

7.3.2 Teoretická rozdělení

Nespojitá (diskrétní) náhodná veličina nabývá pouze konečně mnoha různých hodnot znaku, jinak mluvíme o spojité náhodné veličině.

Pokud budeme u spojité náhodné veličiny třídit hodnoty do stále užších a užších tříd, dostaneme histogramy relativních četností, které se budou stále více blížit hladké spojité křivce, kterou nazýváme frekvenční funkce (hustota pravděpodobnosti) a označujeme ji $f(x)$.

Podobně u kumulativních relativních četností dospějeme ke spojité křivce, kterou značíme $F(x)$ a nazýváme distribuční funkce.

Frekvenční i distribuční funkce spojitých náhodných veličin mají některé důležité vlastnosti:

- obě funkce jsou nezáporné
- frekvenční funkce omezuje nad vodorovnou osou plochu o velikosti 1
- distribuční funkce je na intervalu $\langle 0,1 \rangle$ neklesající
- hodnota distribuční funkce v bodě x_j , tedy $F(x_j)$, představuje plochu, kterou ohraničuje frekvenční funkce na vodorovnou osu od svého začátku do bodu x_j

Poznámka:

V teoretických rozděleních odlišíme aritmetický průměr a směrodatnou odchylku od charakteristik konkrétních souborů změnou symboliky a zavedeme značení μ a σ .

7.3.2.1 Normální rozdělení (Gaussovo rozdělení)

Frekvenční funkce normálního rozdělení

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Parametry μ a σ jsou pro daný soubor konstantní, ale liší se soubor od souboru, proto byla zavedena normovaná náhodná veličina, definovaná vztahem $u = \frac{x-\mu}{\sigma}$. Jejím zavedením získáme frekvenční funkci normovaného normálního rozdělení

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

Porovnáním obou rovnic zjišťujeme, že pro normované normální rozdělení platí $\mu = 0$ a $\sigma = 1$.

Vlastnosti normálního rozdělení:

- Souřadnice vrcholu je $x = \mu$ respektive $u = 0$
- Zvonovitý tvar křivky rozdělení s asymptotou x , znamená, že náhodná veličina může nabývat libovolné hodnoty
 $\mu \pm \sigma$ $p = 0,6827$
- $\mu \pm 2\sigma$ znamená $p = 0,9545$
 $\mu \pm 3\sigma$ $p = 0,9973$
- respektive $p = 0,95$ znamená $\mu \pm 1,95996\sigma \doteq \mu \pm 1,96\sigma$
 $p = 0,99$ znamená $\mu \pm 2,57582\sigma \doteq \mu \pm 2,58\sigma$
- pro hodnoty frekvenční funkce platí $f(u) = f(-u)$

Toto rozdělení je nejdůležitějším teoretickým rozdělením.

7.3.2.2 Binomické rozdělení

V situaci, kdy náhodná veličina nabývá dvou hodnot znaku s pravděpodobnostmi po řadě p, q a při n pokusech doufáme v příznivý výsledek právě x – krát, platí

$$f(x) = \binom{n}{x} p^x q^{n-x}$$

což představuje rozdělení pravděpodobností tzv. binomického rozdělení. Hovoříme o rozdělení pravděpodobností místo o frekvenční funkci, protože náhodná veličina nabývá u binomického rozdělení pouze diskrétních hodnot od 0 do n .

Pro aritmetický průměr a směrodatnou odchylku platí

$$\mu = np \text{ a } \sigma = \sqrt{npq}$$

7.3.2.3 Poissonovo rozdělení

Použití binomického rozdělení pro velká n je velmi pracné, proto za podmínek

1) velké n ($n > 30$)

2) malé p ($p \leq 0,1 \vee p \geq 0,9$) uvažujeme rozdělení pravděpodobností, pro které platí

$$f(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!},$$

kde pro aritmetický průměr a rozptyl je $\mu = \sigma^2 = \lambda$.

Poznámka

Někdy je potřeba ze základního souboru provádět náhodné výběry a na jejich základě pak usuzovat na vlastnosti základního souboru.

Protože náhodný výběr představuje konkrétní soubor o rozsahu n výběrové charakteristiky pak budou \bar{x}, s^2, s , zatímco základní soubor o rozsahu N má parametry μ, σ^2, σ .

