

Cvičení z předmětu Úvod do statistické analýzy

Josef Chudoba

Kapitoly v této cvičebnici jsou řazeny stejně jako na přednáškách. Součástí zadání příkladů je i jejich řešení, které je uvedeno v adresáři řešení. Soubory týkající se určitého příkladu jsou uvedeny pomocí písmenné (číselné) kombinace a platí následující pravidla:

- P – písmenný znak, aby soubor začínal písmenem,
- první dvojčíslí – číslo kapitoly (např. 04 značí příklad ze 4. kapitoly),
- druhé dvojčíslí – číslo příkladu v kapitole (např. 0403 značí 3. příklad ze 4. kapitoly),
- označení res/zad – zad značí vstupní data, res řešení příkladu

K příkladům jsem se snažil psát komentářové poznámky.

Obsah

Cvičení z předmětu Úvod do statistické analýzy	1
1 Kombinatorika.....	2
2 Úvod do teorie pravděpodobnosti	7
3 Náhodná veličina a náhodný vektor	13
4 Diskrétní rozdělení pravděpodobnosti.....	17
5 Spojitá rozdělení pravděpodobnosti.....	22
6 Výběrové charakteristiky.....	26
7 Teorie odhadu.....	31
8 Testy hypotéz.....	35
9 Testy dobré shody.....	43
10 Analýza závislostí	48
11 Úvod do korelační a regresní analýzy.....	50

1 Kombinatorika

1.1 Faktoriál, kombinační čísla

Př. 1:

Jak a pro jaká čísla je definován faktoriál.

Př. 2:

Vypočtěte faktoriál čísla 20 pomocí a) pomocí for cyklu, b) pomocí příkazu počítající faktoriál (funkce **factorial**). Zkuste se zamyslet a spočítat faktoriál 1000000.

[2.4329e+18, 2.4329e+18, 8.2639e+5565708]

Př. 3:

Jak se vypočítají kombinační čísla a co musí pro ně platit. Jak se značí.

Př. 4:

a) Vypočtěte kombinační číslo $\binom{10}{6}$ pomocí faktoriálu a pomocí v matlabu implementované funkce (funkce **nchoosek**).

b) Zkuste obdobně spočítat kombinační číslo $\binom{100}{60}$ a $\binom{1000}{600}$. Zauvažujte, proč nelze vypočítat na počítači pomocí faktoriálu kombinační číslo $\binom{1000}{600}$ a jak lze problém obejít.

[210, 1.374623414580280e+28, 4.965272386254229e+290]

1.2 Pravděpodobnost

Př. 5:

Co je to pravděpodobnost a jak je definována (dle středoškolské matematiky, vysokoškolské definice se naučíte později v kap. 2)?

Př. 6:

Z telefonního čísla, které má devět cifer, jste poslední dvě cifry zapomněli. Víte pouze, že jsou různé. Zkoušte je náhodně. Jaká je pravděpodobnost, že vytočené číslo je správné?

[0.0111]

Př. 7:

Na pěti kartičkách jsou napsány číslice 1, 2, 3, 4, 5. Vyberete tři kartičky, které najednou otočíte. Vypočtěte pravděpodobnost, že číslo je sudé.

[0.4]

Př. 7a:

Jaká je pravděpodobnost, že z balíčku karet obsahující 32 karet vyberete buď spodka nebo filka. V balíčku jsou 4 spodci a 4 filci.

[0.25]

1.3 Variace, permutace, kombinace

Př. 8:

Co to jsou variace a jaký je rozdíl mezi variacemi bez opakování a s opakováním. Zkuste odvodit vzorec na jejich výpočet.

Př. 9:

V mariášovém turnaji je 10 hráčů. Kolika způsoby mohou obsadit první čtyři místa v turnaji?

[5040]

Př. 10:

5x házíme šestistěnnou hrací kostkou a zaznamenáváme na papír výslednou číselnou hodnotu. Kolik existuje možností zapsaných čísel?

[7776]

Př. 11:

V osudí je 10 karet s ciframi 0 až 9. Vybereme 6 cifer a položíme je na stůl. Kolik různých čísel můžeme zaznamenat? První cifra může být 0.

[151200]

Př. 12:

Na papíru máte napsané následující cifry 1,1,1,2,2,2,2,3,3,3,3. Kolik dvanáctimístných čísel z nich můžete vytvořit.

[27720]

Př. 13:

Uveďte vzorce pro výpočet kombinací bez opakování a s opakováním. Odvoďte správnost vzorce pro kombinace bez opakování.

Př. 14:

V prodejně si můžete vybrat ze sedmi druhů pohlednic. Od každé pohlednice mají dostatečný počet exemplářů. Kolika způsoby lze koupit a) 10 pohlednic, b) 5 pohlednic, c) 5 různých pohlednic

[8008, 462, 21]

Př. 15:

Osm přátel si poslalo vzájemně pohlednice z prázdnin. Kolik pohlednic celkem rozeslali?

[56]

Př. 16:

Na florbalovém mistrovství je 12 mužstev. Kolik se odehraje zápasů, jestliže každý hraje s každým.

[66]

1.4 Kombinatorické pravidlo součinu

Př. 17:

Třikrát hodíte hracími kostkami. Jaká je pravděpodobnost, že a) padnou tři šestky, b) padnou všechna stejná čísla, c) padnou dvě šestky a jedno číslo jiné.

[0.0046, 0.0278, 0.0694]

Př. 18:

Máme dva balíčky karet, v prvním jsou pouze červené (8 různých karet) a v druhém pouze žaludy (také 8 různých karet). Z prvního balíčku vyberete 5 karet, z druhého 4 karty. Kolik existuje kombinací výběru karet?

[3920]

Př. 19:

Jaká je pravděpodobnost, že v normálním balíčku mariášových karet, při výběru 4 karet budou samá esa?

- a) Vybranou kartu vracíte do balíčku.
- b) Vybranou kartu necháte venku.

[2.44e-4, 2.78e-5]

Př. 20:

Vyučující připravuje zadání písemné práce. Má k dispozici 10 příkladů, z nichž vybere 3. Studenti znají z loňského roku zadání 5 příkladů, které se mohou objevit v písemné práci.

- a) Zjistěte pravděpodobnost, že učitel vybere z 10 příkladů všechny ty, jejichž zadání studenti znají.
- b) Zjistěte pravděpodobnost, že studenti budou znát pouze 2 zadání příklady.

[0.0833; 0.4167]

Př. 21:

Učitel má připraveno 15 příkladů z pravděpodobnosti a 5 příkladů ze statistiky.

- a) Náhodně zvolí 6 příkladů, kolik existuje možností výběru příkladů?
- b) Jaká je pravděpodobnost, že z 6 vybraných příkladů budou právě 2 ze statistiky.

[38760; 0.3522]

Př. 22:

Dle zadání příkladu 21 dokažte, že součet pravděpodobností přes všechny možné stavy (vybran 0, 1 ... 5 příkladů ze statistiky) bude roven 1.

[a) 0.1291, 0.3874, 0.3522, 0.1174, 0.0135, 0.0004, součet =1]

Př. 23:

Hodí se n kostek. Určete pravděpodobnost, že na všech bude stejné číslo.

$$[\left(\frac{1}{6} \right)^{n-1}]$$

1.5 Sčítání pravděpodobnosti

Př. 23a:

V mariášovém balíčku je 32 karet (8 hodnot po 4 barvách). Vypočtěte pravděpodobnost, že vyberete spodka, nebo filka, nebo žaludovou barvu.

[14/32 = 0.4375]

Př. 24: Mezi 100 a 1000 (čísla na kraji interval započítejte) je 143 prvočísel. Určete pravděpodobnost, že náhodně vybrané celé číslo mezi 100 a 1000 je:

- a) Prvočíslo (funkce **primes**), nebo dělitelné dvěma, nebo dělitelné třemi
- b) není prvočíslo
- c) je dělitelné dvěma nebo pěti

[0.8257, 0.8413, 0.6004]

1.6 Obecné příklady a opakování

Př. 25

Kolika způsoby je možno na šachovnici s 64 poli vložit 5 věží tak, aby se vzájemně neovlivňovaly, tj. žádná neležela ve stejném sloupci a řádku.

[376320]

Př. 26

Vytvořte Pascalův trojúhelník kombinačních čísel pro 8 prvků (funkce **pascal**). Zkuste z něj zjistit následující kombinační čísla $\binom{6}{1}$, $\binom{6}{2}$, $\binom{6}{3}$, $\binom{6}{4}$. Určete základní početní pravidla pro aritmetické operace s kombinačními čísly.

$$(6, 15, 20, 15; \binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k})$$

Př. 27

Ověřte, že součet kombinačních čísel $\sum_{i=0}^m \binom{m}{i}$, pro pět libovolných m je 2^m . (Ověření platnosti binomické věty.)

Př. 28

Máte funkci $y = (1 + x)^{28}$. Spočtěte koeficient u x^6 , x^9 a x^{12} .

[376 740, 6 906 900, 30 421 755]

Př. 29

V atletickém oddíle je 15 chlapců a 10 dívek. Pro reprezentaci je nutné vybrat deset členů (5 chlapců a 5 dívek). Kolik možností výběru existuje?

[756756]

2 Úvod do teorie pravděpodobnosti

2.1 Jevy

Př. 1

Házíme hrací kostkou. Máte následující jevy: $A = \{2,4\}$, $C = \{2,5\}$

- a) Jaké jsou elementární jevy při hodu kostkou? $\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$
- b) Uveďte příklad složeného jevu. $\{\{1,2\}\}$
- c) Jaký je doplněk k jevu A. $\{\{1,3,5,6\}\}$
- d) Vytvořte jev B, který je disjunktní k jevu A. [například $\{3,5\}\}$]
- e) Jaký je průnik jevu A a C. $\{\{2\}\}$
- f) Jaké je sjednocení jevu A a C. $\{\{2,4,5\}\}$
- g) Jaký je rozdíl jevů A a C. $\{\{4,5\}\}$

Př. 2

Jev A je množina čísel dělitelných dvěma beze zbytku. Jev B je množina čísel dělitelných třemi beze zbytku. Jev C je množina čísel dělitelných čtyřmi beze zbytku. Jev D je množina čísel dělitelných pěti beze zbytku. Určete:

- a) $A \cup B$ [množina čísel, která po dělení 6 mají zbytek 0,2,3,4]
- b) $A \cap B$ [množina čísel dělitelných 6 beze zbytku]
- c) $A \cap D$ [množina čísel dělitelných 10 beze zbytku]
- d) \bar{A} [lichá čísla]
- e) $\bar{A} \cap \bar{A}$ [lichá čísla]
- f) $A \cup \bar{A}$ [množina celých čísel]
- g) $A \cap C$ [C]

2.2 Klasická pravděpodobnost

Př. 3:

Dřevěnou kostku o straně 5 cm natřeme na červeno a rozřežeme na krychle o hraně 1 cm. Určete pravděpodobnost, že vylosujete krychličku, která

- a) Nebude obarvena. [0.216]
- b) Bude obarvena na jedné stěně. [0.432]
- c) Bude obarvena na dvou stěnách. [0.288]
- d) Bude obarvena na třech stěnách. [0.064]

Př. 4:

Máme 10 druhů minerálek. 6 je perlivých, zbývající neperlivé. Určete pravděpodobnost, že z náhodně vybraných 3 minerálek budou

- a) Všechny perlivé. [0.167]
- b) Jedna perlivá, ostatní neperlivé. [0.300]
- c) Jaké jsou elementární jevy pokusu? [0 perlivých, 1 perlivá, 2 perlivé, 3 perlivé]

d) Dokažte výpočtem, že součet pravděpodobností přes všechny elementární jevy je roven 1.

