Introduction to Probability and Statistics

Úvod do pravděpodobnosti a statistiky

doc. Mgr. Pasha Zusmanovich, PhD.
Ing. Pavel Rusnok

Ostravská univerzita
2016

Tato studijní opora vznikla při realizaci projektu IRP201605.

Contents Obsah

A Subject of Probability and Statistics

Probability theory is a branch of mathematics which tries to argue rigorously about random and uncertain things.

A word of warning: the words "probability" and "chance" used in the ordinary speech have not much in common with the mathematical probability. Basically, these everyday words invoke the idea of human confidence in an uncertain situation. On the other hand, rigorous measurements of probability, and mathematical handling of the results of these measurements, refer not to the confidence itself which is a psychological factor, but to objective numerical characteristics of the considered phenomena, initially closely related to count.

Though it cannot be said anything definite about outcomes of a single random event, when considering a number of such events in their totality, certain patterns emerge; these patterns are amenable to a rigorous mathematical study.

1 Předmět pravděpodobnosti a statistiky

Teorie pravděpodobnosti je odvětví matematiky, které se snaží argumentovat precizně o náhodných a nejistých věcech.

Jedno varování: slova „pravděpodobnost" a „náhoda", používané v mluvené řeči, nemají moc společného s matematickou pravděpodobností. V podstatě tato každodenní slova vyvolávají představu o lidské důvěře v nejisté situaci. Na druhé straně, přísné měření pravděpodobnosti, a matematická manipulace s výsledky těchto měření, se netýkají lidské důvěry, která je psychologickým faktorem, ale objektivní číselné charakteristiky uvažovaných jevů, které úzce souvisí s počítáním.

I když nelze říci nic konkrétního o výsledcích jednoho náhodného jevu, při uvažování o několika takových jevech v celém rozsahu se určité vzory objevují. Lze je pak poddat přísnému matematickému zkoumání.

Statistics deals with collection, analysis, interpretation, presentation, visualization, and organization of various data. Sometimes the terms "statistics" and "data science" are used interchangeably. Mathematical foundations of statistics are based on probability theory.

However, statistics is not confined to mathematics: its main utility lies in analyzing the "real world" data, be it stock prices, temperature measurements, or results of an opinion poll. It is crucial that the data are collected an a proper way, and that the appropriate methods to analyze it are used. This goes beyond mathematics and often requires methods from the relevant disciplines – biology, sociology, physics, finances, etc. Also, the data is often voluminous, and should be stored and processed efficiently; here a good deal of computer science is involved. As such, statistics is a multidisciplinary enterprise.

Statistika se zabývá sběrem, analýzou, interpretací, prezentací, vizualizací a organizací různých dat. Někdy jsou termíny „statistika“ a „datová analýza“ používány zaměnitelně. Matematické základy statistiky jsou založeny na teorii pravděpodobnosti.

Nicméně statistika není omezená pouze na matematiku: její hlavní použití spočívá v analyzování dat ze „skutečného světa“, jako jsou ceny akcií, měření teploty, nebo výsledky průzkumu veřejného mínění. Je velmi důležité, aby byla data sbírána správným způsobem a byly použity vhodné metody k analýze. Toto přesahuje hranice matematiky a často vyžaduje metod z příslušných disciplín – biologie, sociologie, fyziky, finančnictví, atd. Data jsou také často objemná a měla by být uložena a zpracována efektivně. Na tom se hodně podílí počítačová věda. Jako taková je statistika multidisciplinární iniciativou.

Sometimes, statistical analysis allows to judge in favor of statements, which otherwise may look controversial in some people's eyes. Just two relatively recent examples: the Russian 2012 presidential election was fraudulent (P. Klimek et al., Proc. Nat. Acad. Sci. **109** (2012), 16469–16473); in scientific publications, the top cited papers are not red by the majority of citing authors, but merely copied from one citation list to another (M.V. Simkin and V.P. Roychowdhury, Significance **3** (2006), 179–181).

Někdy statistická analýza umožňuje usuzovat ve prospěch výroků, které by jinak mohly vypadat kontroverzní v očích některých lidí. Zmíníme dva relativně nedávné příklady: Ruské prezidentské volby 2012 byly podvodné (P. Klimek et al., Proc. Nat. Acad. Sci. **109** (2012), 16469–16473); nejvíce citované články ve vědeckých publikacích nejsou čtené většinou citujících autorů, ale pouze kopírovány z jednoho seznamu citací do druhého (M.V. Simkin and V.P. Roychowdhury, Significance **3** (2006), 179–181).

Basic Combinatorics

Counting is heavily involved in a mathematical theory of probability; thus the significance of combinatorics.

Combinatorics is one the branches of mathematics. It is useful not only for probability theory. For example, the so-called Arrow theorem, one of the cornerstone results of Social Choice Theory, says that under reasonable assumptions, any choice in the society, no matter what the procedure to achieve the choice is (e.g., democratic voting), is dictatorial, i.e. is always the choice of a single individual ("dictator"). This somewhat counter-intuitive and disappointing (sociologically, not mathematically!) result is established with the help of not very difficult combinatorics.

The basic notions of combinatorics are permutations, variations, and combinations.

2 Základní kombinatorika

Počítání je důležitou součástí matematické teorie pravděpodobnosti a z toho vyplývá význam kombinatoriky.

Kombinatorika je jedním z odvětví matematiky. Je užitečná nejen pro teorii pravděpodobnosti. Například jeden ze základních kamenů teorie sociálního výběru – tzv. Arrowův teorém – říká, že za rozumných předpokladů je výběr ve společnosti bez ohledu na postup k dosažení volby (např. demokratické hlasování) diktátorský, tj. je vždy výběrem jediného jedince („diktátora"). Tento poněkud neintuitivní výsledek je zklamáním (sociologicky, ne matematicky!). Výsledek je dokázán pomocí nepříliš obtížné kombinatoriky.

Základní pojmy kombinatoriky jsou permutace, variace, a kombinace.

Permutation is the act of arranging elements of a set into some order. The number of permutations of an n-element set is equal to $n!$. Indeed, we can put n different elements on the first place. After one element is placed, we have $n-1$ possibilities to put an element at the second place. Repeating this procedure, on the n-th (last) step we will be left with one element to be put on the n-th place (i.e., there is no choice at all). As on each step the choices are independent, we have

Permutace je akt uspořádání prvků množiny do nějakého pořadí. Celkový počet permutací množiny o n prvcích je $n!$. Můžeme dát n různých prvků na první místo. Poté, co jeden prvek je umístěn, máme $n-1$ možností jaký prvek dát na druhé místo. Opakováním tohoto postupu nám pak na n-té (poslední) místo zbyde jeden prvek, který bude dán na n-tou pozici (tj., nemáme žádnou volbu). Jelikož v každém kroku jsou volby nezávislé, máme

$$n \cdot (n-1) \cdot \ldots \cdot 1 = n!$$

different possibilities.

různých možností.

Variation without repetition is a generalization of permutation, in which not necessary all elements of the set are used. As in the case of permutation, no element of the set occured more than once. The number of such arrangements of a k-element subset of an n-element set is equal to $\frac{n!}{(n-k)!}$. The counting is the same as in the case of permutation ($k = n$), stopping at the k-th step:

Variace bez opakování jsou zobecněním permutací, v nichž není nutné, aby se použily všechny prvky množiny. Stejně jako v případě permutací, žádný prvek množiny se neobjevuje více než jednou. Počet takových uspořádání k-prvkové podmnožiny n-prvkové množiny je rovno $\frac{n!}{(n-k)!}$. Počítáme stejně jako v případě permutací ($k = n$) a zastavíme se v k-tém kroku:

$$n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}$$

When we drop the requirement that no element occurs more than once (i.e., repetitions are allowed), we get the notion of variation with repetition. This is the same as the k-tuple of elements from an n-element set, and the number of such tuples is equal to n^k.

Combination is the way of selecting items from a set, such that the order of selection does not matter. If the set we are choosing from has n elements, and we are choosing k-element subsets, the number of ways it is possible to do is equal to $\binom{n}{k}$, the binomial coefficient. Indeed, to choose k elements out of n-element set is the same as perform a variation without repetition, hence the number of such choices is $\frac{n!}{(n-k)!}$. But since elements of a given k-element set can be chosen in arbitrary order, we have accounted each set the number of times equal to the number of permutations of its elements, i.e. $k!$. Hence the final number of possibilities is

Když vypustíme požadavek, aby se žádný prvek nevyskytoval více než jednou (tj., opakování jsou povolena), dostaneme pojem variace s opakováním. To je stejné jako k-tice prvků z n prvkové množiny. A počet těchto k-tic je roven n^k.

Kombinace je způsob výběru prvků z množiny, tak že nezáleží na pořadí výběru. Pokud má množina, ze které vybíráme n prvků a my vybíráme k-prvkovou podmnožinu, pak je počet způsobů, jak je to možné udělat roven kombinačnímu číslu $\binom{n}{k}$. Ve skutečnosti je výběr k prvků z n-prvkové množiny stejný jako variace bez opakování, protože počet těchto voleb je $\frac{n!}{(n-k)!}$. Ale jelikož prvky dané k-prvkové množiny můžeme vybrat v libovolném pořadí, pak jsme započítali každou množinu tolikrát kolik je počet jejích permutací, tj. $k!$. Proto je konečný počet možností

$$\frac{n!}{(n-k)!k!} = \binom{n}{k}.$$

Example 1 For example, from the standard deck of 52 cards one possible to choose a 5-card hand in

$$\binom{52}{5} = \frac{52!}{47!5!} = 2{,}598{,}960$$

different ways.