Pro nekonečně rozsáhlý základní soubor s parametry μ, σ lze odvodit, že rozdělení výběrových průměrů lze považovat za normální pro které platí

$$\mu_{\bar{x}} = \mu \text{ a } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}. \text{ Toto platí přibližně i pro základní soubory, které mají normální}$$

rozdělení, a to při výběrech s dostatečně velkým rozsahem (cca $n \geq 20$).

Pro práci s aritmetickými průměry a směrodatnými odchylkami v následující kapitole jsou podstatné (a nejen tam) ještě další dvě teoretická rozdělení..

7.2.3.2.4 Rozdělení χ^2

Ze základního souboru s normovaným normálním rozdělením provedeme náhodný výběr n prvků, které označíme u_1, u_2, \dots, u_n . Pro které označíme součet druhých mocnin symbolem χ^2

(čteme „chí kvadrát“) $\chi^2 = \sum_{i=1}^n u_i^2$. Hodnota této funkce je náhodnou veličinou s vlastním

rozdělením, frekvenční funkcí $f_v(\chi^2)$ a distribuční funkcí $F_v(\chi^2)$, kde v obou případech je jediným parametrem funkcí veličina v nazývaná počet stupňů volnosti, která je v tomto případě rovna rozsahu náhodného výběru n . S jeho rostoucí hodnotou se rozdělení χ^2 blíží rozdělení normálnímu.

Poznámka

$$f_v(\chi^2) = f(w) = \frac{1}{\Gamma\left(\frac{n}{2}\right) \cdot 2^{\frac{n}{2}}} \cdot e^{-\frac{w}{2}} \cdot w^{\frac{n}{2}-1}, \text{ kde } \Gamma(\alpha) \text{ je Eulerova Gama funkce definovaná}$$

jako rozšíření faktoriálu nezáporných celých čísel takto: $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$,

přičemž se dá ukázat, že pro přirozená čísla k platí: $\Gamma(k+1) = k!$.

Pro kritické hodnoty χ_p^2 rozdělení χ^2 platí, že je náhodná veličina překročí s pravděpodobností p . Připadají v úvahu tři možné případy a totiž, že plocha leží mezi

$$\text{a) } \chi_0^2 = 0 \text{ a } \chi_{0,05}^2 \quad \text{b) } \chi_{0,95}^2 \text{ a } \chi^2 = \infty \quad \text{c) } \chi_{0,975}^2 \text{ a } \chi_{0,025}^2$$

7.2.3.2.5 Studentovo rozdělení (rozdělení t)

Pro hodnocení odchylek výběrových průměrů \bar{x} od průměru μ základního souboru

s normálním rozdělením je definována náhodná veličina $t = \frac{\bar{x} - \mu}{s} \sqrt{n-1}$, které přísluší

uvedené rozdělení. Frekvenční funkce t rozdělení je souměrná podle osy jdoucí jejím vrcholem, značí se $g_v(t)$, jež má opět jediným parametr, který je tentokrát $\nu = n-1$.

Poznámka:

$$g_v(t) = \frac{1}{\sqrt{\nu} B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left(1 + \frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}},$$

$$\text{kde } B(a, b) \text{ je Eulerova Beta funkce } B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Studentovo rozdělení přechází v normální pro $\nu = \infty$, což v praxi nahrazujeme $\nu \geq 30$. Stejně jako u rozdělení χ^2 hledáme hodnoty, kterými je ohraničena plocha o velikosti p , ovšem většinou symetricky vzhledem k $t = 0$, takže kritické hodnoty $t_p, -t_p$ určují interval hodnot, kterých je nabýváno s pravděpodobností p .

Kritické hodnoty t-rozdělení

ný--p	0,05	0,01	ný--p	0,05	0,01
1	12,706	63,657	18	2,101	2,878
2	4,303	9,925	19	2,093	2,861
3	3,182	5,841	20	2,086	2,845
4	2,776	4,604	21	2,080	2,831
5	2,571	4,032	22	2,074	2,819
6	2,447	3,707	23	2,069	2,807
7	2,365	3,499	24	2,064	2,797
8	2,306	3,355	25	2,06	2,787
9	2,262	3,25	26	2,056	2,779