[0 perlivých 0.033; 1 perlivá 0.300; 2 perlivé 0.500; 3 perlivé 0.167]

Př. 5

Ve hře šťastných deset je v osudí 80 míčků, z nichž se losuje 20. Sázející vybere 10 čísel. Určete pravděpodobnost, že uhádne právě 0 až 10 čísel.

[0.045790700789028	0.179571375643246	0.295256781105723	0.267402367793862
0.147318897071618	0.051427687705001	0.011479394577009	0.001611143098528
0.000135419355264	0.000006120648825	0.000000112211895]	

Př. 6

20 studentů má být rozděleno na 4 stejně početné skupiny. Jaká je pravděpodobnost, že A a B budou ve stejné skupině?

[4/19]

Př. 7

Jaká je pravděpodobnost, že z náhodně poskládaných písmen A,A,B,C,D,E,E sestrojíte slovo ABECEDA?

[7.94 E-4]

Př. 8

Jaká je pravděpodobnost, že mezi 3 mariášovými kartami náhodně vytažených z balíčku bude právě 1 eso?

[0.3048]

Př. 10

Postupně vydávám koule z urny, kde jsou 3 bílé, 5 černých a 4 červené koule. Jaká je pravděpodobnost, že červenou vytáhnu dříve než bílou? Koule nevracíme.

[0.5714]

2.3 Geometrická pravděpodobnost

Př. 11

Ve čtverci o délce hrany 3 cm je kružnice o poloměru 1 cm. Jaká je pravděpodobnost, že náhodně zvolený bod ze čtverce je zároveň uvnitř kružnice.

[0.3491]

Př. 12

Dva lidé si dají schůzku mezi 12 a 13. hodinou. Přijdou v tomto čase zcela náhodně. Jaká je pravděpodobnost, že se setkají, přichází-li nezávisle a čeká-li jeden na druhého přesně 15 minut.

[7/16]

Př. 13

Jaká je pravděpodobnost, že náhodná čísla v intervalu $<0,1>$ budou od sebe vzdáleny na ose méně než 0.1. Jak se změní výsledek úlohy, pokud vzdálenost dvou čísel bude k ($k < 1$)?

[0.19; $1-(1-k)^2$]

2.4 Vlastnosti pravděpodobnosti, nezávislost jevů, Bayesova věta

Př. 14

Pomocí znalosti pravděpodobností jednotlivých jevů a jejich průniků vyjádřete obecně $P(A \cup B \cup C)$. Ověřte výsledek pomocí Vennova diagramu.

$[P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)]$

Př. 15

Mějme $P(A) = 0.3$, $P(B) = 0.5$ a $P(A \cap B) = 0.1$. Určete a zakreslete pomocí Vennova diagramu:

- a) $P(A \cup B)$
- b) $P(A \cap \bar{B})$
- c) $P(\bar{A} \cap \bar{B})$
- d) $P(\bar{A} \cup \bar{B})$

[0.7; 0.2; 0.3; 0.9]

Př. 16

Dvakrát hodíme minci. Jsou výsledky jednotlivých hodů nezávislé?

[ano jsou]

Př. 17:

Z balíčku 32 mariášových karet náhodně vytáhneme jednu kartu. Jev A spočívá ve vytažení žaludové karty, jev B ve vytažení esa. Určete, zda jevy jsou nezávislé a pokud ano určete pravděpodobnost nastolení obou jevů současně.

[ano jsou, 1/32]

Př. 18

Máte zadání z příkladu 15, určete, zda jevy jsou vzájemně nezávislé. Zjistěte podmíněnou pravděpodobnost toho, že nastane jev A za podmínky, že nastal jev B.

[nejsou, 0.2]

Př. 19

V populaci je 20 % lidí se srdeční chorobou; 40 % populace jsou kuřáci. 12 % populace jsou kuřáci se srdeční chorobou. Jaká je pravděpodobnost, že:

- a) člověk má srdeční chorobu za předpokladu, že je kuřák,
- b) člověk má srdeční chorobu za předpokladu, že je nekuřák.

[30%; 13,3 %]

Př. 20

Na závodech na 110 m překážek zvítězí běžec A s pravděpodobností 0.3; běžec B s pravděpodobností 0.2. Při závodu běžec A spadl na překážce. Jaká je pravděpodobnost, že běžec B závod vyhraje.

[0.2857]

Př. 21

Výrobek je s pravděpodobností 0.8 zařazen do 1. jakostní skupiny, s pravděpodobností 0.1 do druhé, s pravděpodobností 0.1 bude do třetí. Pravděpodobnost, že výrobek bude v provozuschopném stavu po stanovenou dobu je pro jednotlivé jakostní skupiny 0.8; 0.6 a 0.3. Jaká je pravděpodobnost, že náhodně získaný výrobek bude v provozuschopném stavu po stanovenou dobu.

[0.73]

Př. 22

Automat A vyrábí za směnu třikrát víc výrobků než automat B. Přičemž automat A má zmetkovitost 0.01 a automat B 0.002. Výrobky se vhazují do stejné krabice. Jaká je pravděpodobnost, že náhodně vybraný výrobek není zmetek.

[0.992]

Př. 23

Neprůhledný pytlík obsahuje 8 černých a 5 bílých kuliček. Budeme provádět náhodný pokus vytažení jedné kuličky, přičemž kuličku do pytlíku nevracíme. Jevy jsou definovány následovně:

- B1 – při první realizaci byla vytažena bílá kulička
- B2 – při druhé realizaci byla vytažena bílá kulička
- C1 – při první realizaci byla vytažena černá kulička
- C2 – při druhé realizaci byla vytažena černá kulička

Co znamenají jevy: $B2|B1$, $B2|C1$, $C2|B1$, $C2|C1$ a jaké jsou jejich pravděpodobnosti.

[$B2|B1$ – při druhé realizaci náhodného pokusu byla vytažena bílá kulička, za předpokladu že při první realizaci byla vytažena také bílá kulička. Dále obdobně;

$$B2|B1 = \frac{4}{12}, B2|C1 = \frac{5}{12}, C2|B1 = \frac{8}{12}, C2|C1 = \frac{7}{12}$$

Př. 24 Pravděpodobnost, že selže aktivní hasicí systém při požáru je 5 %. Pravděpodobnost, že selže signalizace na centrálním pultu je 8 %. Pravděpodobnost, že selžou oba systémy najednou je 3 %.

- a) Jaká je pravděpodobnost, že selže hasicí systém, ale nikoliv signalizace.
- b) Zapůsobí oba dva systémy.

[0.02, 0.90]

Př. 25

160 studentů absolvovalo zkoušky ze statistiky a aplikované matematiky. 20 z nich nesložilo obě zkoušky, dalších 10 nesložilo pouze ze statistiky a dalších 60 nesložilo zkoušku z aplikované matematiky. Určete podmíněnou pravděpodobnost, že náhodně vybraný student:

- a) složil zkoušku z aplikované matematiky, za předpokladu, že nesložil zkoušku ze statistiky,
- b) složil zkoušku ze statistiky, víme-li, že nesložil zkoušku z aplikované matematiky,
- c) složil zkoušku z aplikované matematiky, víme-li, že složil zkoušku ze statistiky.

Uveďte výsledky i se zápisem podmíněných pravděpodobností.

$$P(A|\bar{S}) = \frac{P(A \cap \bar{S})}{P(\bar{S})} = \frac{P(A \cap \bar{S})}{P(A \cap \bar{S}) + P(\bar{A} \cap \bar{S})} = \frac{10}{30}; P(S|A) = \frac{60}{80}; P(A|S) = \frac{70}{130}$$

Př. 26

V běhu na 100 metrů běží 8 závodníků. Šance závodníka A na vítězství je 50 %, u závodníka B 30 % a u závodníka C 15 %. Závodník B ulil start a odstoupil ze závodu. Jaká je nyní pravděpodobnost na vítězství závodníka C.

[21.4 %]

Př. 27

V první urně je 6 bílých a 3 černé koule. V druhé urně je 5 bílých a 1 černá koule. Náhodně zvolíme urnu a vytáhneme jednu kouli. Jaká je pravděpodobnost, že bude bílá.

[0.75]

Př. 28

Jeden ze 3 střelců s pravděpodobnostmi zásahu 0.3, 0.5 a 0.8 vystřelil a zasáhl. Jaká je pravděpodobnost, že střílel druhý střelec.

[0.3125]

Př. 29

Výrobní linka produkuje pouze 60 % kvalitních výrobků. Proto se každý testuje. V případě, že je výrobek nekvalitní, bude s pravděpodobností 90 % odhalen a odstraněn. Naopak kontrola nepropustí 3 % kvalitních výrobků.

- a) jaká je pravděpodobnost, že nekvalitní výrobek projde kontrolou,
- b) jaké je procento neporouchaných výrobků v odstraněných výrobcích.

[6.43 %, 4.76 %]

Př. 30

Na vojenský cíl bylo čtyřikrát vystřeleno. Pravděpodobnost zásahu je 0.1. Jaká je pravděpodobnost, že cíl bude vyřazen, jestliže:

- a) při 3 nebo 4 zásazích je pravděpodobnost zničení 0.9,
- b) při 2 zásazích 0.4,
- c) při 1 zásahu 0.1.

[0.05193]

3 Náhodná veličina a náhodný vektor

3.1 Základní pojmy

Př. 1: Uveďte pět příkladů na diskrétní a spojitou náhodnou veličinu. Uveďte příklady jevů.

Př. 2: Uveďte příklad diskrétní náhodné veličiny, která může mít konečný a nekonečný počet jevů.

3.2 Distribuční funkce

Př. 3: Uveďte vlastnosti distribuční funkce, a jakým způsobem byste ověřovali její vlastnosti.

Př. 4: Máte funkci $y = 1 - e^{-\frac{t}{s}}$. Vykreslete graf funkce (funkce **plot**) a zjistěte, zda se skutečně jedná o distribuční funkci.

[ano]

Př. 4a) Otevřete skript P0304a.mat a určete, které z grafů jsou distribuční funkce. Zdůvodněte.

[priklad 1 a 4]

Př. 5: Máte vygenerovaný soubor 10000 naměřených hodnot, který je uložen v P0305.mat. Vytvořte z nich distribuční funkci. Pro třídění dat použijte příkaz **sort**. Dále vytvořte z dat histogram o 50 sloupcích.