Příklad 1 Například ze standardního balíčku 52 karet je možné vybrat 5 karet

$$\binom{52}{5} = \frac{52!}{47!5!} = 2{,}598{,}960$$

různými způsoby.

Sample Space, Event, Probability

A sample space is a set whose elements represent the possible outcomes of the event we are interested in.

Example 2 When tossing a coin, the sample space is a 2-element set $S = \{\text{head}, \text{tail}\}$. When tossing a dice, the sample space is a 6-element set $\{1, 2, 3, 4, 5, 6\}$. When tossing two coins simultaneously, the sample space is the cartesian product $S \times S$, i.e. the set

$$\{(\text{head}, \text{head}), (\text{head}, \text{tail}), (\text{tail}, \text{head}), (\text{tail}, \text{tail})\}.$$

If we are throwing an (idealized) dart (i.e., a point) at an (idealized) dartboard (say, a circle with radius 1 with the center at the origin), the sample space is

3 Prostor elementárních jevů, náhodný jev, pravděpodobnost

Prostor elementárních jevů je množina takových prvků, které představují výsledky jevu, o který se zajímáme.

Příklad 2 Pokud házíme mincí, pak je prostor elementárních jevů dvouprvková množina $S = \{\text{líc}, \text{rub}\}$. Pokud házíme hrací kostkou, pak prostor elementárních jevů je šestiprvková množina $\{1, 2, 3, 4, 5, 6\}$. Pokud házíme dvěma mincemi najednou, pak je prostor kartézský součin $S \times S$, t.j. množina

$$\{(\text{líc}, \text{líc}), (\text{líc}, \text{rub}), (\text{rub}, \text{líc}), (\text{rub}, \text{rub})\}.$$

Pokud házíme (idealizovanou) šipkou (tj. bod) na (idealizovaný) terč (řekněme, kruh s poloměrem 1 s centrem v počátku souřadnic), pak prostor elementárních jevů je

$$\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}.$$

An event is a subset of the sample space. An elementary event is a subset of the sample space consisting of one element.

Example 3 We may sample birthday dates in a certain group of people. The sample space in this case is the subset of the cartesian product

$$\{1, 2, \ldots, 31\} \times \{\text{Jan}, \text{Feb}, \ldots, \text{Dec}\}$$

(subset, as certain pairs, like February 30 and June 31, are excluded). Now we may be interested, for example, in people born at the specific date, so the one-element sets $\{(1, \text{Jan})\}$ and $\{(10, \text{Mar})\}$ will constitute elementary events, while all birthdays occuring in February:

$$\{1, 2, \ldots, 29\} \times \{\text{Feb}\},$$

or all birthdays occuring at the end of the month:

$$\{(30, \text{Jan}), (28, \text{Feb}), (29, \text{Feb}), \ldots, (31, \text{Dec})\}$$

will constitute (just) events.

Náhodný jev je podmnožinou prostoru elementárních jevů. Elementární jev je jednoprvková podmnožina prostoru elementárních jevů.

Příklad 3 Můžeme vybrat datumy narozenin v určité skupině lidí. Prostor elementárních jevů je v tomto případě podmnožinou kartézského součinu

$$\{1, 2, \ldots, 31\} \times \{\text{Led}, \text{Úno}, \ldots, \text{Pro}\}$$

(podmnožina, protože některé páry, jako 30. února a 31. června, jsou vyloučeny). Nyní můžeme mít zájem například o lidi narozené v určitý den, pak jednoprvkové množiny $\{(1, \text{Led})\}$ a $\{(10, \text{Bře})\}$ budou představovat elementární jevy, zatímco všechny narozeniny vyskytující se v únoru:

$$\{1, 2, \ldots, 29\} \times \{\text{Úno}\},$$

nebo všechny narozeniny vyskytující se na konci měsíce:

$$\{(30, \text{Led}), (28, \text{Úno}), (29, \text{Úno}), \ldots, (31, \text{Pro})\}$$

budou představovat (pouze) náhodné jevy.

In the throwing darts example, the perfect hit, $\{(0,0)\}$, is an elementary event, while hitting, say, the right half of the dartboard:

V příkladu házení šipkami je perfektní zásah $\{(0,0)\}$ elementární jev, zatímco zásah pravé části terče:

$$\{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1, x \geq 0\}$$

will constitute an event.

je náhodný jev.

The set-theoretic operations on events correspond to their logical combinations: the intersection $A \cap B$ of events A and B occurs when both A and B occur; the union $A \cup B$ occurs when either A or B occurs; the complement $S \setminus A$, where S is the whole sample space, occurs when A does not occur.

Množinové operace na náhodných jevech odpovídají jejich logickým kombinacím: k průniku $A \cap B$ jevů A a B dochází, když oba jevy A a B nastanou; sjednocení $A \cup B$ nastane, když buď A nebo B nastane; doplněk $S \setminus A$, kde S je celý prostor elementárních jevů, nastane, když A nenastane.

Example 4 When tossing two coins simultaneously, the event of having head first is

$$A = \{(\text{head}, \text{head}), (\text{head}, \text{tail})\},$$

and the event of having head second is

$$B = \{(\text{head}, \text{head}), (\text{tail}, \text{head})\}.$$

Their intersection is an elementary event having both heads, $\{(\text{head}, \text{head})\}$, and their union is an event of having at least one head:

$$A \cup B = \{(\text{head}, \text{head}), (\text{tail}, \text{head}), (\text{head}, \text{tail})\}.$$

The complement of A is an event of having tails first:

$$S \setminus A = \{(\text{tail}, \text{head}), (\text{tail}, \text{tail})\}.$$

Příklad 4 Když hodíme dvě mince současně, náhodný jev hození nejprve líce je

$$A = \{(\text{líc}, \text{líc}), (\text{líc}, \text{rub})\},$$

a náhodný jev líce na druhém místě je

$$B = \{(\text{líc}, \text{líc}), (\text{rub}, \text{líc})\}.$$

Jejich průnik je elementární jev hození dvakrát líce $\{(\text{líc}, \text{líc})\}$ a jejich sjednocení je náhodný jev hození alespoň jednoho líce

$$A \cup B = \{(\text{líc}, \text{líc}), (\text{rub}, \text{líc}), (\text{líc}, \text{rub})\}.$$

Doplněk k A je náhodný jev, kdy hodíme nejprve rub:

$$S \setminus A = \{(\text{rub}, \text{líc}), (\text{rub}, \text{rub})\}.$$

Probability is a numerical expression of how likely an event occurs. If all outcomes in the sample space S occurs equally likely, then the probability $\Pr(A)$ of an event A is equal to $\frac{|A|}{|S|}$, where $|X|$ is the cardinality (number of elements) of the set X. In particular, the probability of an elementary event is equal to $\frac{1}{|S|}$.

Example 5 In the previous example with two coins tossing, we have

Pravděpodobnost je číselné vyjádření, jaká je šance, že nastane nějaká událost. Pokud všechny výsledky v prostoru S se vyskytují stejně často, pak pravděpodobnost $\Pr(A)$ náhodného jevu A je roven $\frac{|A|}{|S|}$, kde $|X|$ je mohutnost (počet prvků) množiny X. Zejména pravděpodobnost elementárního jevu je rovna $\frac{1}{|S|}$.

Příklad 5 V předchozím příkladu házení dvěma mincemi máme

$$\Pr(A) = \Pr(B) = \Pr(S \setminus A) = \frac{2}{4} = \frac{1}{2}$$
$$\Pr(A \cap B) = \frac{1}{4}$$
$$\Pr(A \cup B) = \frac{3}{4}.$$

More generally, a probability function p on a sample space S is a function from the set of all possible events, i.e. the powerset $P(S)$, to the interval $[0, 1]$, such that $\Pr(S) = 1$ and

Obecněji pravděpodobnost p na prostoru S je funkce z množiny všech možných jevů, tj. potenční množiny $P(S)$, do intervalu $[0, 1]$ tak, že $\Pr(S) = 1$ a

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

if A and B do not occur simultaneously, i.e. $A \cap B = \varnothing$. Thus, the probability of an event can be computed by summing up probabilities of all outcomes (elementary events) comprising it.

pokud A a B se nevyskytují současně, tj. $A \cap B = \varnothing$. To znamená, že pravděpodobnost náhodného jevu lze vypočítat sečtením pravděpodobností všech výsledků (elementárních jevů), které ho tvoří.