10	2,228	3,169	27	2,052	2,771
11	2,201	3,106	28	2,048	2,763
12	2,179	3,055	29	2,045	2,756
13	2,16	3,012	30	2,042	2,75
14	2,145	2,977	40	2,021	2,704
15	2,131	2,947	60	2	2,66
16	2,12	2,921	120	1,98	2,617
17	2,11	2,898	∞	1,96	2,576

7.3.3 Odhady parametrů a intervaly spolehlivosti

Odhad $\hat{\mu}$ aritmetického průměru základního souboru je roven výběrovému aritmetickému průměru \bar{x} , tedy $\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Odhad $\hat{\sigma}$ směrodatné odchylky základního souboru je větší než výběrová směrodatná odchylka a je dán vztahem $\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$.

Lze odvodit, že $\frac{\hat{\sigma}}{s} = \frac{\sqrt{\frac{1}{n-1}}}{\sqrt{\frac{1}{n}}}$ z čehož plyne $\hat{\sigma} = s \cdot \sqrt{\frac{n}{n-1}}$. Protože tohoto odhadu používáme,

když neznáme skutečnou hodnotu směrodatné odchylky σ zřejmě $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n-1}}$.

Interval spolehlivosti pro parametr μ

Pokud rozdělení výběrových průměrů je normální a má průměr $\mu_{\bar{x}} = \mu$ a směrodatnou

odchylku $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. To znamená, že interval $\mu_{\bar{x}} \pm 1,96\sigma_{\bar{x}}$ zahrne 95% všech výběrových průměrů. Lze psát, že $\mu_{\bar{x}} - 1,96\sigma_{\bar{x}} \leq \bar{x} \leq \mu_{\bar{x}} + 1,96\sigma_{\bar{x}}$.

Použitý součinitel je hodnotou normované náhodné veličiny u , kterou budeme nadále označovat symbolem u_p . Zřejmě $u_{0,05} = 1,96$, proto bude náhodná veličina u s pravděpodobností 0,025 menší než -1,96 a s touž pravděpodobností větší než 1,96. Poznamenejme ještě, že $u_{0,01} = 2,58$.

Zřejmě

$$\mu_{\bar{x}} - u_p \sigma_{\bar{x}} \leq \bar{x} \leq \mu_{\bar{x}} + u_p \sigma_{\bar{x}}$$

$$\mu_{\bar{x}} - u_p \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_{\bar{x}} + u_p \frac{\sigma}{\sqrt{n}} \text{ z čehož dostáváme } \bar{x} - u_p \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + u_p \frac{\sigma}{\sqrt{n}}.$$

Protože hodnotu σ zpravidla neznáme, nahrazujeme ji jejím odhadem

$$\bar{x} - u_p \frac{s}{\sqrt{n-1}} \leq \mu \leq \bar{x} + u_p \frac{s}{\sqrt{n-1}} .$$

Pro výběry s malým rozsahem musíme nahradit hodnotu u_p hodnotou t_p , takže

$$\bar{x} - t_p \frac{s}{\sqrt{n-1}} \leq \mu \leq \bar{x} + t_p \frac{s}{\sqrt{n-1}} ,$$

kde kritické hodnoty stanovujeme pro $\nu = n - 1$ stupňů volnosti.

7.3.4 Testy významnosti

Zjišťujeme:

- a) zda zkoumaný výběr pochází ze základního souboru s určitým rozdělením
- b) zda dva výběry pochází z jednoho a téhož základního souboru
- c) zda lze pohlížet na odlehlou hodnotu jako na důsledek hrubé chyby při „měření“
- d) zda je možno považovat daný soubor za náhodně uspořádaný atd.

Obecný postup

- a) volba hladiny významnosti
- b) formulace nulové hypotézy
- c) volba testovacího kritéria (testy parametrické a neparametrické)
- d) interpretace výsledku

add c)

Po výpočtu příslušných hodnot testovacího kritéria je porovnáme s příslušnou kritickou hodnotou odpovídající náhodné veličiny. Poté mohou nastat dvě možnosti:

I) Kritická hodnota je menší než vypočítaná hodnota testovacího kritéria. To znamená, že nastává případ, který jsme za předpokladu, že je nulová hypotéza správná, očekávali s tak malou pravděpodobností (obvykle 0,05 nebo 0,01), že jej můžeme považovat takřka za nemožný. Soudíme proto, že testovaná odchylka nemá náhodný charakter a je způsobena účinkem zkoumaných faktorů. Proto **nulovou hypotézu zamítáme** na zvolené hladině významnosti a tvrdíme, že sledovaný rozdíl **je statisticky významný**.