Př. 6: Máte výsledky 10000 hodů šestistěnnou kostkou. Data jsou uložena v P0306.mat. Vytvořte distribuční funkci z výsledků.

3.3 Diskrétní náhodná veličina

Př. 7: Nakreslete do grafu pravděpodobnostní funkci, která je dána předpisem $P(x) = \binom{10}{x} 0.1^x 0.9^{10-x}$.

Př. 8: Obdrželi jste pravděpodobnostní funkci, která je dána v tabulce. Vytvořte graf pravděpodobnostní funkce.

Př. 9: V souboru P0309.mat je 10000 naměřených dat. Vytvořte graf, kde budou vyobrazeny 4 histogramy. První bude obsahovat 10, druhý 100, třetí 500 a čtvrtý 5000 sloupců. Jaký byste z nich označily jako nejlepší?

3.4 Spojitá náhodná veličina

Př. 10: Máte distribuční funkci ve tvaru $F(x) = 1 - e^{-(ax)^b}$, kde $a,b > 0$. Vypočtěte hustotu distribuční funkce. Užijte knihovnu **symbolic**.

[a*b*exp(-(a*x)^b)*(a*x)^(b - 1)]

Př. 11: Hustota pravděpodobnosti má tvar $f(x) = \lambda e^{-\lambda x}$. Zjistěte distribuční funkci.

[1-exp(-lambda*t)]

3.5 Funkce náhodné veličiny

Př.12: V souboru P0312.mat máte vygenerováno 1000 dvojic dat (první sloupec vektor x, druhý sloupec vektor y), která byla z rovnoměrného rozdělení v intervalu $\langle 0,1 \rangle$. Nakreslete 4 grafy:

- v prvním bude distribuční funkce náhodné veličiny z vektoru x,
- v druhém bude distribuční funkce náhodné veličiny ($x+y$),
- ve třetím bude distribuční funkce náhodné veličiny ($x*y$),
- ve čtvrtém bude distribuční funkce náhodné veličiny (x/y). Vodorovnou osu dekadicky zlogaritmujte.

Uvědomte si, jaký vliv mají transformace náhodné proměnné na výsledky. Distribuční funkci můžete vygenerovat pomocí funkce ecdf.

3.6 Číselné charakteristiky náhodné veličiny

Př. 14: Máte distribuční funkci ve tvaru $F(x) = 1 - e^{-ax}$. Uvažujte, že parametr $a=0.1, 0.2$ a 0.5 ($t>0$). Vypočtěte hustotu pravděpodobnosti, intenzitu náhodného jevu ($\lambda = \frac{f(x)}{1-F(x)}$), střední hodnotu, rozptyl, směrodatnou odchylku, šikmost a špičatost. Zkuste na základě výsledků odvodit vzorce pro výše uvedené veličiny.

[střední hodnota je $1/a$; rozptyl $1/a^2$; směrodatná odchylka $1/a$; šikmost=2, špičatost = 9]

Př. 15: Máte hustotu pravděpodobnosti ve tvaru $f(x) = \frac{1}{b-a}$, pro $a < x < b$. Vypočtěte střední hodnotu, rozptyl, směrodatnou odchylku, šikmost a špičatost.

$$\left[\frac{a+b}{2}, \frac{(a-b)^2}{12}, \sqrt{\frac{(a-b)^2}{12}}, 0, \frac{9}{5} \right]$$

Př. 16: Máte rozdělení s distribuční funkcí ve tvaru $\frac{1}{2} + \frac{1}{\pi} \arctg \frac{x}{2}$. Zjistěte, zda se jedná o statistické rozdělení. Pokud ano, určete median a střední hodnotu. Nakreslete graf rozdělení.

[ano je statistické rozdělení, median = 0, střední hodnota neexistuje]

Př. 16a: Náhodná veličina X má distribuční funkci $\frac{x^2}{4}$ na intervalu $(0,2)$, nulovou pro $x<0$ a jednotkovou pro $x>2$. Najděte hustotu funkce, medián a střední hodnotu. Určete pravděpodobnost $P(0.5 < x < 1.5)$.

[$x/2, 1.4142, 1, 0.5$]

Př. 18: Náhodná veličina má distribuční funkci $F(x) = \frac{1}{8}x^3$, pro $x \in \langle 0,2 \rangle$. Určete střední hodnotu a rozptyl náhodné veličiny. Jaká je pravděpodobnost, že náhodná veličina bude mít výsledek v intervalech:

- ⟨0,1⟩
- ⟨0.5,1.5⟩
- ⟨0,1⟩ ∪ ⟨1.5,2⟩

[střední hodnota=3/2; rozptyl = 3/20; a) 0.1250; b) 0.4063; c) 0.7031]

Př. 19: Nechť X je spojitá náhodná veličina definována hustotou pravděpodobnosti $f(x)$:

$$f(x) = c(2 - x)(2 + x)$$

V intervalu $<-2,2>$. Hustota pravděpodobnosti je nulová jinde. Úkolem je

- Nalézt konstantu c tak, aby $f(x)$ byla korektně zadána. Uvažte, že při integraci příčítáte i posun.
- Nakreslit do jednoho grafu hustotu pravděpodobnosti a distribuční funkci
- Určete pravděpodobnost $P(X < 0.3)$; $P(0 < X < 1)$ a $P(X > 1)$

[$1/2 - (x^*(x^2 - 12))/32$; $P(X < 0.3) = 0.6117$; $P(0 < X < 1) = 0.3438$; $P(X > 1) = 0.1563$]

3.7 Statistiké charakteristiky kvalitativních proměnných

Př. 20: Vypočtěte pravděpodobnost, že uhádnete právě n čísel ve hře šťastných desek (viz příklad P0205). Výsledky zobrazte graficky.

Př. 21: Předmět na univerzitě si zapsalo 80 studentů, z nichž 4 vykonaly zkoušku na první termín, 16 na druhý, 25 na třetí a 10 studentů zkoušku nesplnilo ani na třetí termín (slopec 4). Zbytek studentů neobdržel zápočet (slopec 5). Zjistěte relativní četnosti a vytvořte sloupcový graf výsledků. Vytvořte distribuční funkci pomocí funkce `cdfplot` a histogram (funkce `hist`).

Př. 22: Byly provedeny 2 průzkumy. Prvního se zúčastnilo 5 respondentů a výsledky byly 40 % ano a 60 % ne. Druhého výzkumu na stejně otázky se zúčastnilo 50 respondentů. Výsledky byly 24 % ano, 62 % ne a zbytek neví. Jaký test má dle vašeho názoru vyšší váhu a proč.

[druhý test]

Př. 23: Máte k dispozici záznamy o poruchách na jednotlivých zařízeních v rámci jednoho podniku (soubor P0323.xlsx). Data obsahují datum poruchy a typ zařízení, kde byla porucha nalezena. Vytvořte histogram pro poruchovost jednotlivých komponent.

3.8 Statistiké charakteristiky numerických proměnných

Př. 24: Proč se výběrové statistiky rozptylu a dalších odlišují od způsobu výpočtu rozptylu počítaných pomocí integrálu.

Př. 25: Od kolika hodnot lze provést výpočet výběrové střední hodnoty, výběrového rozptylu, výběrové směrodatné odchylky, výběrové šíkmosti a výběrové špičatosti. Uvědomte si kolikáté centrální (obecné) momenty využíváte pro výpočet.

[1, 2, 2, 3, 4]

Př. 26: Máte naměřená diskrétní data uložená v souboru P0326.xlsx. Zjistěte z nich střední hodnotu, medián, modus, rozptyl a směrodatnou odchylku. Pro kontrolu zkuste vytvořit histogram a distribuční funkci (příkaz `cdfplot`). Proveďte diskuzi, zda výsledky mohou být správné.

[12.44, 11, 0.94.56, 9.72]

Př. 27: Místní odborová organizace zveřejnila hrubé mzdy ve firmě, která jsou uložená v souboru P0327.xlsx. Zjistěte z nich aritmetický a geometrický průměr. Zjistěte medián, vytvořte histogram a distribuční funkci z naměřených dat. Proč není účelné zjišťovat modus.

[aritmetický 22 815 Kč, geometrický 21 374 Kč, medián 21 579 Kč]

Př. 28: Máte data uložená v souboru 0328.mat. Zjistěte z nich aritmetický průměr a výběrový rozptyl. Dále určete 5% a 95% kvantil a dolní a horní kvartil. Jistě nepohrdnete i znalostí mediánu. Vytvořte empirickou distribuční funkci.

Př. 29: Byly zjištěny výsledky ze 100 teplotních čidel (údaje jsou v K). V jakých jednotkách bude střední hodnota, medián, rozptyl, směrodatná odchylka, šíkmost a špičatost.

[rozptyl K^2 , šíkmost 1, špičatost 1, vše ostatní K]

Př. 30: Mějte výsledky měření z 5 aparatur, které jsou uloženy v souboru P0330.xlsx. Vypočtěte pro každý typ aparatury střední hodnotu, rozptyl a směrodatnou odchylku.

Př. 31: Máte naměřená data uložená v souboru P0331.mat. Určete metodou vnitřní hradby, z-souřadnice a $x_{0,5}$ souřadnice hranici pro odlehlá měření. Zkuste vysvětlit, proč zjištěná odlehlá data nejsou pomocí všech tří testů shodná.

Pozn. Data vypadají obdobně, protože byly vygenerovány z normálního rozdělení.

[686, 1308 677,1311 610,1366]

Př. 32: Máte naměřená data, která jsou uložena v souboru P0332.mat. Zjistěte 5%, 50% a 95% kvantil těchto hodnot. Užijte funkci **quantile**. Vykreslete distribuční funkci z naměřených dat.

[10.82, 24.08, 37.39]

Př. 33: Máte naměřená data, která jsou uložena v souboru P0333.mat. Zjistěte 5%, 50% a 95% kvantil těchto hodnot. Určete střední hodnotu a median. Zkuste zauvažovat, co značí výrazná odlišnost střední hodnoty a mediánu. Může Vám pomoci histogram, použijte 50 sloupců.