Example 6 If we are throwing a crooked dice, where 6 can occur with the probability $\frac{1}{5} = 0.2$ (instead of the fair $\frac{1}{6}$), and the rest of points, from 1 till 5, can occur with the equal probability 0.16, the probability to get an even number of points is equal to

Příklad 6 Pokud házíme křivou kostkou, u níž se 6 vyskytuje s pravděpodobností $\frac{1}{5} = 0{,}2$ (namísto vyvážené $\frac{1}{6}$), a zbytek bodů od 1 do 5, může nastat se stejnou pravděpodobností 0,16, pak pravděpodobnost, že dostaneme sudý počet bodů, se rovná

$$\Pr(\{2, 4, 6\}) = \Pr(\{2\}) + \Pr(\{4\}) + \Pr(\{6\}) = 2 \cdot 0.16 + 0.2 = 0.52.$$

Still, these notions of event and probability are not entirely satisfactory, as we can run into problems with infinite sets. When the sample space S is infinite, the appropriate notion of event appears to be not an arbitrary subset of S, but an element of a σ-algebra, i.e. a set of subsets of S closed with respect to complements, and countable unions and intersections. Then the probability function p is defined as a measure on the σ-algebra, normalized by the condition $\Pr(S) = 1$.

Example 7 In the throwing darts example, the probability to hit any measurable subset A of our idealized dartboard is equal to $\frac{\mu(A)}{\pi}$. For example, the probability of the perfect hit is zero (as the measure of a set consisting of a single point is zero), while the probability to hit the right half of the dartboard is $\frac{1}{2}$.

Stále nejsou pojmy náhodného jevu a pravděpodobnosti zcela uspokojivé, protože můžeme narazit na problémy u nekonečných množin. Je-li prostor elementárních jevů S nekonečný, pak vhodný pojem náhodného jevu se nezdá být libovolná podmnožina S, ale nějaký prvek σ-algebry, tj. souboru podmnožin S uzavřenému vůči doplňkům a spočetným sjednocením a průnikům. Pravděpodobnost p je pak definována jako míra na σ-algebře, normalizovaná podmínkou $\Pr(S) = 1$.

Příklad 7 V příkladu házení šipkou je pravděpodobnost zásahu jakékoli měřitelné podmnožina A našeho idealizovaného terče rovna $\frac{\mu(A)}{\pi}$. Například, pravděpodobnost perfektního zásahu je nula (tak jako míra množiny, která obsahuje jediný bod, je nula), zatímco pravděpodobnost zásahu pravé poloviny terče je $\frac{1}{2}$.

Conditional Probability, Bayes' Formula, Independent Events

A conditional probability, denoted by $\Pr(A|B)$, is a probability of an event A assuming that another event B has occured. It is defined as

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

(Of course, we assume here that $\Pr(B) > 0$.)

4 Podmíněná pravděpodobnost, Bayesův vzorec, nezávislé náhodné jevy

Podmíněná pravděpodobnost $\Pr(A|B)$ je pravděpodobnost náhodného jevu A předpokládající, že jiný náhodný jev B nastal. Je definována jako

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

(Samozřejmě předpokládáme, že $\Pr(B) > 0$.)

Example 8 Assuming a non-leap year, let A be an event that a person has a birthday at the first day of a month, and B an event that a person has a birthday at an odd-numbered day at summer. Then $A \cap B$ is an event that a person has a birthday at the first day of a summer month,

Příklad 8 Za předpokladu nepřechodného roku, nechť A je náhodný jev, že osoba má narozeniny první den v měsíci a B je náhodný jev, kdy osoba má narozeniny v lichém dni během léta. Pak $A \cap B$ je jev, kdy osoba má narozeniny v prvním dni letního měsíce,

$$\Pr(A \cap B) = \frac{3}{365},$$
$$\Pr(B) = \frac{15+16+16}{30+31+31} = \frac{47}{92},$$
$$\Pr(A|B) = \frac{\frac{3}{365}}{\frac{47}{92}} = \frac{276}{17155} \approx 0.016.$$

Compare this with the value of unconditional probability

Porovnejte to s hodnotou nepodmíněné pravděpodobnosti

$$\Pr(A) = \frac{12}{365} \approx 0.033.$$

Having two events A and B with nonzero probabilities, along with the conditional probability $\Pr(A|B)$, we may consider the conditional probability

Pokud máme dva náhodné jevy A a B s nenulovými pravděpodobnostmi společně s podmíněnou pravděpodobností $\Pr(A|B)$, pak můžeme uvažovat o podmíněné pravděpodobnosti

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)},$$

what implies

což implikuje

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}.$$

This is known as Bayes' formula.

To je známé jako Bayesův vzorec.

Two events A and B are called independent, if one of the following equivalent equalities holds:

Dva náhodné jevy A a B se nazývají nezávislé, pokud jedna z následujících ekvivalentních rovností platí:

$$\begin{aligned}\Pr(A|B) &= \Pr(A),\\ \Pr(B|A) &= \Pr(B),\\ \Pr(A \cap B) &= \Pr(A)\Pr(B).\end{aligned}$$

The equivalence follows from the definition of conditional probability, and Bayes' formula.

Ekvivalence plyne z definice podmíněné pravděpodobnosti a Bayesova vzorce.

Example 9 For simplicity of calculation in this example assume that every month of year has 30 days. For example, an event of having a birthday specified in terms of the day of the month (e.g., at the 10th day of the month, at odd days, from 10th till 15th day, etc.) is independent from the event of having birthday specified in terms of the month (e.g., at January, at spring, at the last 3 months of the year, etc.). On the other hand, the events of having birthday at summer, and at the odd-numbered months are not independent (intuitively this is clear, but check it numerically!)

Příklad 9 Pro jednodušší počítání v tomto příkladu předpokládejme, že každý měsíc v roce má 30 dní. Náhodný jev mít narozeniny specifikované pomocí dnů měsíce (např. 10. den měsíce, liché dny, od 10. do 15. dne, atd.) je nezávislý na náhodném jevu mít narozeniny specifikované pomocí měsíce (např. v lednu, na jaře, ve třech posledních měsících v roku, atd.). Na druhé straně, náhodné jevy narozenin v létě, a v lichých měsících nejsou nezávislé (intuitivně to je jasné, ale zkontrolujte to numericky!)

Discrete Random Variable, Distribution Function

In some situations, we may be interested not in the sample space itself, but only in some of its features. This leads us to the notion of a random variable.

A discrete random variable is a function on the sample space S with values in $\mathbb{R}$, accepting finite or countable number of different values.

Of course, if the sample space if finite, then any random variable defined on it is discrete.

Example 10 When throwing pair of dices, we may be interested not in the exact outcome, but merely in the sum of two throws, or in the maximum of two throws. These are examples of a discrete random variable.

5 Diskrétní náhodná proměnná, distribuční funkce

V některých situacích se nemusíme zajímat o prostor elementárních jevů samotný, ale pouze o některé jeho vlastnosti. To nás vede k pojmu náhodné proměnné.

Diskrétní náhodná proměnná je funkce z prostoru elementárních jevů S do $\mathbb{R}$ s konečným nebo spočetným množstvím různých hodnot.

Samozřejmě, pokud je prostor elementárních jevů konečný, pak jakákoli náhodná proměnná definována na něm je diskrétní.

Příklad 10 Když házíme dvěma kostkami, nemusíme se zajímat o přesný výsledek, ale pouze o součet dvou hodů, nebo o maximum z dvou hodů. To jsou příklady diskrétní náhodné proměnné.

Let $X : S \to \mathbb{R}$ be a discrete random variable defined on a sample space S. The mass function of X is the function $f_X : \mathbb{R} \to [0, 1]$ defined for any $x \in \mathbb{R}$ as

Nechť $X : S \to \mathbb{R}$ je diskrétní náhodná proměnná definována na prostoru elementárních jevů S. Pravděpodobnostní funkce náhodné proměnné X je funkce $f_X : \mathbb{R} \to [0, 1]$ definována pro libovolné $x \in \mathbb{R}$

$$f_X(x) = \Pr(X = x).$$

The distribution function of X is the function $F_X : \mathbb{R} \to [0, 1]$ defined for any $x \in \mathbb{R}$ by

Distribuční funkce proměnné X je funkce $F_X : \mathbb{R} \to [0, 1]$ definována pro libovolné $x \in \mathbb{R}$ pomocí

$$F_X(x) = \Pr(X \leq x).$$

(Note that formally the right-hand sides of the last two formulas had to be written as

(Všimněte si, že formálně by pravé strany posledních matematických vzorců měly být zapsané jako

$$\Pr(\{s \in S \mid X(s) = x\}),$$

and

a

$$\Pr(\{s \in S \mid X(s) \leq x\}),$$

respectively, but here and in similar situations below, we use universally accepted shorthands).

ale tady a v podobných situacích níže používáme obecně přijaté zkratky).

By definition, the mass function attains possibly non-zero values in the finite or countable number of points (the values of the discrete random variable X), and is zero elsewhere. We have

Dle definice nabývá pravděpodobnostní funkce případných nenulových hodnot v konečném nebo spočetném počtu bodů (hodnoty diskrétní náhodné proměnné X), a je nulová všude jinde. Máme

$$F_X(x) = \sum_{t \leq x} f_X(t)$$

for any $x \in \mathbb{R}$, so the distribution function is always non-decreasing, and has "jumps" only in a finite or countable number of points.

pro libovolné $x \in \mathbb{R}$. Distribuční funkce je tudíž vždy neklesající, a má „skoky“ pouze v konečném nebo spočetném množství bodů.