II) Kritická hodnota je větší než vypočítaná hodnota testovacího kritéria. Tedy jde o případ, který jsme očekávali, za předpokladu, že nulová hypotéza je správná, s tak velkou pravděpodobností, že jeho výskyt můžeme považovat takřka za jistý. Předpokládáme proto, že testovaná odchylka má charakter náhodný, z čehož usuzujeme, že účinek studovaných faktorů se neuplatňuje. Proto **nulovou hypotézu nezamítáme** na zvolené hladině významnosti a tvrdíme, že sledovaný rozdíl **není statisticky významný**.

Používání testovacích kritérií pro testy rozdílu a testy shody přesahuje rámec základního seznámení s problematikou.

Test extrémních odchylek

Za předpokladu normality základního souboru.

Grubbsův test

Je parametrickým testem. Určíme výběrový průměr \bar{x} a výběrovou směrodatnou odchylku s a uspořádáme v souboru hodnoty znaku $x_1 \leq x_2 \leq \dots \leq x_n$.

Za testovací kritérium používáme vztahů $T_n = \frac{x_n - \bar{x}}{s}$ nebo $T_1 = \frac{\bar{x} - x_1}{s}$ podle toho,

zda testujeme největší nebo nejmenší hodnotu.

Kritické hodnoty $T_{1:p} = T_{n:p}$ jsou uvedeny v tabulce.

Kritické hodnoty pro Grubbsův test

n--p	0,05	0,01	n--p	0,05	0,01
3	1,412	1,414	15	2,493	2,8
4	1,689	1,723	16	2,523	2,837
5	1,869	1,955	17	2,551	2,871
6	1,996	2,13	18	2,577	2,903
7	2,093	2,265	19	2,6	2,932
8	2,172	2,374	20	2,623	2,959
9	2,237	2,464	21	2,644	2,984
10	2,294	2,54	22	2,664	3,008
11	2,343	2,606	23	2,683	3,03
12	2,387	2,663	24	2,701	3,051
13	2,426	2,714	25	2,717	3,071
14	2,461	2,759			

Dixonův test

Je neparametrickým testem. Opět uspořádáme v souboru hodnoty znaku $x_1 \leq x_2 \leq \dots \leq x_n$.

Testovací kritérium je dáno vztahy $Q_n = \frac{x_n - x_{n-1}}{x_n - x_1}$ nebo $Q_1 = \frac{x_2 - x_1}{x_n - x_1}$.

Kritické hodnoty $Q_{1:p} = Q_{n:p}$ jsou opět tabelovány.

Kritické hodnoty pro Dixonův test

n--p	0,05	0,01	n--p	0,05	0,01
3	0,941	0,988	17	0,32	0,416
4	0,765	0,889	18	0,313	0,407
5	0,642	0,78	19	0,306	0,398
6	0,56	0,698	20	0,3	0,391
7	0,507	0,637	21	0,295	0,384
8	0,468	0,59	22	0,29	0,378
9	0,437	0,555	23	0,285	0,372
10	0,412	0,527	24	0,281	0,367

11	0,392	0,502	25	0,277	0,362
12	0,376	0,482	26	0,273	0,357
13	0,361	0,465	27	0,269	0,353
14	0,349	0,45	28	0,266	0,349
15	0,338	0,438	29	0,263	0,345
16	0,329	0,426	30	0,26	0,341

Test náhodnosti

Zjišťují zda hodnoty měření kolísají vlivem náhodných změn či nikoli.

Test pomocí iterace

Každému prvku výběru přiřadíme určitým klíčem symbol A nebo B . Často se volí pro hodnoty pod mediánem A a nad mediánem B . (Při lichém počtu prvků medián neuvažujeme). V pořadí měření sestavíme posloupnost symbolů A, B , kde $n_1(A) + n_2(B) = n$.