[2021,118, nesymetrická data]

4 Diskrétní rozdělení pravděpodobnosti

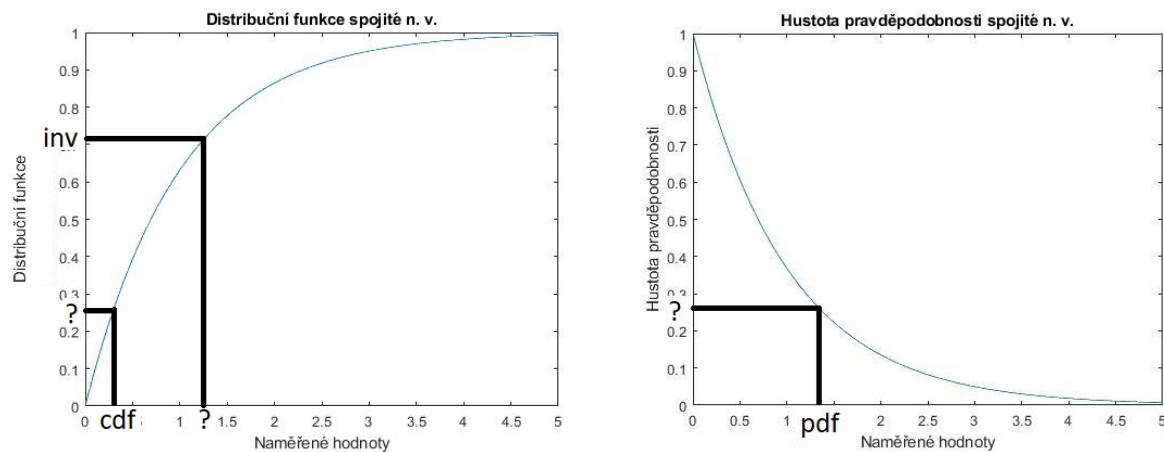
4.1 Alternativní rozdělení

Př. 1: Vypočtěte z pravděpodobnostní funkce ($P(X = 1) = p; P(X = 0) = 1 - p$) střední hodnotu, směrodatnou odchylku, rozptyl, šikmost a špičatost.

$$[E=p; D=p*(1-p); \text{sigma}=(p*(1-p))^{(1/2)}; a3=(1-2*p)/(p*(1-p))^{(1/2)}$$

$$a4=(3*p^2-3*p+1)/(p*(1-p))]$$

4.2 Binomické rozdělení



Binocdf Distribuční funkce binomického rozdělení

Binopdf Pravděpodobnostní funkce binomického rozdělení

Binoinv inverzní funkce k distribuční funkci binomického rozdělení

Binostat střední hodnota a rozptyl binomického rozdělení

Binofit odhad parametrů binomického rozdělení

Binornd náhodná čísla z binomického rozdělení

Př. 2: Jakým způsobem se odlišuje binomické rozdělení od alternativního.

Př. 3: Napište si vzorec pro pravděpodobnostní funkci Binomického rozdělení a uvědomte si význam jednotlivých členů.

Př. 5: Pětkrát hodíme mincí. Určete pravděpodobnost, že orel padne právě dvakrát. Určete pravděpodobnost, že padne alespoň 4.

$$[0.3125, 0.1875]$$

Př. 5a: Pravděpodobnost, že ve sportce uhádnu první cenu je $\frac{1}{13983816}$. Vypočtěte pravděpodobnost, že když vsadím za rok 1000x, že vyhraji právě dvakrát první cenu.

[2.5542e-09]

Př. 5b:

Zásilka obsahuje 80 % kvalitních a 20 % nekvalitních výrobků. Náhodně s vracením vybereme 5 výrobků. Určete pravděpodobnost, že:

- a) právě 3 budou kvalitní,
- b) alespoň 3 budou kvalitní

[0.2048, 0.9421]

Př. 6: Pravděpodobnost narození dívčete je 0.49. Určete pravděpodobnost, že ve třídě mající 25 dětí bude (neuvážujte jednopohlavní třídy):

- a) Právě 10 dívek,
- b) Alespoň 10 a více dívek,
- c) Více než 15 dívek,
- d) Kolik dívek bude ve třídě nejpravděpodobněji.

Nakreslete graf, kde bude vynesena pravděpodobnost počtu dívek ve třídě.

[0.1071; 0.8646; 0.0964; 12]

Př. 7: Pravděpodobnost narození chlapce je 0.51. Určete minimální počet dětí, aby pravděpodobnost, že mezi nimi bude alespoň jeden chlapec, byla větší než 0.99.

[7 dětí]

Př. 8: Víme, že mezi výrobky je 10 % vadných. Určete pravděpodobnost, že u 20 náhodně vybraných výrobků bude:

- a) právě 0 vadných
- b) více než 5 ks vadných.

[12.16 %, 1.13 %]

Př. 8a: Mějme mariášové karty (32 karet, které obsahují 4 esa, 4 krále, 4 filky, ..., 4 sedmičky). Losujete karty z balíčku a vracíte je zpět. Určete pravděpodobnost, že z prvních 7 vylosovaných karet dostanete právě 4 esa nebo krále. A poslední osmá vylosovaná karta bude 7.

[0.0072]

4.3 Hypergeometrické rozdělení

Hygecdf Distribuční funkce hypergeometrického rozdělení

Hygepdf Pravděpodobnostní funkce hypergeometrického rozdělení

hygeinv Inverzní funkce k distribuční funkci hypergeometrického rozdělení

Př. 9: Jak se odlišuje hypergeometrické rozdělení od binomického. Jak byste rozdíl těchto rozdělení simulovali u pokusu, kde máte v krabici m černých a n bílých koulí.

Př. 10: Vypočtěte pravděpodobnost, že z 32 karetního balíčku budou při vylosování 3 karet právě 2 esa. Jak se změní pravděpodobnost, jestliže karty do balíčku vracíme a pokud je nevracíme.

[0.0339, 0.0410]

Př. 11: V loterii je v osudí 200 čísel, z nichž se losuje 30. Jaká je pravděpodobnost, že vybereme-li náhodně 10 čísel, bude z nich právě 5 vylosovaných. Řešte

- a) Pomocí hypergeometrického rozdělení,
- b) Aproximace na binomického rozdělení,
- c) Kolik čísel uhádneme nejpravděpodobněji.

[0.0071, 0.0085, 1]

Př. 12: Vypočtěte pravděpodobnost, že při výběru 10 karet z 32 karetního balíčku bude právě 8 vyšších karet (spodek, filek, král nebo eso). Balíček obsahuje 4 spodky, 4 filky, 4 krále a 4 esa.

- a) Řešte pomocí hypergeometrického rozdělení, za předpokladu, že karty nevracíte.
- b) Řešte pomocí binomického rozdělení, za předpokladu, že karty vracíte.
- c) Odůvodněte rozdíl mezi výše uvedenými výsledky.

[0.0239, 0.0439]

Př. 12a: V osudí je 10000 bílých a 7000 černých koulí. Losujete z nich 30 koulí. Jaká je pravděpodobnost, že vylosujete právě 22 bílých a 8 černých. Odůvodněte prakticky shodnost výsledků.

[0.041118, 0.041180]

4.4 Geometrické rozdělení

Př. 13: Házíte kostkou. Určete pravděpodobnost, že právě u pátého hodu Vám padne poprvé šestka.

[0.0804]

Př. 14: Dva hráči střídavě házejí kostkou. Vyhrává ten, kdo první hodí šestku. Jaká je pravděpodobnost, že vyhraje ten, který začínal.

[0.5455, 0.4545]

Př. 15: Distributor prodává knihu. 10 % knihkupců ji zakoupí. Jaká je pravděpodobnost, že distributor bude poprvé úspěšný:

- a) Právě u 5 návštěvy knihkupectví,
- b) Do 5 návštěvy (5. již neuvažujeme),
- c) Při osmé a více návštěvě.

[0.0656, 0.3439, 0.4783]

4.5 Negativně binomické rozdělení

Př. 15a: Pravděpodobnost výskytu krevní skupiny A+ je 0.35. V nemocnici potřebují najít 3 dárce s touto krevní skupinou. Dárcové však neznají svojí krevní skupinu. Jaká je pravděpodobnost, že pro nalezení právě 3. dárce s krevní skupinou A+ budou muset vyšetřit:

- a) právě 10 dárců,
- b) více jak 9 dárců,
- c) aspoň 6 (včetně) a nejvýše 10 dárců (včetně).

Obdobný příklad, ale chceme vědět, že mezi 10 dárci budou právě 3 s krevní skupinou A+.

[0.0757, 0.3373, 0.5032, 0.2522]

4.6 Multinomické rozdělení

Př. 15b: Máte balíček mariášových karet. 10x losujete kartu, kterou následně vracíte do balíčku. Určete pravděpodobnost, že:

- a) Vylosujete 8 karet a to buď spodky, filky, krále nebo esa.
- b) Vylosujete právě 2 esa, 3 krále, 2 filky a 1 spodka.
- c) Vypočtěte příklad ad a) a ad b), jestliže karty nevracíte zpět.

[0.0439, 0.00112, 0.0239, 0.00107]

4.7 Poissonovo rozdělení

Př. 16: Na 100 metrech látky se nachází 10 kazů. Jestliže vybereme 20 metrový úsek látky, jaká je pravděpodobnost, že zde

- a) není žádný kaz,
- b) jsou zde právě 2 kazy,
- c) je zde více než 5 kazů.

[0.1353, 0.2706, 0.0165]

Př. 17: Při sledování poruchovosti provozu se zjistilo, že za 1 rok zde bylo na 10 strojích zaznamenáno 5 poruch. Určete pravděpodobnost, že v následujících 2 letech bude na 25 strojích zaznamenáno:

- a) Méně než 12 poruch,
- b) Právě 20 poruch,
- c) Více než 25 poruch.

[0.0014, 0.0519, 0.4471]

Př. 18: Průměrný telefonní hovor trvá 1,5 min. Dochází-li průměrně k 600 hovorům za hodinu, jaká je pravděpodobnost, že se bude současně konat více než 30 hovorů.

[0.000 197]

Př. 19: Průměrný telefonní hovor trvá 1,5 min. Kolik linek musí ústředna mít, dochází-li průměrně ke 240 hovorům za hodinu a pravděpodobnost ztráty volání nesmí překročit 0,01.

[12]

4.8 Aproximace binomického a hypergeometrického rozdělení na Poissonovo

Př. 20: Tisíckrát se hodilo mincí. Jaká je pravděpodobnost, že mezi 480x až 520x padne orel. Proveďte výpočet pomocí:

- a) Binomického rozdělení,
- b) Poissonova rozdělení,
- c) Proč nejsou splněny předpoklady převedení binomického rozdělení na poissonovo?

[0.8052, 0.6408]

Př. 21: Ve sportce se táhne 6 čísel ze 49. Sázíme 6 čísel. Jaká je pravděpodobnost, že jsme uholili právě 2 čísla.

- a) Řešte pomocí hypergeometrického rozdělení.
- b) Řešte pomocí approximace na binomické rozdělení.
- c) Řešte pomocí approximace na Poissonovo rozdělení.

[0.1324, 0.1334, 0.1295]

Př. 22: Korektura tisíce stran textu prokázala 1500 chyb. Určete pravděpodobnost, že na náhodně vybrané stránce se nachází 4 chyby. Odhadněte, kolik stran v tisícistránkové knize bude bez chyby.

[0.0471, 223]

Př. 23: Bankovní úředník zjistil, že u 20 % návrhů na půjčku zákazníci zamlčí důležité informace. Určete pravděpodobnost, že mezi 100 návrhy budou 25 se zamlčenými informacemi.