Example 11 In the tossing coin example, let us assign numerical values of 0 and 1 to tail and head respectively, and on the sample space S of outcomes of tossing 3 coins simultaneously, consider the random variable X equal to the sum of all 3 outcomes, so the possible values of X are $0, 1, 2, 3$. Let us compute the corresponding mass and distribution functions.

Příklad 11 V příkladu házení mincí přiřaďme číselné hodnoty 0 a 1 rubu a líci a na prostoru elementárních jevů S výsledků házení třemi mincemi najednou uvažujeme náhodnou proměnnou X rovnou součtu všech 3 výsledků. Proto možné hodnoty X jsou $0, 1, 2, 3$. Spočítejme příslušné pravděpodobnostní a distribuční funkce.

$$f_X(0) = \Pr(X = 0) = \Pr(\{(0,0,0)\}) = \tfrac{1}{8}; \qquad F_X(0) = f_X(0) = \tfrac{1}{8};$$

$$f_X(1) = \Pr(X = 1) = \Pr(\{(1,0,0),(0,1,0),(0,0,1)\}) = \tfrac{3}{8}; \qquad F_X(1) = f_X(0) + f_X(1) = \tfrac{1}{8} + \tfrac{3}{8} = \tfrac{1}{2};$$

$$f_X(2) = \Pr(X = 2) = \Pr(\{(1,1,0),(1,0,1),(0,1,1)\}) = \tfrac{3}{8}; \qquad F_X(2) = f_X(0) + f_X(1) + f_X(2) = \tfrac{1}{8} + \tfrac{3}{8} + \tfrac{3}{8} = \tfrac{7}{8};$$

$$f_X(3) = \Pr(X = 3) = \Pr(\{(1,1,1)\}) = \tfrac{1}{8}; \qquad F_X(3) = f_X(0) + f_X(1) + f_X(2) + f_X(3) = \tfrac{1}{8} + \tfrac{3}{8} + \tfrac{3}{8} + \tfrac{1}{8} = 1.$$

At any other points, the value of f_X is zero, and

Ve všech zbylých bodech jsou hodnoty f_X rovny nule a

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0; \\ F_X([x]) & \text{if } 0 \le x \le 3; \\ 1 & \text{if } x > 3. \end{cases}$$

Here $[x]$ denotes the integer part of x.

Pomocí $[x]$ označujeme celočíselnou část x.

Continuous Random Variable, Density Function

If a real-valued function defined on the sample space S attains not a discrete, but a continuous range of values, we arrive at the notion of a continuous random variable. Formally, a random variable $X : S \to \mathbb{R}$ is continuous, if

$$\Pr(a \leq X \leq b) = \int_a^b f_X(t)\,\mathrm{d}t$$

for some function $f_X : \mathbb{R} \to \mathbb{R}$, and any $a, b \in \mathbb{R}$, $a \leq b$. The function f_X is called the density function of X. It is a continuous analog of the mass function of a discrete random variable.

Note that any density function attains only non-negative values, and satisfies

$$\int_{-\infty}^{\infty} f_X(t)\,\mathrm{d}t = 1.$$

6 Spojitá náhodná proměnná, hustota pravděpodobnosti

Pokud reálná funkce definovaná na prostoru elementárních jevů S nenabývá diskrétních hodnot, ale spojitého intervalu hodnot, pak se dostáváme k pojmu spojité náhodné proměnné. Formálně, náhodná proměnná $X : S \to \mathbb{R}$ je spojitá, pokud

pro nějakou funkci $f_X : \mathbb{R} \to \mathbb{R}$ a libovolné $a, b \in \mathbb{R}$, $a \leq b$. Funkce f_X se nazývá hustota pravděpodobnosti proměnné X. Je to spojitá analogie pravděpodobnostní funkce diskrétní náhodné proměnné.

Všimněte si, že hustota pravděpodobnosti nabývá pouze nezáporných hodnot a splňuje, že

The distribution function $F_X : \mathbb{R} \to [0,1]$ of a continuous random variable X is defined the same way as for a discrete one:

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$$

for any $x \in \mathbb{R}$.

Example 12 In the throwing darts example, a distance from a given point to the origin (or any other "good behaving" real function defined on the unit circle), is a continuous random variable.

In the real world, we are dealing with discrete random variables, even with a particular case of them which involves only finite number of possible values. Continuous random variables are very useful mathematical abstractions helping to capture important properties of the discrete case when the number of possible values is becoming huge. This explains a big similarity between discrete and continuous random variables: as a rule of thumb, any formula, result, or reasoning involving the discrete case can be turned into the continuous one, by replacing summation by integration.

Distribuční funkce $F_X : \mathbb{R} \to [0,1]$ spojité náhodné proměnné X je definována stejným způsobem jako pro diskrétní:

pro libovolné $x \in \mathbb{R}$.

Příklad 12 V příkladu házení šipkami je vzdálenost od daného bodu k počátku (nebo jakákoli jiná „dobře se chovající" reálná funkce definována na jednotkovém kruhu) spojitá náhodná proměnná.

V reálném světě máme co do činění s diskrétními náhodnými proměnnými, a to i v konkrétním případě, který zahrnuje pouze konečný počet možných hodnot. Spojité náhodné proměnné jsou velmi užitečné matematické abstrakce pomáhající zachytit důležité vlastnosti diskrétního případu, kdy je počet možných hodnot obrovský. To vysvětluje velkou podobnost mezi diskrétními a spojitými náhodnými proměnnými: jako pravidlo, libovolný vzorec, výsledek, nebo úvaha zahrnující diskrétní případ může být přenesená do spojitého, nahrazením sumace integrací.

We can operate with random variables defined on the same sample space, both discrete and continuous, the same way as we operate with functions: we can add them, multiply them, apply other functions to them, etc.

Můžeme operovat s náhodnými proměnnými definovanými na stejném prostoru elementárních jevů, a to jak s diskrétními tak spojitými, stejně jako operujeme s funkcemi: můžeme je sčítat, násobit, aplikovat na ně jiné funkce, atd.

Numerical Characteristics of a Random Variable

Random variables may contain a huge amount of data in a very complicated form, so sometimes one wants to summarize that or another property of a random variable by a single number.

The expected value, or mean, of a random variable X, denoted by $E[X]$, is its average value, or, in other words, the center of the corresponding distribution function. For a discrete random variable attaining values $x_1, x_2, \ldots$, the expected value is just the weighted mean of the values, with weights being the respective probabilities:

$$E[X] = \sum_{i=1,2,\ldots} f_X(x_i)x_i.$$

Note that, generally, we are dealing here with an infinite sum, which may not exist. However, it does exist in most of the important cases occuring on practice. Of course, if the random variable X attains only finite number of values, the sum is finite and thus exists always.

7 Číselné charakteristiky náhodné proměnné

Náhodné proměnné mohou obsahovat velké množství dat ve velmi složité formě, takže někdy můžeme chtít shrnout tuto nebo jinou vlastnost náhodné proměnné jediným číslem.

Očekávaná hodnota nebo průměr náhodné proměnné X, který značíme $E[X]$, je její průměrná hodnota, nebo jinými slovy, střed odpovídající distribuční funkce. Pro diskrétní náhodnou proměnnou nabývající hodnot $x_1, x_2, \ldots$ je očekávaná hodnota jen váženým průměrem těchto hodnot – váženými příslušnými pravděpodobnostmi:

Všimněte si, že obecně zde pracujeme s nekonečným součtem, který nemusí existovat. Nicméně, existuje ve většině důležitých případů objevujících se v praxi. Samozřejmě, že v případě, že náhodná proměnná X nabývá pouze konečnéno počtu hodnot, suma je konečná, a tudíž existuje vždycky.

Example 13 The expected value of the random variable equal to the number of points got in one throw of a dice is equal to

Příklad 13 Očekávaná hodnota náhodné proměnné, která je rovna počtu bodů na hozené kostce, se rovná

$$\frac{1}{6}\cdot 1+\frac{1}{6}\cdot 2+\frac{1}{6}\cdot 3+\frac{1}{6}\cdot 4+\frac{1}{6}\cdot 5+\frac{1}{6}\cdot 6=3.5.$$

The expected value of a continuous random variable X is defined as

Očekáváná hodnota spojité náhodné proměnné X je definována

$$E[X]=\int_{-\infty}^{\infty} t f_X(t)\,\mathrm{d}t.$$

The p-th quantile of a random variable X, where p is a number between 0 and 1, is the smallest number q_p such that

p-tý kvantil náhodné proměnné X, kde p je číslo mezí 0 a 1, je nejmenší číslo q_p takové, že

$$\Pr(X\leq q_p)=p.$$

A particular case of quantile is median, which is defined as 0.5th quantile. Sometimes mean is not an adequate characteristic of a random variable. For example, the mean of the yearly income per household in countries with very unevenly distributed wealth (think US, not Czech Republic) would exhibit values much higher then "expected", due to a relatively small number of embarrassingly wealthy individuals. In such cases, a more adequate representation of a "mean" value would be given by median. Informally, the median is the value which "sits in the middle", and it is much less sensitive than mean to extreme values in the data. Another frequently used in practice quantiles are quartiles, which are defined as 0.25th, 0.50th, and 0.75th quantiles.