Iterací pak rozumíme skupinu stejných symbolů. Počet iterací pro konkrétní hodnoty $n_1(A), n_2(B)$ má určité teoretické rozdělení, pomocí něhož byl tabelován nejmenší a největší počet iterací, který se již s danou pravděpodobností nemůže vyskytnout u náhodně uspořádaných posloupností. Tyto hodnoty opět nazveme kritické, přičemž případně přeznačíme tak, aby platilo $n_1(A) \geq n_2(B)$.

Kritické hodnoty počtu iterací

$p = 0,05$ $n_1 \geq n_2$

n_1--n_2	3	4	5	6	7	8	9	10
3	-6							
4	-7	-8						
5	-7	2-8	2-9					
6	2-7	2-8	3-9	3-10				
7	2-7	2-9	3-10	3-11	3-12			
8	2-7	3-9	3-10	3-11	4-12	4-13		
9	2-7	3-9	3-11	4-12	4-13	5-13	5-14	
10	2-7	3-9	3-11	4-12	5-13	5-14	5-15	6-15
11	2-7	3-9	4-11	4-12	5-13	5-14	6-15	6-16
12	2-7	3-9	4-11	4-12	5-13	6-15	6-15	7-16
13	2-7	3-9	4-11	5-13	5-14	6-15	6-16	7-17
14	2-7	3-9	4-11	5-13	5-14	6-15	7-16	7-17
15	3-7	3-9	4-11	5-13	6-14	6-15	7-17	7-17
16	3-7	4-9	4-11	5-13	6-15	6-16	7-17	8-18
17	3-7	4-9	4-11	5-13	6-15	7-16	7-17	8-18
18	3-7	4-9	5-11	5-13	6-15	7-16	8-17	8-18
19	3-7	4-9	5-11	6-13	6-15	7-16	8-17	8-19
20	3-7	4-9	5-11	6-13	6-15	7-16	8-17	9-19

n1--n2	11	12	13	14	15	16	17	18	19	20
11	7-16									
12	7-17	7-18								
13	7-18	8-18	8-19							
14	8-18	8-19	9-19	9-20						
15	8-18	8-19	9-20	9-21	10-21					
16	8-19	9-20	9-20	10-21	10-22	11-22				
17	9-19	9-20	10-21	11-22	11-22	11-23	11-24			
18	9-19	9-20	10-21	11-23	11-23	11-24	12-24	12-25		
19	9-20	10-21	10-22	11-23	11-23	12-24	12-25	13-25	13-26	
20	9-20	10-21	10-22	12-24	12-24	12-24	13-25	13-26	13-26	14-27

Tento test lze použít i jako test shody dvou výběrů (tj. jako test, zda dva výběry pocházejí ze stejného základního souboru). V tomto případě uspořádáme prvky obou výběrů do společné uspořádané posloupnosti, pak podle příslušnosti k prvnímu nebo druhému výběru označíme jednotlivé prvky A nebo B a opět stanovíme počet iterací. Nulovou hypotézu zamítneme za týchž podmínek, jako u testu náhodnosti.

7.3.5 Závislost dvou náhodných veličin

Závislosti

- A) funkční – určité hodnotě nezávisle proměnné odpovídá pevná hodnota veličiny závisle proměnné ($s = \frac{1}{2}gt^2$)
- B) **statistické** – určité hodnotě nezávisle proměnné odpovídá určitý obor hodnot veličiny závisle proměnné

Add B

Mějme výběr o n prvcích, na kterých měříme hodnoty dvou kvantitativních znaků, takže získáme n dvojic $x_1, y_1; x_2, y_2; \dots; x_n, y_n$.

Graficky výsledky znázorníme tečkovým diagramem.

Rozeznáváme tři různé varianty statistických závislosti (tj. korelací)

- I. statistická nezávislost – tečkový diagram tvaru kruhu
- II. lineární korelace – tečkový diagram tvaru elipsy
- III. nelineární korelace

Add I, II

Pro stanovení „jakosti“ lineární korelace určíme koeficient korelace r

Sylabus

$$r = \frac{s_{xy}^2}{s_x \cdot s_y} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y} = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{s_x \cdot s_y} = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2\right) \left(\frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}^2\right)}}$$

Koeficient korelace může nabývat hodnot od -1 do 1. Z toho mají hodnoty -1,0,1 speciální význam. Hodnota 0 znamená, že mezi znaky není lineární vztah. Pro ± 1 dostáváme extrémní případ, kdy statistická závislost přechází v závislost funkční.