > \underbrace{b(k;n,p) + b(k+1;n,p) \ldots + b([(n+1)p];n,p)}_{[(n+1)p]+1-k \text{ terms}}$$

Now we are able to make some estimates.

$$1 > \underbrace{b(k;n,p) + b(k+1;n,p) \ldots + b([(n+1)p];n,p)}_{[(n+1)p]+1-k \text{ terms}}$$

$$> \underbrace{b(k;n,p) + b(k;n,p) + \ldots + b(k;n,p)}_{[(n+1)p]+1-k \text{ terms}} > b(k;n,p)((n+1)p-k)$$

Now we are able to make some estimates.

$$1 > \underbrace{b(k;n,p) + b(k+1;n,p) \ldots + b([(n+1)p];n,p)}_{[(n+1)p]+1-k \text{ terms}}$$

$$> \underbrace{b(k;n,p) + b(k;n,p) + \ldots + b(k;n,p)}_{[(n+1)p]+1-k \text{ terms}} > b(k;n,p)((n+1)p-k)$$

Hence

$$b(k;n,p) < \frac{1}{(n+1)p-k},$$

$$Q(k;n,p) < \frac{(n-k+1)p}{((n+1)p-k)^2}$$

Finally, let's estimate the probability that

$$\frac{\mu(A;n)}{n} < p - \varepsilon$$

for some small $\varepsilon > 0$. In this case $\mu(A;n) < n(p-\varepsilon) \approx k$ and

$$Q(k;n,p) < \frac{(n+1-n(p-\varepsilon))p}{((n+1)p-n(p-\varepsilon))^2} = \frac{np(1-p)+(1-\varepsilon)p}{(n\varepsilon-p)^2} \to 0$$

as $n \to \infty$. In the similar way it can be shown that the probability of inequation

$$\frac{\mu(A;n)}{n} > p + \varepsilon$$

tends to zero as $n \to \infty$. In sum, the probability that

$$\left|\frac{\mu(A;n)}{n} - p\right| > \varepsilon$$

can be made arbitrary small for sufficiently large n.

Law of large numbers

As n increases, the probability that the average number $\mu(A;n)/n$ of occurrences of event A deviates from $p = \mathrm{P}(A)$ by more than any preassigned ε tends to zero. In other words,

$$\mathrm{P}\left(\left|\frac{\mu(A;n)}{n} - p\right| < \varepsilon\right) \to 1$$

We see that an empirical fact about statistical stability is embedded into theory. At the same time it would be an error to say that theory of probability *proved* the stability of frequencies.

What is a random variable?

Let's start with a few examples.

Example. Let a person be taken out of a population of n persons. $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ denotes here the set of all persons. Denote by $A(\omega)$ the age of person ω. Thus

$$\omega \to A(\omega)$$

is a function with domain Ω.

Any person has a height, a weight and an income:

$$\omega \to H(\omega), \quad \omega \to W(\omega), \quad \omega \to I(\omega)$$

At census times, agents want to know

$$\omega \to \frac{I(\omega)}{N(\omega)},$$

where $N(\omega)$ is the number of family members. The ration describes "per capita income".

Example. Let Ω denote the set of all molecules in a vessel. (Yes, the number of molecules can be large, say, 10^{25}, but it's still finite.) Then any molecule has its mass m, velocity v, momentum M, kinetic energy E. We have eventually

$$\omega \to m(\omega), \quad \omega \to v(\omega),$$

$$\omega \to M(\omega) = m(\omega)v(\omega), \quad \omega \to E(\omega) = \frac{1}{2}m(\omega)v(\omega)^2.$$

In an experiment with two dice the sample space is

$$\begin{aligned}\Omega = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\ &(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}\end{aligned}$$

The number of points on the first die can be defined as

$$\omega = (x, y) \to X(\omega) = x$$

In the examples above elementary outcomes were assigned with *values* obtained from *measurements*. Suppose for now Ω is a denumerable set.

Definition

A function X of ω:

$$X(\cdot) : \Omega \to \mathrm{R}$$

is called *random variable*.

This variable is a function because it depends on an outcome (maps the set of elementary outcomes into real line). This variable is random because its argument ω is randomly sampled from Ω.

Theorem 1

Let X and Y be be random variables. Then

$$X+Y, \quad X-Y, \quad ,X\cdot Y,$$

and $aX+bY$, where a, b are real numbers, are random variables.

The proof follows from definition of a random variable.

Theorem 2

Let φ be a function of two variables, X and Y be two random variables. Then

$$\omega \to \varphi(X(\omega), Y(\omega))$$

is also a random variable.

Let X and Y be the tangential and normal velocities of a gas molecule, then $\varphi(X,Y)=\sqrt{X^2+Y^2}$ is its absolute velocity.