- a) Řešte pomocí binomického rozdělení
- b) Řešte pomocí approximace na Poissonovo rozdělení

[0.0439, 0.0446]

Př. 24: Řešte obdobný příklad 23 se změněnými parametry. Bankovní úředník zjistil, že u 50 % návrhů na půjčku zákazníci zamlčí důležité informace. Určete pravděpodobnost, že mezi 10 návrhy bude 5 se zamlčenými informacemi.

- a) Řešte pomocí binomického rozdělení
- b) Řešte pomocí approximace na Poissonovo rozdělení
- c) Jaké předpoklady approximace nejsou splněny.

[0.2461, 0.1755]

5 Spojitá rozdělení pravděpodobnosti

5.1 Rovnoměrné rozdělení

Př. 6: Funkce náhodné číslo generuje data z rovnoměrného rozdělení s parametry $a=0$, $b=1$. Transformujte tato data tak, aby $a=10$ a $b=15$. Tj. byla rovnoměrně rozdělena mezi $<10,15>$.

Př. 7: Ověřte výpočtem správnost střední hodnoty, rozptylu a směrodatné odchylky rovnoměrného rozdělení. Zjistěte dále šikmost a špičatost.

$$[\frac{1}{2}(a+b), \frac{1}{12}(b-a)^2, 0, -\frac{6}{5}]$$

Př. 8: Náhodná veličina X má rovnoměrné rozdělení. Jaké jsou parametry a a b , jestliže jste zjistily z dat, že střední hodnota výběru je 1 a rozptyl je 3.

$$[\text{řeší se soustava rovnic: } \frac{1}{2}(a+b), \frac{1}{12}(b-a)^2 = 3; a = 4, b = -2]$$

Př. 9: Uveďte příklady, kde data jsou z rovnoměrného rozdělení.

5.2 Exponenciální rozdělení

Př. 10: Uveďte příklady, kde data jsou z exponenciálního rozdělení.

Př. 11: Zkuste vysvětlit, co představuje intenzita náhodného jevu (u tohoto rozdělení je konstantní).

Př. 12: Doba do poruchy zařízení lze popsat exponenciálním rozdělením. Data o poruchách máte uvedeny v souboru P0512.mat. Vypočtěte parametry exponenciálního rozdělení a střední dobu do poruchy (střední hodnota rozdělení).

$$[\lambda = \frac{1}{497}; 497 \text{ h}]$$

Př. 13: Doba do poruchy zařízení je popsána exponenciálním rozdělením, kde střední doba do poruchy $E(T)=2000$ h. Vygenerujte 10, 100 a 1000 dat z tohoto rozdělení a vypočtěte z dat střední hodnotu a rozptyl. Všimněte si, že pravděpodobně 1000 vstupních dat bude mít nejblíže k střední hodnotě 2000 h.

Pro vektor 1000 dat vypočtěte medián. Všimněte si, že medián není roven střední hodnotě.

Př. 14: Vypočtěte z hustoty pravděpodobnosti střední hodnotu, rozptyl, směrodatnou odchylku, šikmost a špičatost.

$$[\frac{1}{\lambda}, \frac{1}{\lambda^2}, \frac{1}{\lambda}, 2, 6]$$

Př. 15: Doba opravy má exponenciální rozdělení. Určete střední dobu opravy, jestliže do 60 minut je opraveno 30 % výrobků.

$$[\lambda=0.00594, EX=168 \text{ h}]$$

Př. 16: Doba do poruchy zařízení má exponenciální rozdělení s parametrem λ . Jaká je pravděpodobnost, že mezi dvěma po sobě jdoucími poruchami uběhne alespoň $\frac{3}{\lambda}$ hodin.

[0.0498]

Př. 17: Výrobek má střední dobu do poruchy 3 roky. Jaká je pravděpodobnost, že se porouchá v záruce, tj. v prvních dvou letech provozu.

[48.66 %]

Př. 17a: V souboru P0517a.xlsx máte data o poruchách komponenty. Data jsou v prvním sloupci. Určete:

- a) Parametry exponenciálního rozdělení
- b) Zjistěte pravděpodobnost, že se komponenta porouchá do 10000 hodin
- c) Zjistěte pravděpodobnost, že na výrobku budou dvě poruchy komponenty za jeden rok, tj. 8640 h.

[5954 h, 81.35 %, 24.66 %]

Př. 17b: Testuje se životnost 100 výrobců. Doba zkoušky trvá 10000 hodin. Celkem bylo zaznamenáno 79 poruch, viz soubor P0517b.xlsx. Výrobky po poruše nejsou nahrazeny novými. V čase 10000 hodin je zkouška ukončena (21 výrobců). Určete parametry exponenciálního rozdělení.

[6354 h]

5.3 Weibullovo rozdělení

Př. 18: Uveďte příklady, kde data jsou popsána Weibullovým rozdělením

Př. 19: Nakreslete graf distribuční funkce, kde parametr $\alpha = 1$ a parametr β bude po řadě 0.8; 1, 1.5, 2, 2.5, 3 a 10. Zdůvodněte výsledky.

Př. 20: Poruchovost degradujícího zařízení je popsána Weibullovým rozdělením s parametry $\alpha = 3$ roky a parametr $\beta = 1.5$. Určete střední dobu do poruchy (střední hodnota) zařízení. A určete pravděpodobnost, že se zařízení porouchá v době záruky, tj. do dvou let.

[2.708, 41.97 %]

Př. 21: Zjistěte z dat o poruchovosti výrobku, které jsou uloženy v souboru P0521.mat, parametry Weibullovova rozdělení.

[a=269, b=5.46]

Př. 21a: Zjistěte z dat o poruchovosti výrobku, které jsou uloženy v souboru P0521a.mat, parametry Weibullovova rozdělení.

[a=242, b=1.09]

Př. 21b: Příklad vychází z P0517b.

Testuje se životnost 100 výrobků. Doba zkoušky trvá 10000 hodin. Celkem bylo zaznamenáno 79 poruch. Výrobky po poruše nejsou nahrazeny novými. V čase 10000 hodin je zkouška ukončena (21 výrobků). Určete parametry Weibullovova rozdělení. Vstupní data viz soubor P0521b.xlsx

[a=6336, b=0.953]

5.4 Normální rozdělení

Př. 22: Nakreslete graf, kde budou vyneseny hustoty pravděpodobnosti z normálního rozdělení s následujícími parametry:

- a) $\mu(X) = 0; \sigma^2 = 1$
- b) $\mu(X) = 0; \sigma^2 = 4$
- c) $\mu(X) = 4; \sigma^2 = 1$
- d) $\mu(X) = 4; \sigma^2 = 4$

Rozsah vodorovné osy volte $<-10, 10>$. Uvědomte si, jakým způsobem se odlišují jednotlivé příklady od základního uvedeného v bodu a.

Př. 23: Jak se odlišuje normované normální rozdělení od obecného normálního rozdělení. Uveďte transformační vztah, abyste obdrželi normované normální rozdělení.

$$[N(\mu, \sigma^2), N(0, 1), z = \frac{x - \mu}{\sigma}]$$

Př. 24: Máte normální rozdělení s parametry $N(\mu = 5; \sigma^2 = 4)$. Vypočtěte následující hodnoty:

- a) 20% kvantil
- b) 50% kvantil
- c) Ze znalosti výsledku z bodu a) z paměti 80% kvantil
- d) $F(x = 3.5)$
- e) $F(x = 8)$
- f) Z paměti ze znalosti výsledku z bodu d) $F(x = 6.5)$

[3.317, 5, 6.683, 0.2266, 0.9332, 0.7734]

Př. 25: Délka výrobku v mm má $N(\mu = 50 \text{ mm}; \sigma^2 = 0.49 \text{ mm}^2)$. Určete pravděpodobnost, že rozměr výrobku bude mezi 49 a 51 mm.

[0.8469]

Př. 26: Výsledky měření jsou zatíženy jen normálně rozdělenou náhodnou chybou se směrodatnou odchylkou 3 mm.

- a) Jaká je pravděpodobnost, že při měření bude chyba v intervalu (-2 mm, 5 mm).
- b) Máte 3 výrobky, jaká je pravděpodobnost, že alespoň u jednoho výrobku bude chyba mimo tento interval.

[0.6997, 0.6574]

Př. 27: Výsledky radarového měření jsou zatíženy normálně rozdelenou náhodnou chybou s nulovou střední hodnotou, která s pravděpodobností 0.95 nepřesahuje ± 20 m. Určete směrodatnou odchylku měření.

[10.2043 m]

Př. 30: X je náhodná veličina s rozdelením $N(\mu = 10; \sigma^2 = 20)$. Jak velké musí být číslo x, aby náhodná veličina nabyla hodnoty z intervalu (3, x) s pravděpodobností 25 %.

[7.7668]

Př. 31: Pravděpodobnost, že náhodná veličina nabude vyšší hodnoty než 59.6 je 0.2119. Pravděpodobnost, že nabude hodnoty menší než 57.2 je 0.7258. Náhodná veličina je z normálního rozdelení. Vypočtěte hodnoty parametrů.

[$\mu = 49.98, \sigma = 12.01$]

Př. 32: Nalezněte 1, 5, 10, 50, 90, 95 a 99% kvantil normálního rozdelení s parametry $N(\mu = 10; \sigma^2 = 80)$. Určete pravděpodobnost, že náhodná veličina bude záporná.

[-10.808, -4.712, -1.4625, 10, 21.4625, 24.712, 30.808, 0.1318]

Př. 33: Máte naměřená data v souboru P0533.mat. Odstraňte data, která jsou odlehlá a následně zjistěte parametry normálního rozdělení.

[10.38, 41.85]

Př. 34: Jaká je pravděpodobnost, že po 200 hodinách provozu budou fungovat alespoň 3 výrobky z 5, jestliže doba do poruchy v hodinách je popsána $N(\mu = 180 \text{ hodin}; \sigma^2 = 400 \text{ hodin}^2)$.

[3.1 %]

5.5 Logaritmicko - normální rozdělení

Př. 36: Nechť X je náhodná veličina s logaritmicko-normálním rozdelením s parametry $\mu = 3; \sigma^2 = 16$. Vypočtěte pravděpodobnost, že data jsou v intervalu <2,4>. Vykreslete hustotu pravděpodobnosti. Zvolte maximum 5 a krok 0.001. Všimněte si nesymetričnosti dat.

[0.0613]

5.6 Přesnost statistických charakteristik kvantitativních proměnných

Př. 37: Vygenerujte 100 dat z logaritmicko-normálního rozdělení s parametry $\mu = 4; \sigma^2 = 5$. Vytvořte z nich krabicový graf, dale pomocí Weibullovou a normálního papíru ověřte, zda byste mohli použít dané rozdělení.

Př. 38: Pro data uložená v souboru P0538.mat vytvořte krabicový graf. Co představují jednotlivé čáry v grafu.

Př. 29: Vygenerujte 1000 dat z následujících rozdělení a vytvořte z každého z nich krabicový graf. Následně ověřte pro data z bodu d), že data pochází z Weibullovou nebo normálního rozdělení podle příslušného papíru.