The standard deviation expresses the idea how "spread" the random variable is. The standard deviation of a random variable X (both discrete and continuous), denoted by $\sigma(X)$, is defined as

Jedním příkladem kvantilu je medián, který je definován jako 0,5-tý kvantil. Průměr někdy není dostačující charakteristikou náhodné proměnné. Například průměrný roční příjem domácnosti v zemích s velmi nerovnoměrně rozloženým bohatstvím (například USA, nikoli Česká republika) by vykazoval hodnoty mnohem vyšší než jsou „očekávány", vzhledem k relativně malému počtu až trapně bohatých jednotlivců. V takových případech by vhodnější zastoupení „průměrné" hodnoty bylo dáno mediánem. Neformálně je medián hodnotou, která „sedí uprostřed", a je mnohem méně citlivý na extrémní hodnoty v datech než průměr. Další často používané v praxi kvantily jsou kvartily, které jsou definovány jako 0,25-tý, 0,50-tý a 0,75-tý kvantil.

Směrodatná odchylka vyjadřuje úvahu o tom, jak „široká" je náhodná proměnná. Směrodatná odchylka náhodné proměnné X (diskrétní i spojité), označována $\sigma(X)$, je definována jako

$$\sigma(X) = \sqrt{E[(X - E[X])^2]}.$$

If X is a discrete random variable attaining, with the equal probability $\frac{1}{n}$, a finite number of n distinct values $x_1, \ldots, x_n$, then

Pokud je X diskrétní náhodná proměnná nabývající se stejnou pravděpodobností $\frac{1}{n}$ konečného počtu n různých hodnot $x_1, \ldots, x_n$, pak

$$E[X] = \frac{x_1 + \cdots + x_n}{n}$$

and

a

$$\sigma(X) = \sqrt{\frac{(x_1 - E[X])^2 + \cdots + (x_n - E[X])^2}{n}}.$$

The latter formula explains why indeed the standard deviation is a good measure of how spread the data is: the more the values x_i stay away from their mean $E[X]$, the bigger $\sigma(X)$ would be.

Druhý vzorec vysvětluje, proč je skutečně směrodatná odchylka dobrým měřítkem toho, jak široká jsou data: čím více jsou hodnoty x_i daleko od jejich průměru $E[X]$, tím větší $\sigma(X)$ bude.

Discrete Distributions: Uniform, Binomial, Poisson, Hypergeometric

Some types of distributions are of utmost importance, as they appear often on practice, and provide a convenient material for building effective statistical models.

Perhaps the simplest possible distribution is an uniform one. A discrete random variable is distributed uniformly, if its mass function attains the same value at the finite number of n points. The mass function of the uniform distribution is of the form

$$f_n(k) = \frac{1}{n},$$

where $k = 1, 2, \ldots, n$. The expected value and the standard deviation of an uniformly distributed random variable are equal to $\frac{n+1}{2}$ and $\sqrt{\frac{n^2-1}{12}}$, respectively.

8 Diskrétní rozdělení: rovnoměrné, binomické, Poissonovo, hypergeometrické

Některé typy distribucí (rozdělení) jsou nanejvýš důležité, protože se objevují často v praxi, a poskytují pohodlný materiál pro stavění efektivních statistických modelů.

Snad nejjednodušší rozdělení je rovnoměrné rozdělení. Diskrétní náhodná proměnná je rozložena rovnoměrně, pokud její pravděpodobnostní funkce nabývá stejné hodnoty na konečném počtu n bodů. Pravděpodobnostní funkce rovnoměrného rozdělení je ve tvaru

$$f_n(k) = \frac{1}{n},$$

kde $k = 1, 2, \ldots, n$. Očekávaná hodnota a směrodatná odchylka rovnoměrně rozložené náhodné proměnné se rovná $\frac{n+1}{2}$ respektive $\sqrt{\frac{n^2-1}{12}}$.

Example 14 Our favorite random variable examples of throwing a single (fair) dice or tossing a single (fair) coin are uniformly distributed.

The binomial distribution with parameters $n = 1, 2, \ldots$ and p, where $0 \leq p \leq 1$, is the discrete distribution of the number of successes in a sequence of n experiments with a binary outcome (success/failure), each of which yields success with probability p. The mass function of the binomial distribution has the form

Příklad 14 Naše oblíbené příklady náhodných proměnných házení jednou (vyváženou) kostkou nebo házení jednou (vyváženou) mincí jsou rovnoměrně rozloženy.

Binomické rozdělení s parametry $n = 1, 2, \ldots$ a p, kde $0 \leq p \leq 1$, je diskrétní rozdělení počtu úspěchů v posloupnosti n experimentů s binárním výsledkem (úspěch/neúspěch). Každý experiment dává úspěch s pravděpodobností p. Pravděpodobnostní funkce binomického rozdělení má tvar

$$f_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where $k = 0, 1, 2, \ldots, n$.

kde $k = 0, 1, 2, \ldots, n$.

Example 15 Suppose that a biased coin comes up heads with probability 0.3. Then the probability to have 4 heads after 6 tosses is equal to

Příklad 15 Předpokládejme například, že u nevyvážené mince vyjde líc s pravděpodobností 0,3. Pak pravděpodobnost mít 4 líce po 6 hodech se rovná

$$f_{6,0.3}(4) = \binom{6}{4} 0.3^4 (1-0.3)^{6-4} \approx 0.0595.$$

The expected value of a binomially distributed random variable is equal to

$$\sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = np,$$

and the standard deviation is equal to

$$\sqrt{np(1-p)}.$$

The Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event. It is defined as a discrete distribution with parameter $\mu > 0$ (estimated number of events) whose mass function has the form

$$f_\mu(k) = \frac{\mu^k}{k!} e^{-\mu},$$

where $k = 0, 1, 2, \ldots$.

Očekáváná hodnota proměnné s binomickým rozdělením je rovna

$$\sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = np,$$

a směrodatná odchylka je rovna

$$\sqrt{np(1-p)}.$$

Poissonovo rozdělení vyjadřuje pravděpodobnost určitého počtu událostí vyskytujících se v pevném časovém intervalu a/nebo prostoru, jestliže se tyto události vyskytují se známou průměrnou mírou a nezávisle na době od poslední události. Je definováno jako diskrétní rozdělení s parametrem $\mu > 0$ (odhadovaný počet událostí), jehož pravděpodobnostní funkce má tvar

$$f_\mu(k) = \frac{\mu^k}{k!} e^{-\mu},$$

kde $k = 0, 1, 2, \ldots$.

The expected value of a random variable whose distribution function is Poisson, is equal to

Očekávaná hodnota náhodné proměnné, jejíž rozdělení je Poissonovo, je rovna

$$\sum_{k=0}^{\infty} k \frac{\mu^k}{k!} e^{-\mu} = \mu,$$

and the standard deviation is equal to $\sqrt{\mu}$.

a směrodatná odchylka se rovná $\sqrt{\mu}$.

The hypergeometric distribution with parameters N, K, n, where N is a non-negative integer, and K and n are integers ranging from 0 till N, describes the number of successes in n binary (success/failure) draws, without replacement, from a finite set of N elements, that contains exactly K successes. The mass function of the hypergeometric distribution has the form

Hypergeometrické rozdělení s parametry N, K, n, kde N je nezáporné celé číslo a K a n jsou celá čísla v rozmezí od 0 do N, popisuje množství úspěchů mezi n binárními (úspěch/neúspěch) tahy, bez vracení, z konečné množiny N prvků, který obsahuje přesně K úspěchů. Pravděpodobnostní funkce hypergeometrického rozdělení má tvar

$$f_{N,K,n}(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

where $k = 0, 1, 2, \ldots, \min(n, K)$. The expected value is equal to $\frac{nK}{N}$ and the standard deviation is

kde $k = 0, 1, 2, \ldots, \min(n, K)$. Očekávaná hodnota je rovna $\frac{nK}{N}$ a směrodatná odchylka je

$$\frac{1}{N}\sqrt{\frac{nK(N-K)(N-n)}{N-1}}.$$

Continuous Distributions: Uniform, Normal, Exponential

Again, among continuous distributions the continuous uniform distribution has the most simple form: its density function is a constant within a given range. More precisely, the density function of the uniform distribution on the interval $[a, b]$ is defined as

$$f_{a,b}(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

The expected value and the standard deviation of an uniformly distributed continuous random variable are equal to $\frac{a+b}{2}$ and $\frac{b-a}{2\sqrt{3}}$, respectively.

9 Spojité rozdělení: rovnoměrné, normální, exponenciální

Opět platí, že mezi spojitými distribucemi spojité rovnoměrné rozdělení má nejjednodušší tvar: jeho hustota je konstantní v rámci daného intervalu. Přesněji řečeno je hustota pravděpodobnosti rovnoměrného rozdělení na intervalu $[a, b]$ definována jako

$$f_{a,b}(x) = \begin{cases} \frac{1}{b-a} & \text{pokud } a \leq x \leq b, \\ 0 & \text{jinak.} \end{cases}$$

Očekávaná hodnota a směrodatná odchylka rovnoměrně rozdělené spojité náhodné proměnné je rovna $\frac{a+b}{2}$ a $\frac{b-a}{2\sqrt{3}}$.