Let $\varphi(x, y) = x + y$, then Theorem 1 is a corollary of Theorem 2. Another example: let $X_1, X_2, \ldots, X_n$ denote numbers of points in n consecutive rolls of a die, then

$$S_n(\omega) = X_1(\omega) + X_2(\omega) + \ldots + X_n(\omega)$$

gives the total number of points thrown in n rolls.

Let $\varphi(x, y) = x + y$, then Theorem 1 is a corollary of Theorem 2.
Another example: let $X_1, X_2, \ldots, X_n$ denote numbers of points in n consecutive rolls of a die, then

$$S_n(\omega) = X_1(\omega) + X_2(\omega) + \ldots + X_n(\omega)$$

gives the total number of points thrown in n rolls.
Specific events are generated by random variables. For example, if $A(\omega)$ is the age of a person ω, then this age can be between 20 and 40 years,

$$\{20 \leqslant A \leqslant 40\} = \{\omega\colon 20 \leqslant A(\omega) \leqslant 40\},$$

and can be not within the range. This is an example of a random event. Typical events are

$$\{\omega\colon A(\omega) = a\}, \quad \{\omega\colon a < A(\omega) < b\}, \quad \text{etc}$$

Sometimes we think of a random variable *before* we have any formal model of an experiment (such as a triple $(\Omega, \mathfrak{F}, \mathrm{P})$).

Example. The cost of a book copy is \$3 if the print run is below 1000, then \$2 per an additional copy with number from 1001 to 5000 and \$1 for further copies. Assume the initial print run was 1000 copies, each copy at a \$5 price. In this situation the number of books sold is random. Denote it by X.
An income Y can be computed as

$$Y = \begin{cases} 5X - 3000, & \text{if } X \leqslant 1000, \\ 2000 + 3(X - 1000), & \text{if } 1000 < X \leqslant 5000, \\ 14000 + 4(X - 5000), & \text{if } X > 5000. \end{cases}$$

- What is the probability that the book is unprofitable? It is equal to the probability of the event

$$\{5X - 3000 < 0\} = \{X < 600\}.$$

- What is the probability that the book is unprofitable? It is equal to the probability of the event

$$\{5X - 3000 < 0\} = \{X < 600\}.$$

- What is the probability that the profit with be above \$10000? It is equal to the probability of the event

$$\begin{aligned} &\{2000 + 3(X - 1000) \geqslant 10000\} \cup \{X > 5000\} = \\ &= \{X \geqslant \frac{8000}{3} + 1000\} \cup \{X > 50000\} = \\ &= \{X \geqslant 3667\}. \end{aligned}$$

These probabilities depend on behavior of X. We need to know the probabilities of different possible values of X.

What a sample space Ω can be here? Since the main protagonist of this example is random variable X, we can have its possible values $0, 1, \ldots$ as outcomes. Then any outcome ω is a non-negative integer number, and $Y(\omega)$ is a random variable on a set of non-negative integers. What we need here is a *distribution* of random variable X.

What a sample space Ω can be here? Since the main protagonist of this example is random variable X, we can have its possible values $0, 1, \ldots$ as outcomes. Then any outcome ω is a non-negative integer number, and $Y(\omega)$ is a random variable on a set of non-negative integers. What we need here is a *distribution* of random variable X.

Let A be a subset of real numbers (a subset of a real line R). Then we shall write

$$\mathrm{P}(X \in A) = \mathrm{P}(\{\omega \colon X(\omega) \in A\}).$$

E.g., if $A = [a, b]$ (a segment), then

$$\mathrm{P}(X \in A) = \mathrm{P}(a \leqslant X \leqslant b)$$

If Ω is a denumerable set, X can have only denumerable number of different values

$$x_1, x_2, \ldots, x_n, \ldots$$

Put

$$p_n = \mathrm{P}(X = x_n), \quad , n = 1, 2, \ldots$$

If the numbers $p_1, p_2, \ldots, p_n, \ldots$ are known, then it is possible to compute any probability concerning random variable X on its own right. For example,

$$\mathrm{P}(A \leqslant X \leqslant b) = \sum_{a \leqslant x_n \leqslant b} p_n; \quad \mathrm{P}(X \in A) = \sum_{x_n \in A} p_n.$$

We say that pairs $\{(x_1, p_1), (x_2, p_2), \ldots, (x_n, p_n), \ldots\}$ give *the probability distribution* of random variable X, When random variable X has a finite number of possible values x_1, x_2, $\ldots$, x_n it is convenient to place probability distribution in a table:

a	x_1	x_2	$\ldots$	x_n
$\mathrm{P}(X = a)$	p_1	p_2	$\ldots$	p_n

When random variable X has a finite number of possible values x_1, x_2, $\dots$, x_n it is convenient to place probability distribution in a table:

a	x_1	x_2	$\dots$	x_n
$\mathrm{P}(X = a)$	p_1	p_2	$\dots$	p_n

Uniform distribution

a	1	2	$\dots$	n
$\mathrm{P}(X = a)$	$\frac{1}{n}$	$\frac{1}{n}$	$\dots$	$\frac{1}{n}$

$$\mathrm{P}(X = a) = \frac{1}{n}$$

Binomial distribution

a	0	1	$\dots$	n
$\mathrm{P}(X=a)$	$C_n^0 p^0(1-p)^n$	$C_n^1 p^1(1-p)^{n-1}$	$\dots$	$C_n^n p^n(1-p)^0$