- a) Exponenciální rozdělení se střední hodnotou 100 hodin
- b) Weibullovo rozdělení s parametrem $\alpha = 100, \beta = 1.5$
- c) Weibullovo rozdělení s parametrem $\alpha = 100, \beta = 3$
- d) Normální rozdělení s parametry $N(\mu = 100; \sigma^2 = 900)$
- e) Vyneste vygenerovaná data z bodů a až d do jednoho grafu (sloupce 1 až 4)

Výsledky porovnejte z pohledu rozptylu a šiknosti.

Př. 39: Bylo testováno, zda se rozměry výrobků mění v závislosti na intervalu mezi seřízením stroje. Intervaly mezi seřízením stroje byly 1, 2, 3 a 4 dny. Zkuste analyzovat data pomocí krabicového grafu a odhadněte z charakteru výsledků, zda dochází k posunu přesnosti rozměrů (změna středních hodnot) a k rozptylu přesnosti dat. Data jsou uložena v souboru P0539.mat.

[střední hodnota se zvyšuje, stejně tak se opticky zvyšuje rozptyl] neboli lze interpretovat

[stroj je nutno dříve setřídit, protože rozptyl se opticky zvyšuje.]

6 Výběrové charakteristiky

6.1 Výběrové charakteristiky

Př. 1:

- a) Vygenerujte 10000 náhodných čísel z rovnoměrného rozdělení $<0,1>$ a vytvořte histogram o 100 sloupcích, které vynesete do grafu 1.
- b) Dále vygenerujte 10000 dvojic náhodných čísel, které sečtete a vydělíte dvěma. Opět vyneste histogram o 100 sloupcích do grafu 2.
- c) Obdobně vygenerujte 10000 pětic náhodných čísel, které sečtete a vydělíte pěti. Opět vyneste histogram o 100 sloupcích do grafu 3.
- d) Obdobně vygenerujte 10000 desetic náhodných čísel, které sečtete a vydělíte pěti. Opět vyneste histogram o 100 sloupcích do grafu 4.
- e) Vysvětlete, proč data se „shlukují“ v blízkosti střední hodnoty.

Všimněte si, že 1) se zvyšujícím se počtem vygenerovaných dat se více podobají normálnímu rozdělení, 2) blíží se k průměru a zmenšuje se rozptyl..

Př. 2: Vygenerujte 1000000 desetic náhodných čísel z rovnoměrného rozdělení $<0,1>$. Prvky v desetících sečtěte a vyneste do histogramu o 1000 sloupcích. Zároveň do stejného grafu vyneste hustotu pravděpodobnosti normálního rozdělení s parametry $N(\mu = 5; \sigma^2 = \frac{10}{12})$, kterou vynásobíte 8000. Vysvětlete, proč se obě funkce tvarově relativně dobře překrývají.

Př. 3: Vygenerujte 10000 čísel z normálního rozdělení s parametry $N(\mu = 5; \sigma^2 = 4)$ a vyneste je do grafu ve formě histogramu o 100 sloupcích. Do druhého obdobného grafu vyneste vygenerovaných 10000 čísel z normálního rozdělení s parametry $N(\mu = -5; \sigma^2 = 4)$.

Odhadněte, jaké parametry bude mít rozdělení, jestliže hodnoty sečtete. Ověřte výpočtem správnost Vašeho řešení.

$[N(\mu = 0; \sigma^2 = 8)]$

Př. 3a: Vygenerujte 10000x1000 dat z logaritmicko normálního rozdělení (značně nesymetrické) s parametry $\mu = 3$; $\sigma^2 = 4$. V řádku hodnoty sečtěte a udělejte z nich průměr, který vynesete do histogramu o 100 sloupcích.

Všimněte si, že i průměr značně nesymetrických vstupních dat může se blížit normálnímu rozdělení.

Př. 4: Náhodná veličina A má $E(X)=5$ a $D(X)=4$, náhodná veličina B má $E(X)=3$ a $D(X)=6$, náhodná veličina C má $E(X)=2$ a $D(X)=8$. (Jsou nezávislé.) Vypočtěte střední hodnotu a rozptyl výsledné náhodné veličiny X, která je dána vzorcem $X=A+B+C$ a $Y=A+B-C$.

$$[E(X)=10, D(X)=18; E(Y)=6, D(Y)=18]$$

Př. 5: Dokažte následující tvrzení. Vynásobením naměřených dat se střední hodnotou $E(X)$ a rozptylem $D(X)$ konstantou c, kde $c > 0$, bude střední hodnota naměřených dat rovna $cE(X)$ a rozptyl $c^2D(X)$. Zkuste ověřit na datech.

Př. 6: Náhodná veličina A má $E(X)=5$ a $D(X)=4$, náhodná veličina B má $E(X)=0$ a $D(X)=16$. Vypočtěte střední hodnotu a rozptyl výsledné náhodné veličiny X, která je dána vzorcem $X=A+3*B$.

$$[E(X)=5, D(X)=148]$$

6.2 Centrální limitní věta

Jsou-li X_1, X_2, \dots, X_N , pro N velká, ze stejněho rozdělení s konečným průměrem μ_x a rozptylem σ_x^2 , potom

$$X = X_1 + X_2 + \dots + X_n \sim N(n \cdot \mu_x, n \cdot \sigma_x^2)$$

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right)$$

$$\text{Platí, že } \frac{\bar{X} - \mu_x}{\sigma_x} \sqrt{n} \sim N(0, 1)$$

Př. 7: Máte vygenerováno 1000 náhodných čísel z rovnoměrného rozdělení $<0,1>$. Určete pravděpodobnost, že průměr všech vygenerovaných čísel bude vyšší než 0.520.

$$[0.0142]$$

Př. 8: Životnost komponenty má exponenciální rozdělení se střední hodnotou 5 let. Určete pravděpodobnost, že 100 náhodně vybraných komponent bude mít v průměru životnost nižší než 4 roky.

$$[0.0228]$$

Př. 9: Zatížení letadla s 64 místy nemá překročit 6000 kg. Jaká je pravděpodobnost, že při plném obsazení bude tato hodnota překročena, má-li hmotnost cestujícího střední hodnotu 90 kg a směrodatnou odchylku 10 kg.

$$[0.0013]$$

Př. 10: Počet chyb na jedné straně textu má střední hodnotu 3 a rozptyl 4. Jaká je pravděpodobnost, že na 400 stranách bude méně než 1000 chyb.

[2.867e-07]

Př. 11: Stokrát hodíme šestistěnnou kostkou. Jaká je pravděpodobnost, že součet hodů bude mezi 320 a 380.

[0.9259]

Př. 12: 600 krát hodíme kostkou. Pomocí binomického rozdělení, Poissonova rozdělení a centrální limitní věty určete, jaká je pravděpodobnost, že šestka padne 105 a vícekrát.

[0.3078; 0.3216; 0.3110]

Př. 13: V osudí je 16 bílých a 14 černých koulí. Jaká je pravděpodobnost, že při 150 tazích jedné koule (s vracením) vytáhneme bílou právě 77x.

- a) Řešte pomocí binomického rozdělení.
- b) Řešte pomocí Poissonova rozdělení.
- c) Řešte pomocí centrální limitní věty.

[0.0577, 0.0429, 0.0578]

6.3 Rozdíl výběrových průměrů

Př. 14: Průměrný plat v České republice je 27 000 Kč se směrodatnou odchylkou 8000 Kč. Průměrné náklady na bydlení jsou 7000 Kč se směrodatnou odchylkou 2000 Kč. Vypočtěte pravděpodobnost, že člověku zůstane alespoň 25000 Kč, jestliže z platu odečteme náklady na bydlení.

[0.2721]

Př. 15: V roce 2015 a 2016 probíhal průzkum ohledně měsíčních výdajů za pivo. Zjistěte pravděpodobnost, že v roce 2016 dávají lidé za pivo více než v roce 2015.

x2015=[587,124,651,1212,1074,523,273,800,485,961,1683,2411]

x2016=[121,524,2612,847,1310,1521,951,1000,521,12,190,263,321,587,953]

[0.0061]

Př. 16: Odvodte vzorec pro rozdíl výběrových průměrů: $\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ ~ N(0,1)

Př. 17: Zeptali jsme se 1000 respondentů na oblibu místního cholerickeho politika. Obdrželi jsme kladný výsledek od 168 respondentů. Místní cholerickej politik však říká, že jeho obliba je 55 procent. Určete pravděpodobnost, že jeho tvrzení je pravdivé a jeho obliba je 55 %.

[1.5 E-130, místní cholerickej politik nemluví pravdu]

Př. 18: Zeptali jsme se 1000 respondentů na určitý výrok. 60 % řeklo, že s ním souhlasí. Určete pravděpodobnost, že po zeptání celé společnosti bude výsledek minimálně: 45%, 50%, 55%, 59 %, 60 %, 61 %, 65 %, 70 %.

Porovnejte výsledky mezi sebou. Lze vidět, že výsledné pravděpodobnosti 45 % i 70 % jsou velmi málo pravděpodobné.

[45 %: 1.0000 50 %: 1.0000 55 %: 0.9994 59 %: 0.7407 60%: 0.5000 61 %: 0.2593
65 %: 0.0006 70%: 0.0000]

Př. 19: Jak se změní výsledky z příkladu 18, jestliže se zeptáme pouze 100 lidí?

[45 %: 0.9989 50 %: 0.9794 55 %: 0.8463 59 %: 0.5809 60 %: 0.5000 61 %: 0.4191
65 %: 0.1537 70 %: 0.0206]

Př. 20: V roce 2015 jsme se zeptali 250 respondentů na určitý názor – 62 odpovědělo souhlasně. Obdobně v roce 2016 jsme se zeptali 340 respondentů na stejnou otázku – 141 odpovědělo souhlas. Zjistěte pravděpodobnost, že rozdíl výběrových četností roků 2016 a 2015 je kladný.

[0.9999936]

Př. 21: Řešte př. 20 s následující úpravou. Určete pravděpodobnost, že se zvýšila podpora tohoto názoru minimálně o 10 %.

[0.9596]

6.4 χ^2 rozdělení

Př. 22: Zjistěte 5 a 95% kvantil chí kvadrát rozdělení s 10 stupni volnosti

[3.9403, 18.3070]

Př. 23: Vykreslete graf hustoty pravděpodobnosti pro chí kvadrát rozdělení s 2, 4 a 6 stupni volnosti.

Př. 24: Mějme data z χ^2 rozdělení s 12 stupni volnosti. Určete pravděpodobnost: $P(X > 20)$.

[0.0671]

6.5 Studentovo rozdělení (t-rozdělení)

Př. 25: Určete pravděpodobnost, že Studentovo rozdělení s 2, 4, 10, 100 stupni volnosti nabývá $P(X > 1)$. Určete pravděpodobnost i pro normované normální rozdělení.

[0.2113, 0.187, 0.1704, 0.1599, 0.1587]

Př. 26: Zjistěte 5 a 95% kvantil Studentova rozdělení s 10 stupni volnosti.