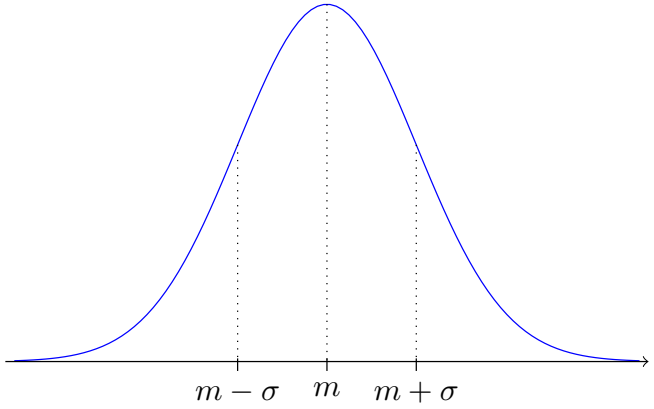
The normal distribution with parameters m and σ is a continuous distribution with the density function of the form

Normální rozdělení s parametry m a σ je spojité rozděleni s hustotou pravděpodobnosti ve tvaru

$$f_{m,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2}.$$

The graph of this density function has the famous "bell-shaped" form, with the maximum around $x = m$.

Graf této hustoty pravděpodobnosti má slavný „zvonovitý" tvar, přičemž maximální je kolem $x = m$.



This is, perhaps, the single most important distribution, due to the Central Limit Theorem, one of the cornerstones results in probability and statistics. Roughly, this theorem says that, under certain natural conditions, the average of a large number of identically distributed random variables is distributed normally, no matter what the initial distribution was. This is the reason why normally distributed random variables appear so often on practice.

Toto je snad jedno z nejdůležitějších rozdělení vzhledem k centrální limitní větě – jednomu ze základních výsledků pravděpodobnosti a statistiky. Zhruba tato věta říká, že za určitých přirozených podmínek, průměr z velkého počtu stejně rozdělených náhodných proměnných má normální rozdělení, bez ohledu na to, jaké bylo původní rozdělení. To je důvod, proč se normálně rozdělené náhodné proměnné objevují v praxi tak často.

The expected value of a normally distributed random variable is equal to

Očekáváná hodnota normálně rozdělené náhodné proměnné je rovna

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} te^{-\frac{1}{2}(\frac{t-m}{\sigma})^2} \,\mathrm{d}t = m,$$

and the standard deviation is equal to σ.

a směrodatná odchylka je rovna σ.

The exponential distribution describes the time between events in a process in which events occur continuously and independently at a constant average rate $\lambda > 0$. It is defined as the continuous distribution with the density function of the form

$$f_\lambda(x) = \begin{cases} 0 & \text{if } x < 0, \\ \lambda e^{-\lambda x} & \text{if } x \geq 0. \end{cases}$$

The expected value of an exponentially distributed random variable is equal to

$$\lambda \int_0^\infty t e^{-\lambda t} \, \mathrm{d}t = \frac{1}{\lambda},$$

and the standard deviation is equal to $\frac{1}{\lambda}$ too.

Exponenciální rozdělení popisuje časy mezi událostmi v procesu, ve kterém se události objevují spojitě a nezávisle v konstantní průměrné míře $\lambda > 0$. Je definováno jako spojité rozdělení s hustotou pravděpodobnosti ve tvaru

$$f_\lambda(x) = \begin{cases} 0 & \text{když } x < 0, \\ \lambda e^{-\lambda x} & \text{když } x \geq 0. \end{cases}$$

Očekávaná hodnota exponenciálně rozdělené náhodné proměnné se rovná

$$\lambda \int_0^\infty t e^{-\lambda t} \, \mathrm{d}t = \frac{1}{\lambda},$$

a směrodatná odchylka se rovněž rovná $\frac{1}{\lambda}$.

Population, Its Numerical Characteristics

A statistical population is a set of similar items or events which is of interest for some question or experiment. Examples of populations: human population of a given region or country, harvest of given crop in a given region, traffic going through a given transport hub, financial transactions performed on a given stock exchange during a given timeframe, stars in a galaxy, particles in a suspension, etc. In each concrete situations, we are interested in certain numerical data attached to members of the population, such as height and weight of humans, amount of harvest, amount of traffic, brightness of stars, size of particles, etc. A common aim of statistics is to produce information about such data in a chosen population.

10 Populace, její číselné charakteristiky

Populace je množina podobných předmětů nebo událostí, která jsou zajímavá ve vztahu k nějaké otázce nebo experimentu. Příklady populací: lidská populace daného regionu nebo země, sklizeň nějaké plodiny v daném regionu, dopravní provoz procházející určitým dopravním uzlem, finanční transakce prováděné na dané burze během určitého časového rámce, hvězdy v galaxii, částice v suspenzi, atd. V každé konkrétní situaci se zajímáme o určité číselné údaje týkající se členů populace, jako jsou výška a hmotnost člověka, množství sklizně, intenzita provozu, jas hvězd, velikost částic, atd. Obecným cílem statistiky je předložit informace o těchto datech ve vybrané populaci.

One can think about population and the data attached to it as a "real-word" implementation of the abstract notions of sample space and discrete random variable. As such, one can assign to them the same numerical characteristics as to discrete distributions: the population mean μ, and the population standard deviation σ. If the population of size n is represented by numerical values $(x_1, \ldots, x_n)$, the values for μ and σ are defined by the same formulas as on page 31. Expanding the squares in the formula for the standard deviation, we get an alternative expression

O populaci a datech k ní připojených můžeme uvažovat jako o implementaci abstraktních termínů prostoru elementárních jevů a diskrétní náhodné proměnné v „reálném světě". Jako takovým jim můžeme přiřadit stejné číselné charakteristiky jako diskrétním rozdělením: populační průměr μ, a směrodatnou odchylku σ. V případě, že je populace velikosti n zastoupena číselnými hodnotami $(x_1, \ldots, x_n)$, pak hodnoty μ a σ jsou definovány stejnými vzorci, které jsou na straně 31. Rozepsáním druhé mocniny ve vzorci pro směrodatnou odchylku dostaneme alternativní vyjádření

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2}.$$

Sometimes one considers also the population total τ, which is merely the sum of all numerical values in question:

Někdy uvažujeme rovněž o populačním úhrnu τ, což je pouze součet všech příslušných číselných hodnot:

$$\tau = \sum_{i=1}^{n} x_i,$$

and the population variance, which is the square of the population standard deviation.

a populačním rozptylu, což je druhá mocnina populační směrodatné odchylky.

Example 16 Let the population consist of n flippings of a coin, with head and tail outcomes encoded by 0 and 1 respectively, and assume that head occurs with probability p (so, if $p \neq \frac{1}{2}$, the coin is biased). For big values of n, the population mean is very close to p, and the population standard deviation is very close to

Příklad 16 Nechť se populace skládá z n hodů mincí s lícem a rubem zakódovaným pomocí 0 a 1, ve stejném pořadí, a předpokládáme, že líc se objevuje s pravděpodobností p (pokud $p \neq \frac{1}{2}$ pak je mince nevývážená). Pak pro velké hodnoty n je populační průměr velmi blízko p a populační směrodatná odchylka je velmi blízko

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2} = \sqrt{p - p^2} = \sqrt{p(1-p)}.$$

Sample, Random Sampling

On practice, when collecting statistical data, in most of the cases it is unfeasible to collect it for every member of the given population. For example, when conducting an opinion poll, it is impossible to contact every person living in the country; when performing a financial audit of a large company, it is impractical to look at every financial transaction or record; when performing a destructive test of the output of a manufacturing process, one wants to destroy as little items as possible. That is why it is important to choose a relatively small subset of a given population, a sample (or data sample, or dataset), which should be representative enough to make a justified extrapolation on the whole population, as far as the data we are interested in are concerned. This is achieved via survey sampling.

11 Výběr, náhodný výběr

V praxi, při sběru statistických dat, je ve většině případů neproveditelné sesbírat data pro každého člena dané populace. Například při provádění ankety, je nemožné kontaktovat každou osobu žijící ve státu. Při provádění finančního auditu velké společnosti je nepraktické se dívat na každou finanční transakci nebo záznam. Při provádění destrukčního testu produkce z výrobního procesu, chceme zničit tak málo položek, jak je to jen možné. To je důvod, proč je důležité zvolit relativně malou podmnožinu dané populace – výběr (nebo vzorek dat nebo datový soubor), který by měl být dostatečně reprezentativní, abychom provedli ospravedlněnou extrapolaci na celé populaci, do takové míry, jak nám data, o která se zajímáme, dovolí. Toho je dosaženo pomocí anketního výběru.

The most standard and widely accepted procedure to perform a survey sampling is random (or probability) sampling. Under random sampling, members of the population are selected according to certain probabilistic criteria; that allows to give an estimate of sampling error.