,

$$\mathrm{P}(X=a)=C_n^a p^a(1-p)^{n-a},$$

Hyper-geometric distribution

a	0	2	$\dots$	n
$\mathrm{P}(X=a)$	$\frac{C_M^0 C_{N-M}^{n-0}}{C_N^n}$	$\frac{C_M^1 C_{N-M}^{n-1}}{C_N^n}$	$\dots$	$\frac{C_M^n C_{N-M}^0}{C_N^n}$

$$\mathrm{P}(X=a)=\frac{C_M^a C_{N-M}^{n-a}}{C_N^n}$$

Suppose a symmetric coin it tossed until a Heads comes out. Denote by X the number of tosses. Thus $\{X = n\}$ means that there were exactly $n - 1$ Tails and last one was Heads. Since Heads and Tails in different tosses are show up independently

$$p_n = \mathrm{P}(X = n) = \frac{1}{2^n}.$$

If the probability of Heads is p, then

$$p_n = \mathrm{P}(X = n) = (1 - p)^{n-1} p, \quad n = 1, 2, \ldots$$

Random variable X is said to have **geometric distribution**. X is the number of Bernoulli trials until the first success.

Given pairs $\{(x_1, p_1), (x_2, p_2), \ldots, (x_n, p_n), \ldots\}$ such that $p_1 > 0$, $p_2 > 0, \ldots$ and $p_1 + p_2 + \ldots = 1$ we can construct a random variable with exactly this distribution. Let $\Omega = \{x_1, x_2, \ldots\}$, $p(\omega) = p_n$ if $\omega = x_n$, and put $X(\omega) = \omega$.
Then we have a discrete probability sense in Kolmogorov's sense and a proper defined random variable with given distribution.

Consider Bernoulli trials with small probability p of a success. Let this probability $p = p(n)$ change so that $n \cdot p \to \lambda$ as $n \to \infty$ (and this means that $p(n) \to 0$). Then

$$b(0; n, p) = (1 - p)^n \to e^{-\lambda}$$

and

$$R(k; n, p) = \frac{p}{1 - p} \cdot \frac{n - k + 1}{k} \to \frac{\lambda}{k},$$

so true is the following

Limit Theorem of Poisson

Let $pn \to \lambda$ as $n \to \infty$, then

$$b(k; n, p) \to \frac{\lambda^k}{k!} e^{-\lambda}.$$

Proof. We have

$$b(0;n,p) \to e^{-\lambda},$$

$$b(1;n,p) = R(1;n,p)b(0;n,p) \to \lambda e^{-\lambda},$$

$$b(2;n,p) = R(2;n,p)b(1;n,p) \to \frac{\lambda}{2}\lambda e^{-\lambda} = \frac{\lambda^2}{2!}e^{-\lambda},$$

$$b(3;n,p) = R(3;n,p)b(2;n,p) \to \frac{\lambda}{3}\frac{\lambda^2}{2!}e^{-\lambda} = \frac{\lambda^3}{3!}e^{-\lambda},$$

and so on.

It is remarkable that numbers

$$p(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}, \quad k = 0, 1, \ldots \tag{*}$$

sum up to unity:

$$\sum_{n=0}^{\infty} \frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda}\sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda}e^{\lambda} = 1.$$

It means that pairs $\{(0, p(0;\lambda)), (1, p(0;\lambda), \ldots\}$ defined in (*) form a proper probability distribution called the **Poisson distribution**.

Consider a sequence of random events occurring in time, such as *radioactive disintegrations*, or *incoming calls at a telephone exchange*, or *incoming claims to an insurance company*. Each event is represented by a point on the time axis. The simplest physical assumptions lead to $p(n;\lambda)$ as the probability of finding exactly n points (events) within a fixed interval of specified length.

Imagine a unit interval partitioned into n subintervals of length $1/n$. Let each subinterval have the same probability $p(n)$ to contain one or more points. A subinterval is then either occupied or empty, and the assumed independence of non-overlapping time intervals implies that we are dealing with Bernoulli trials. We assume that the probability of exactly k occupied subintervals is given by $b(k;n,p(n))$.

As $n \to \infty$, $b(k; n, p(n)) \to p(k; \lambda)$, and since we are dealing with individual points, the number of occupied cells agrees in the limit with the number of points contained in our unit time interval.

In applications it is necessary to replace the unit interval by an interval of arbitrary length t. If we divide it again into subintervals of length $1/n$, the number of subintervals no is nt. Then the probability of finding exactly k points in a fixed interval of length t equals

$$p(k; \lambda t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$