[-1.8125, 1.8125]

Př. 27: Vykreslete graf hustoty pravděpodobnosti pro Studentovo rozdělení s 1, 2 a 4 stupni volnosti. Do jednoho grafu nakreslete zároveň hustotu pravděpodobnosti normovaného normálního rozdělení. Uvědomte si, že Studentovo rozdělení konverguje k normovanému normálnímu, jestliže stupeň volnosti se blíží ∞ .

6.6 Fisherovo-Schnedecorovo rozdělení (F rozdělení)

Př. 28: Zjistěte 5 a 95 % kvantil F rozdělení s 10 a 5 stupni volnosti. Dále určete 5 a 95% kvantil F rozdělení s 5 a 10 stupni volnosti. Zkuste odhadnout jaký je mezi výsledky vztah.

[0.3007, 4.7351, 0.2112, 3.3258]

7 Teorie odhadu

7.1 Bodový odhad

Př. 2: Mějte data 0.1; 0.2; 0.3; ... ; 0.8; 0.9 a 1. Vypočtěte vyběrovou střední hodnotu, rozptyl, směrodatnou odchylku a výběrovou šíkmost. Jak se změní tyto charakteristiky, jestliže data budou vynásobena 10.

[střední hodnota=0.55 vs. 5.5; rozptyl=0.0917 vs. 9.17; směrodatná odchylka=0.3028 vs. 3.028; šíkmost=1.7758 je shodná, špičatost 0 a je shodná; střední hodnota a směrodatná odchylka se zvětší 10x, rozptyl 100x, šíkmost a špičatost se nezmění]

7.2 Intervalový odhad střední hodnoty normálního rozdělení

Př. 3: Deset balíčků mouky pocházející z balicího stroje mělo hmotnost v gramech: 987, 1001, 993, 994, 993, 1005, 1007, 999, 995 a 1002. Sestrojte:

- a) 95% interval spolehlivosti pro střední hodnotu,
- b) 90% interval spolehlivosti pro střední hodnotu,
- c) 95% interval spolehlivosti pro minimální hmotnost.

[993.1, 1002.1; 993.98, 1001.2; 993.98, ∞]

Př. 4: Z 12 pozorování doby trvání montážní operace byl zjištěn průměr 44 s a směrodatná odchylka 4 s. Sestrojte 90% interval spolehlivosti pro očekávanou délku operace, jestliže daná operace má normální rozdělení.

Protože nemáte naměřená data, musíte počítat dle vzorců z přednášek.

[41.926 s, 46.074 s]

Př. 5: Naměřili jsme 10 údajů o životnosti žárovky: 380, 402, 408, 412, 454, 459, 472, 481, 491, 502 hodin. Odhadněte, zda data jsou z normálního rozdělení (například pravděpodobnostním papírem) a dale určete 95% intervalový odhad střední hodnoty životnosti žárovky. Určete i 95% jednostranný intervalový odhad pro minimální a maximální odhad střední hodnoty. Interpretujte výsledky.

[Data jsou z normálního rozdělení, 415.4 až 480.7, levostranný 0 až 474.6, pravostranný 421.6 až ∞ .]

Oboustranný: S pravděpodobností 95 % bude střední doba do poruchy v intervalu mezi 415.4 až 480.7 hodin.

Left: s pravděpodobností 95 % bude střední doba do poruchy kratší než 474.6 hodin.

Right: s pravděpodobností 95 % bude střední doba do poruchy delší než 421.6 hodin.]

Př. 6: V prodejně si udělali průzkum, kolik zákazníků přijde do obchodu během jednoho dne. Byly zjištěny následující data:

x=[541,574,585,596,612,618,632,641,654,671,681,692,711,713,718,719,754,796,812,815,835,858];

Ověřte, že data jsou z normálního rozdělení. Zjistěte 99% interval spolehlivosti odhadu střední hodnoty.

[637.6, 746.8 hod]

Př. 7: Automat vyrábí pístové kroužky o daném průměru. Při kontrole kvality bylo náhodně vybráno 80 kroužků a zjištěna střední hodnota průměru 12.01 mm. A dále vypočtena směrodatná odchylka jejich průměru 0.04 mm. Určete 95% oboustranný intervalový odhad střední hodnoty. Uvažujte dva případy a) směrodatná odchylka je definována na 0.04 mm, b) směrodatná odchylka byla vypočtena 0.04 mm. Odůvodněte rozdíl výsledků.

(Předpokládejte, že průměr pístových kroužku lze modelovat pomocí normálního rozdělení.)

[a) $\langle 12.0012, 12.0188 \rangle$, b) $\langle 12.0011, 12.0189 \rangle$]

7.3 Intervalový odhad rozptylu normálního rozdělení

Př. 8: Obdoba zadání z příkladu 1.

Deset balíčků mouky pocházející z balícího stroje mělo hmotnost v gramech: 987, 1001, 993, 994, 993, 1005, 1007, 999, 995 a 1002. Sestrojte

- a) 95% interval spolehlivosti pro rozptyl a směrodatnou odchylku hmotnosti.
- b) 95% jednostranný interval spolehlivosti pro odhad maximální hodnoty rozptylu.

[18.4200, 129.7591; 4.29, 11.39; 0,105.38]

Př. 9: U 100 náhodně vybraných výrobků činila průměrná hmotnost materiálu 150 g a výběrový rozptyl byl 16 g². Sestrojte 95% interval spolehlivosti pro očekávanou hmotnost materiálu a jeho rozptyl.

[$\mu = \langle 149.2, 150.8 \rangle$, $\sigma^2 = \langle 12.33, 21.59 \rangle$]

7.4 Intervalový odhad relativní četnosti

Př. 10: Při provádění průzkumu 400 respondentů uvedlo 12 %, že by volilo Stranu mírného pokroku v mezích zákona. Vypočtěte 95% interval spolehlivosti pro očekávanou relativní četnost. Jak se změní interval spolehlivosti, jestliže se budeme ptát 1600 respondentů.

Zdůvodněte, proč je šířka intervalu u 1600 respondentů poloviční, oproti 400 respondentů.

[$\mu_{400} = \langle 8.82\%, 15.18\% \rangle$, $\mu_{1600} = \langle 10.41\%, 13.59\% \rangle$]

Př. 11: Při kontrole data spotřeby určitého druhu masové konzervy ve skladech produktů masného průmyslu bylo náhodně vybráno 320 z 20 000 konzerv a zjištěno, že 59 z nich má prošlou záruční lhůtu. Stanovte se spolehlivostí 95% intervalový odhad podílu konzerv s prošlou záruční lhůtou. A dále 95 % intervalový odhad počtu konzerv s prošlou záruční lhůtou.

[$p = \langle 0.1419, 0.2269 \rangle$, $n = \langle 2837, 4537 \rangle$]

7.5 Rozsah výběru

Př. 13: Při odhadu volebních výsledků chceme, aby šířka intervalu volebního výsledku mající odhadem 20 % hlasů byla maximálně 2 %. Určete rozsah výběru pro 95% intervalový odhad.

[6146 respondentů]

Př. 14: Ze zadání příkladu 13 proveďte diskuzi, při jaké pravděpodobnosti volebního výsledku strany musí být rozsah největší, a kdy naopak nejmenší. Ověřte grafem výsledek.

[při 50 %]

Př. 14a: Jak velký by měl být rozsah výběru, jestliže chceme, aby 95% intervalový odhad měl šířku intervalu relativní četnosti menší než 0.01.

[38400 respondentů]

7.6 Intervalový odhad mediánu

Př. 15: Životnost výrobku je popsána exponenciálním rozdělením. Byly zjištěny následující data doby do poruchy:

$x=[37, 61, 98, 135, 162, 194, 222, 235, 256, 287, 317, 345, 400, 412, 484, 495, 510, 528, 612, 711, 787, 843, 911, 987, 1014, 1218, 1512]$ hodin.

Opticky ověřte, že data nejsou z normálního rozdělení, ale z exponenciálního. Určete 95% intervalový odhad mediánu.

[<248,576> hodin]

Př. 16: Vygenerujte si 10000 dat z normálního rozdělení s parametry $\mu = 170, \sigma = 50$. Data jsou symetrická kolem střední hodnoty. Vypočítejte 95% odhad střední hodnoty, jestliže víte, že data jsou z normálního rozdělení. A obdobně vypočítejte 95% odhad mediánu, jestliže informaci o typu rozdělení nemáte.

Odhadněte, proč je šířka intervalu přibližně stejná?

[a) přibližně (169.1, 171.1), b) přibližně (169.4, 171.5), vlivem symetričnosti a velkého počtu dat]

[výsledky se mohou mírně lišit, protože se v příkladu generují vstupní data]

7.7 Intervalový odhad parametrů spojitých rozdělení

Př. 16a: Doba do poruchy nedegradujícího výrobku je popsána exponenciálním rozdělením. Prováděla se zkouška 50 výrobků po dobu 1000 hodin. Po poruše nebyly výrobky nahrazovány. Bylo zjištěno 10 poruch, u ostatních 40 výrobků byla zkouška ukončena v čase 1000 hodin. Určete parametr exponenciálního rozdělení a jeho 95% intervalový odhad.

Tporuch=[80,160,240,320,400,560,720,800,900,960]

[střední hodnota je 4514 h, intervalový odhad <2642,9413> h]

Př. 16b: Rozšířené zadání příkladu z kap. 5, př. 21.

Zjistěte z dat o poruchovosti výrobku, které jsou uloženy v souboru P0716b.mat, parametry Weibullovova rozdělení a jejich 95% intervalový odhad.

[a=269, b=5.46; a=<266.2,272.7>, b=<5.22,5.72>]

7.8 Intervalový odhad poměru rozptylů dvou výběrů s normálním rozdelením

Př. 17: Stroj vyrábějící komponenty potřebuje jednou za čas setřídit. Při testování několika výrobků po 20 hodinách nepřetržitého provozu jsme obdrželi následující hodnoty určitého rozměru. Obdobně jsme testovali také po 50 hodinách nepřetržitého provozu.

$x_{20} = [3.96, 4.03, 4.07, 4.12, 4.16, 4.18, 4.20, 4.22, 4.23, 4.24, 4.24, 4.25, 4.29, 4.32, 4.35, 4.38, 4.41, 4.44];$

$x_{50} = [4.02, 4.07, 4.11, 4.16, 4.22, 4.28, 4.32, 4.36, 4.40, 4.42, 4.46, 4.48, 4.51, 4.52, 4.54, 4.58, 4.62, 4.73];$

Velikost rozptylu u stroje ukazuje, zda je třeba stroj setřídit či nikoliv. Zjistěte 99% intervalový odhad podílu rozptylu.

[podíl rozptylů je 0.4141; intervalový odhad podílu rozptylů je $\langle 0.1117, 1.5351 \rangle$]

7.9 Intervalový odhad rozdílu středních hodnot dvou výběrů s normálním rozdelením

Př. 19: V roce 1980 jsme se dostali rychlíkem z Prahy do Brna za $t_{1980} = [243, 251, 257, 257, 259, 261, 263, 265, 284, 293]$ minut. Obdobně v roce 2015 jsme stejnou cestu absolvovali za $t_{2015} = [191, 193, 193, 195, 195, 195, 197, 198, 199, 202, 202, 203, 204, 205, 207, 208]$ minut. Zjistěte 99% intervalový odhad zrychlení cesty.