Example 17 If one needs to select one person from each given group (e.g., each household in a given area, or each class in a university), one may assign to each member of the group a random number from the uniform distribution between 0 and 1, and select the person assigned the highest number. Or, in a simple random sample with replacement, we may randomly select a unit out of 100,000 units (the size of the population), and repeat the procedure 100 times (the size of the sample). Or, when conducting a poll about customers' buying habits, on may choose every 15th visitor of the supermarket within a given timeframe.

Nejstandardnější a široce přijatá procedura pro anketní výběr je náhodný (nebo pravděpodobnostní) výběr. U náhodného výběru jsou členové populace vybráni podle určitého pravděpodobnostního kritéria, které nám dovoluje odhadnout výběrovou chybu.

Příklad 17 Jestliže je třeba vybrat jednu osobu z každé skupiny (např. každá domácnost v dané oblasti, nebo každá třída na univerzitě), můžeme přiřadit každýmu členu skupiny náhodné číslo z rovnoměrného rozdělení mezi 0 a 1 a zvolit osobu, které jsme přiřadili nejvyšší číslo. Nebo v jednoduchém náhodném výběru s opakováním, můžeme náhodně vybrat jednotku ze 100,000 jednotek (velikost populace), a opakovat postup 100 krát (velikost vzorku). Nebo pokud provádíme průzkum o nákupní zvyklosti zákazníků, můžeme si vybrat každého 15. návštěvníka supermarketu v daném časovém rámci.

Generally, a random sample is defined as a collection of random variables which have the same probability distribution and are mutually independent. The distribution in question can be any – normal, uniform, exponential, Poisson, or even a distribution not described by an explicitly known formula; it is only important that each random variable has the same distribution. It is the task of inferential statistics to determine which theoretical statistical distribution (sometimes called a model distribution) is most suitable to describe the empirical data, and to determine the distribution parameters (so-called parameters estimation). This can be achieved with a variety of methods, ranging from simple ones to pretty much sophisticated.

Obecně je náhodný výběr definován jako soubor náhodných proměnných, které mají stejné pravděpodobnostní rozdělení a jsou vzájemně nezávislé. Příslušné rozdělení může být libovolné – normální, rovnoměrné, exponenciální, Poissonovo, nebo i rozdělení, které není popsáno pomocí explicitně známého vzorce. Je pouze důležité, aby každá náhodná proměnná měla stejné rozdělení. Je úkolem inferenční statistiky určit, které teoretické statistické rozdělení (někdy nazýváno model distribuce) je nejvhodnější pro popsání empirických dat, a určení parametrů distribuce (tzv. odhad parametrů). Toho může být dosaženo různými metodami, od jednoduchých až po poměrně hodně propracované.

The most primitive "visual" method may run as follows. We plot the histogram of our data, and try to guess the distribution by the histogram shape; due to the central limit theorem, in many cases the normal distribution would be a good guess. Then we may estimate the normal distribution parameters as the population mean and variance, plot the graph of the so obtained normal distribution, and assess visually how good it matches the histogram. Alternatively, we may play with the parameters of the chosen distribution – a normal or another one – and assess visually which set of parameters produces a curve which fits the empirical data best.

Nejprimitivnější „vizuální" metoda může být provedena následujícím způsobem. Zobrazíme graf histogramu z našich dat a pokusíme se odhadnout rozdělení podle tvaru histogramu. Kvůli centrální limitní větě bude v mnoha případech normální rozdělení dobrým odhadem. Pak můžeme odhadnout parametry normálního rozdělení jako jsou populační průměr a rozptyl. Vykreslíme graf takto získaného normálního rozdělení, a posoudíme vizuálně, jak dobře graf odpovídá histogramu. Případně si můžeme hrát s parametry zvoleného rozdělení – normálního či jiného – a posoudit vizuálně, které nastavení parametrů vytváří křivku, která se hodí empirickým datům nejlépe.

In more sophisticated methods, we may need to compute other characteristics of our empirical datasets: for example, various quantiles; or relative frequencies, i.e. the quantities

Ve více propracovaných metodách budeme potřebovat spočítat další charakteristiky našich empirických dat: například různé kvantily, nebo relativní četnosti, t.j veličiny

$$\frac{|\{i \mid X_i = a\}|}{n}$$

for different values of a, which approximate the probability mass function; or higher moments, i.e. the sums of the form

pro různé hodnoty a, které aproximují pravděpodobnostní funkci; nebo vyšší momenty, t.j. součty ve tvaru

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k;$$

or to take into account skewness (i.e., a possible non-symmetricity around the mean) of the data.

nebo vzít v potaz vychýlenost (t.j. možná nesymetričnost okolo průměru) dat.

Hypothesis Testing, Null and Alternative Hypotheses

The more sophisticated methods of choosing a suitable distribution and estimation of the distribution parameters are usually performed in the framework of hypothesis testing. Hypothesis testing is also used for establishing a relationship (or lack thereof) between two datasets, or, more generally, in deriving any kind of statistical observation about one or more datasets.

12 Testování hypotéz, nulové a alternativní hypotézy

Propracovanější metody výběru vhodného rozdělení a odhadu parametrů rozdělení jsou běžně prováděny v rámci testování hypotéz. Testování hypotéz je rovněž používáno pro stanovení vztahu (nebo jeho zamítnutí) mezi dvěmi datovými množinami, nebo obecněji v odvozování jakéhokoli statistického pozorování o jedněch nebo více datech.

Usually, this is done by specifying two rival and mutually exclusive hypotheses, the null and alternative hypotheses, and their subsequent comparison by certain statistical procedures. There is no rule of thumb how null and alternative hypotheses should be formed. However, the usual statistical practice stipulates that the null hypothesis states that the phenomenon being studied produces no effect or makes no difference. The null hypothesis is also usually the hypothesis one wants to reject, or "nullify". For example, when investigating relationship between two datasets, the null hypothesis should state that there is no relationship at all, while the alternative hypothesis should indicate the existence of such relationship. Or, say, when investigating the impact of smoking on lung cancer, the null hypothesis would state that smoking does not have any impact.

Obvykle se to provádí určením dvou konkurujících a vzájemně se vylučujících hypotéz, nulové a alternativní hypotézy, a jejich následného porovnání pomocí nějakých statistických postupů. Neexistuje žádné pravidlo, jak by měla být nulová a alternativní hypotéza vytvořena. Nicméně, obvyklá statistická praxe vyžaduje, aby nulová hypotéza uváděla, že studovaný jev nemá žádný vliv ani nehraje žádnou roli. Nulová hypotéza je obvykle taková hypotéza, kterou chceme zamítnout nebo „vynulovat". Například při zkoumání vztahu mezi dvěma soubory dat, nulová hypotéza by měla uvést, že neexistuje vůbec žádný vztah. Zatímco alternativní hypotéza by měla naznačovat existenci takového vztahu. Nebo, řekněme, že když zkoumáme vliv kouření na rakovinu plic, nulová hypotéza by měla konstatovat, že kouření nemá žádný vliv.

A type I error is the (incorrect) rejection of a true null hypothesis. The probability of type I error is called the significance level of a test. A type II error is the (incorrect) acceptance of a false null hypothesis. The probability of not making a type II error, i.e. the (correct) rejection of a false null hypothesis, is called the power of a test. The probabilities of making type I and type II errors are traded off against each other: for any given sample set, the effort to reduce one type of error generally results in increasing the other type of error. For a given test, the only way to reduce both error rates is to increase the sample size, and this may not be feasible.

One of the most used test statistics in hypothesis testing is p-value, which is defined as the probability of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis H_0 is true. What is "more extreme" and how it is measured, depends on the context.

Chyba typu I je (nesprávné) zamítnutí pravdivé nulové hypotézy. Pravděpodobnost chyby typu I se nazývá hladina významnosti testu. Chyba typu II je (nesprávné) přijetí nepravdivé nulové hypotézy. Pravděpodobnost, že neuděláme chybu typu II, to jest (správně) zamítneme nepravdivou nulovou hypotézu, se nazývá síla testu. Pravděpodobnosti provedení chyby typu I a chyby typu II jsou převáděny mezi sebou: pro libovolný výběr dat má snaha snížit jeden typ chyby obecně za následek zvýšení chyby druhé. Pro daný test je jediným způsobem, jak snížit obě chyby, zvýšení velikosti vzorku, a to nemusí být proveditelné.

Jednou z nejčastěji používaných testovacích statistik v testování hypotéz je p-hodnota, která je definována jako pravděpodobnost získání výsledku shodného nebo „více extrémního“ tomu, co bylo ve skutečnosti pozorováno, když je nulová hypotéza H_0 pravdivá. Co je „extrémnější“ a jak je měřena, závisí na kontextu.

Example 18 For a "double-tailed" event, the p-value of a random variable X assuming values "more extreme" than x (the observed value), might be defined as

$$2 \cdot \min\{ \Pr(X \geq x \mid H_0), \Pr(X \leq x \mid H_0) \},$$

while for "left-tailed" events the same value might be defined as

$$\Pr(X \leq x \mid H_0),$$

and similarly for "right-tailed" ones.

The p-value measures statistical significance of the test, but it should not be confused with the probability of the hypothesis being true, the probability of observing the given data, etc. p-values are often misused and misinterpreted.

Příklad 18 Pro „dvoustranné" události může být p-hodnota náhodné proměnné X, za předpokladu „extrémnějších" hodnot než x (pozorovaná hodnota), definována jako

$$2 \cdot \min\{ \Pr(X \geq x \mid H_0), \Pr(X \leq x \mid H_0) \},$$

zatímco pro „levostranné" události může být stejná hodnota definována jako

$$\Pr(X \leq x \mid H_0),$$

a obdobně pro „pravostranné".

p-hodnota měří statistickou významnost testu, ale neměla by být zaměňována s pravděpodobností, že hypotéza je pravdivá, pravděpodobností pozorování daných dat, atd. p-hodnoty bývají často zneužívány a nesprávně interpretovány.

Example 19 In the flipping coin example, suppose that the null hypothesis specifies that the coin is fair. In a double-tailed model, the alternative hypothesis would be that the coin is biased either way, while in an one-tailed model the alternative hypothesis says that the coin is biased towards, say, heads. Suppose that one gets 5 heads in a row in one experiment. In the one-tailed model this is the most extreme possible value, with a p-value equal to

$$\left(\frac{1}{2}\right)^5 = \frac{1}{32} \approx 0.03\,.$$

In the double-tailed model, the corresponding p-value would be twice as that:

$$2 \cdot \left(\frac{1}{2}\right)^5 = \frac{1}{16} \approx 0.06\,.$$

One frequently sets 0.05 as the threshold for the p-value of the test to be statistically significant. Under this assumption, we should reject the null hypothesis in the first case, while we cannot do that in the second one.

Příklad 19 Předpokládejme, že v příkladu házení mincí nulová hypotéza tvrdí, že mince je vyvážená. V oboustranném modelu by alternativní hypotéza byla, že mince je nevyvážená v kterémkoli směru, zatímco v jednostranném modelu alternativní hypotéza říká, že mince je nakloněna směrem, řekněme, lícům. Předpokládejme, že hodíme 5 líců v řadě v jednom experimentu. V jednostranném modelu se jedná o nejextrémnější možnou hodnota s p-hodnotou rovnající se

V dvoustranném modelu by byla odpovídající p-hodnota dvakrát větší:

Často se nastavuje 0,05, jako mezní hodnota pro p-hodnotu testu, pro to aby byla statisticky významná. Za takového předpokladu bychom měli zamítnout nulovou hypotézu v prvním případě, zatímco v druhém případě to nemůžeme udělat.

An essentially equivalent procedure, but not using the concept of a p-value, would run as follows:

1. Choose a test statistics (for example, just the number of heads in the flipping coin example);

2. Derive the distribution of the test statistics under the null hypothesis (the binomial distribution on our example);

3. Select the significance level of the test (the common values are 0.05 and 0.01);

4. Determine the critical (or rejection) region – the values of the test statistics for which the null hypothesis is rejected;

5. Perform the test, derive from it the empirical value of test statistics, and see whether it falls into the critical region or not.

V principu ekvivalentní postup, který nepoužívá konceptu p-hodnoty, by byl následující:

1. Vyberte testovací statistiku (například pouze počet líců v příkladu házení mincí);

2. Odvoďte rozdělení testovací statistiky za předpokladu nulové hypotézy (například binomické rozdělení);

3. Určete úroveň signifikance (významnosti) testu (běžné hodnoty jsou 0,05 a 0,01);

4. Určete kritický (nebo zamítací) obor hodnot – hodnoty testovací statistiky, pro které je nulová hypotéza zamítnuta;

5. Proveďte test, vypočítejte z empirických hodnot testovací statistiku a podívejte se, jestli padla do oboru kritických hodnot či nikoli.

From these examples and descriptions it should be clear that the matter of rejecting the null hypothesis is highly sensitive to a big number of factors, some of them of a highly subjective character: the choice of the null and alternative hypotheses themselves, initial assumptions about the form of statistical distribution, the choice of significance level, the sample size, etc. Any statistical claim without mentioning all these factors should be taken with a big grain of salt. See, for example, a curious (but edifying) Journal of Articles in Support of the Null Hypothesis, where such practices are often (and justly) criticized.

Z těchto příkladů a popisu by mělo být zřejmé, že záležitost zamítnutí nulové hypotézy, je vysoce citlivá na velké množství faktorů. Některé z nich jsou vysoce subjektivního charakteru: samotná volba nulové a alternativní hypotézy, počáteční předpoklady o tvaru statistického rozdělení, volba hladiny významnosti, velikost výběru, atd. Každé statistické tvrzení bez uvedení všech těchto faktorů by mělo být přijato s velkou dávkou skepse. Podívejte se například na zvláštní (ale poučný) Journal of Articles in Support of the Null Hypothesis (Časopis článků pro podporu nulové hypotézy), kde takové postupy jsou často (a oprávněně) kritizovány.

There is a direct relationship between the critical region and the confidence interval. The confidence interval of a certain statistical parameter is the interval, calculated from the sample, that contains the specified value of the parameter with the specified probability. A typical situation when this notion occurs naturally is estimation of the average of the mean of identically distributed random variables. For example, if a certain random variable is normally distributed with the same mean μ and standard deviation σ, then it is known that the average $\overline{x}$ of n observations is normally distributed around μ with standard deviation $\frac{\sigma}{\sqrt{n}}$. A 95% confidence interval for μ is determined then as

Existuje přímá souvislost mezi oborem kritických hodnot a konfidenčním intervalem. Konfidenční interval určitého statistického parametru je interval, vypočítaný z výběru, který obsahuje zadanou hodnotu parametru s určitou pravděpodobností. Typickou situací, kdy se tento pojem vyskytuje přirozeně, je odhad průměrné střední hodnoty stejně rozdělené náhodné proměnné. Například, pokud určitá náhodná proměnná má normální rozdělení se stejným průměrem μ a směrodatnou odchylkou σ, pak je známo, že průměr $\overline{x}$ o n pozorováních má normální rozdělení se středem μ a směrodatnou odchylkou $\frac{\sigma}{\sqrt{n}}$. A 95% konfidenční interval μ je určený jako jako

$$\overline{x} + N_{0.025}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + N_{0.975}\frac{\sigma}{\sqrt{n}},$$

where $N_{0.975} \approx 1.96$ and $N_{0.025} = -N_{0.975}$ are the 97.5% and 2.5% quantiles in the standard (i.e. with parameters $\mu = 0$ and $\sigma = 1$) normal distribution.

kde $N_{0,975} \approx 1{,}96$ a $N_{0,025} = -N_{0,975}$ jsou 97,5% a 2,5% kvantily standardního (tj. s parametry $\mu = 0$ a $\sigma = 1$) normálního rozdělení.

Now, suppose that for some parameter θ and its value θ_0, we test the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$ (one-tailed test). Then we reject H_0 in favor of H_1 (i.e., θ is not in the critical region) at the significance level α if and only if θ_0 is not in the $100(1-\alpha)\%$ one-tailed confidence interval for θ. A similar statement is true in the case of double-tailed test.

Nyní předpokládejme, že pro nějaký parametr θ a jeho hodnotu θ_0 testujeme nulovou hypotézu $H_0 : \theta = \theta_0$ oproti alternativní hypotéze $H_1 : \theta > \theta_0$ (jednostranný test). Potom zamítáme H_0 ve prospěch H_1 (t.j. θ není v oboru kritických hodnot) na úrovni α pravě tehdy když θ_0 není v $100(1-\alpha)\%$ jednostranném konfidenčním intervalu pro θ. Podobný výrok je pravdivý v případě dvoustranného testu.

Sources 13 Zdroje

In compiling this material, we have used the following sources liberally.

Při přípravě tohoto materiálu byly volně použity následující zdroje.

▷ P. Dalgaard, Introductory Statistics with R, 2nd ed., Springer, 2008.

▷ F.M. Dekking, C. Kraaikamp, H.P. Lopuhaä, and L.E. Meester, A Modern Introduction to Probability and Statistics, Springer, 2005.

▷ Yu.I. Manin, Mathematical knowledge: internal, social and cultural aspects, in: Mathematics as Metaphor. Selected Essays of Yuri I. Manin, AMS, 2007, 3–26.

▷ J.A. Rice, Mathematical Statistics and Data Analysis, 3rd ed., Thomson Brooks/Cole, 2007.

and also (English) Wikipedia articles:

a také články v (anglické) Wikipedii:

- ▷ Alternative hypothesis
- ▷ Binomial distribution
- ▷ Discrete uniform distribution
- ▷ Combination
- ▷ Confidence interval
- ▷ Exponential distribution
- ▷ Hypergeometric distribution
- ▷ Normal distribution
- ▷ Null hypothesis
- ▷ p-value
- ▷ Permutation
- ▷ Poisson distribution
- ▷ Probability distribution
- ▷ Probability theory
- ▷ Sample (statistics)
- ▷ Sampling (statistics)
- ▷ Simple random sample
- ▷ Statistical hypothesis testing
- ▷ Statistical population
- ▷ Statistical power
- ▷ Statistics
- ▷ Survey sampling
- ▷ Systematic sampling
- ▷ Type I and type II errors