[52.8, 75.4 minut]

7.10 Intervalový odhad pro rozdíl relativních četností dvou populací

Př. 21: HDD dvou velkých výrobců - DISK a EMEM byly podrobeny zkoušce kvality. HDD obou výrobců jsou baleny po 20 kusech. Ve 40 balících firmy DISK bylo nalezeno 24 vadných HDD, ve 30 balících EMEM bylo nalezeno 14 vadných HDD. Se spolehlivostí 0,95 určete intervalový odhad rozdílu relativních četností (procent) vadných HDD v celkové produkci firem DISK a EMEM.

[-0.0105, 0.0239]

Př. 22: Stranu mírného pokroku v mezích zákona by v roce 2020 volilo 60 z 845 respondentů. Obdobný průzkum proběhl i v roce 2021 s výsledkem: 57 z 541 respondentů. Určete intervalový odhad na hladině významnosti 95 %, o kolik se zvýšila podpora této strany.

[0.0044, 0.0644]

Myslíte si, že se na základě výsledků intervalových odhadů statisticky prokazatelně zvýšila podpora strany mírného pokroku v mezích zákona?

[na hladině významnosti 5 % ano].

8 Testy hypotéz

8.1 Jednovýběrové testy

8.1.1 Jednovýběrový test – test rozptylu normálního rozdělení

[H,P,CI,STATS] = vartest(X,V,alpha,tail)

X – vstupní data

V – hodnota veličiny, se kterou porovnáváme

Alpha – hladina významnosti

Tail – jednostranný (oboustranný) interval – ‘both’, ‘right’, ‘left’

H – výsledná hypotéza

P – p-value

Ci – konfidenční interval

Stats – velikost testovací veličiny a počet stupňů volnosti

Př. 1: Stroj na sáckování smažených brambůrků vyrábí 100 gramové sáčky. Stroj má povolenou maximální směrodatnou odchylku 1.5g na jeden sáček (rozptyl 2.25 g^2). Data jsou uložena v souboru P0801.mat.

Vypočtěte z dat rozptyl. Uveďte hypotézu H₀ a H₁. Otestujte na hladině významnosti 5 %, zda je tato směrodatná odchylka splněna.

[Hypotézu H₀, že rozptyl je menší nebo roven 2.25 g^2 přijímáme, pvalue=0.3798]

Př. 1a: Uvědomte si souvislost mezi testováním hypotéz a intervaly spolehlivosti.

[Interval spolehlivosti je mezi <1.875,∞>. Jestliže bychom měli hypotézu H₀: $\sigma^2 \leq 1.875$, potom pval=0.05. Jestliže bychom měli hypotézu $\sigma^2 \leq 1.8$, potom hypotézu H₀ zamítáme.]

Př. 2: Pro bavlněnou přízi je předepsána horní mez rozptylu pevnosti vlákna (data jsou z normálního rozdělení), která nemá překročit $H_0: \sigma^2 \leq 0.6$. Otestujte velikost rozptylu na 5% hladině významnosti?

Při zkoušce 16 vzorků byly zjištěny výsledky: x=[2.22, 3.54, 2.37, 1.66, 4.74, 4.82, 3.21, 5.44, 3.23, 4.79, 4.85, 4.05, 3.48, 3.89, 4.90, 5.37]

[hypotézu H₀ zamítáme, pvalue=0.0038]

[H₀ bychom přijali, jestliže bychom testovali: $H_0: \sigma^2 \leq$ více než 0.8086.]

Př. 2a: Vypočtěte příklad 2 pomocí vzorců.

[vypočteme hodnotu testovací statistiky, porovnáme s hodnotou chí kvadrát rozdělení]

Př. 3: Na hladině významnosti 5 % otestujte, zda je směrodatná odchylka $\sigma = 300$. Laborant Vám však nepředal naměřená data, ale ve snaze ušetřit Vám práci pouze vzorků $n=25$, výběrovou střední hodnotu $\bar{x} = 3118$ a výběrovou směrodatnou odchylku $s = 357$.

Poznámka: nutno počítat podle vzorců.

[H_0 přijímáme, pvalue=0.1698]

8.1.2 Jednovýběrový test – test střední hodnoty normálního rozdělení

[H,P,CI,STATS] = ttest(X,V,alpha,tail)

[H,P,CI,STATS] = ztest(X,střední hodnota,směrodatná odchylka,alpha,tail)

rozptyl je předem definován – extrémně vzácný případ

Př. 5: Spotřeba téhož auta byla testována u 11 řidičů s výsledky

Spotreba=[8.8, 8.9, 9.0, 8.7, 9.3, 9.0, 8.7, 8.8, 9.4, 8.6, 8.9] (l/100 km). Lze přijmout hypotézu danou výrobcem, že spotřeba je rovna 8.8 l/100 km? Lze na hladině významnosti 5 % přijmout tvrzení, že rozptyl spotřeby je 0,1?

[střední hodnota: hypotézu H_0 na hladině významnosti 5 % přijímáme, pvalue=0.1455.]

[rozptyl: hypotézu H_0 na hladině významnosti 5 % přijímáme, pvalue 0.3976]

Př. 6: Při kontrole životnosti 50 výrobků bylo z dat zjištěno, že střední doba do poruchy výrobcu je 27400 hodin a směrodatná odchylka 5400 hodin (popsáno normálním rozdělením). Určete na hladině významnosti 5 %, zda lze přijmout fakt výrobce, že střední doba do poruchy je rovna 30 000 hodin.

[H_0 nepřijímáme,pvalue=0.0013]

8.1.3 Párový test

Př. 7: Mějme následující data, kde první řádek představuje hodnotu parametru před tepelnou úpravou (vzorek 1, řádek 1) a v druhém řádku jsou uvedeny výsledky na stejných kusech po tepelné úpravě (vzorek 2, řádek 2). Data jsou z normálního rozdělení. Zjistěte na hladině významnosti 5 %, zda:

- A) Je shodná hodnota parametru u obou výběrů $(H_0: \mu_1 = \mu_2, H_1: \mu_1 \neq \mu_2)$
B) došlo ke zvýšení parametru po tepelné úpravě. $(H_0: \mu_1 \geq \mu_2, H_1: \mu_1 < \mu_2)$

x=[35.0,36.0,36.3,36.8,37.2,37.6,38.3,39.1,39.3,39.6,39.8;

37.2,38.1,38.2,37.9,37.6,38.3,39.2,39.4,39.7,39.9,39.9];

[a) Hypotézu H_0 o shodě parametrů na hladině významnosti 5% zamítáme , pval=0.0024

b) Hypotézu H_0 na hladině významnosti 5 % zamítáme, pval=0.0012. Prokázali jsme na hladině významnosti 5 % vliv tepelné úpravy.]

8.1.4 Znaménkový test

Př. 8: Mějme data: $x=[-6,-3,-1,0,2,3,5,6,7,8,9,11,12,14,15,18,22,28,32,37,41]$. Otestujte na hladině významnosti 5 % znaménkovým testem, zda medián je roven 25.

Pozn: Nepředpokládáme, že data jsou z normálního rozdělení

[Na hladině významnosti 5 % zamítáme H_0 , že median je roven 25, $pval=0.0072$].

Př. 9: Mějme data z příkladu 5 o spotřebě auta, kdy byla testována spotřeba u 11 řidičů. Otestujte na hladině významnosti 5 % (předpokládáte, že data nemusí pocházet z normálního rozdělení), zda medián spotřeby může být 8.8. Porovnejte výsledky s příkladem 5.

Spotreba=[8.8, 8.9, 9.0, 8.7, 9.3, 9.0, 8.7, 8.8, 9.4, 8.6, 8.9]

[Hypotézu H_0 na hladině významnosti 5 % nezamítáme, $pval=0.5078$]

8.1.5 Kvantilový test

Př. 10: Mějme data: $x=[2,3,4,5,6,7,7,8,8,9,11,12,13,15,16,18,19,22,25,28,31,34,37,39,42,45,48]$. Otestujte na hladině významnosti 1 %, zda dolní kvartil může být 3.5. Vykreslete hypotetickou distribuční funkci.

[H_0 , $pval=0.0415$]

Př. 11: Mějme data $x=[2,3,4,5,6,7,7,8,8,9,11,12,13,15,16,18,19,22,25,28,31,34,37,39,42,45,48]$. Otestujte na hladině významnosti 5 %, zda 10% kvantil může být 1.5 (tj. menší než minimální hodnota).

[H_0 , $pval=0.1163$].

Př. 12: Byla zkoumána životnost 50 silně namáhaných výrobků, životnost nelze popsat žádným jednoduchým rozdělením. Otestujte znaménkovým testem na hladině významnosti 5 %, zda medián životnosti je 220 hodin.

Data životnosti jsou následující (soubor P0812.mat).

Data pro ukázkou: $x=[12, 15, 24, 32, 63, 69, 75, 87, 95, 121, 154, 159, 162, 187, 191, 201, 212, 218, 223, 241, 246, 249, 253, 259, 263, 269, 273, 291, 312, 313, 318, 323, 352, 356, 361, 368, 369, 371, 395, 521, 523, 561, 785, 800, 823, 837, 844, 954, 991, 1023]$;

[H_0 na hladině 5 % nezamítáme, $pvalue=0.0649$]

Př. 12a: Vypočtěte data z příkladu 12 pomocí vzorců. Ověřte správnost výsledků.

8.1.6 Jednovýběrový Wilcoxonův test

$P = \text{signrank}(X, M)$

Př. 13: Byla zkoumána životnost 50 silně namáhaných výrobků, životnost nelze popsat žádným jednoduchým rozdělením. Otestujte Wilcoxonovým testem, zda medián životnosti je 220 hodin.

Data životnosti jsou následující (soubor P0812.mat).

x=[12, 15, 24, 32, 63, 69, 75, 87, 95, 121, 154, 159, 162, 187, 191, 201, 212, 218, 223, 241, 246, 249, 253, 259, 263, 269, 273, 291, 312, 313, 318, 323, 352, 356, 361, 368, 369, 371, 395, 521, 523, 561, 785, 800, 823, 837, 844, 954, 991, 1023];

Pozn. U Wilcoxonova testu je předpoklad symetrie dat, která zde není bez odstranění odlehlych hodnot splnena. Proto lepsi je test znaménkový. Například pomocí krabicového grafu, nebo výpočtem šikmosti se lze přesvědčit, že data nejsou symetrická.

[H0 na hladině významnosti 5 % zamítame, pvalue=0.0171]

Př. 14: Zdůvodněte, proč Wilcoxonův test má vyšší váhu než znaménkový test.

[Wilcoxonův test uvažuje rozdíly od střední hodnoty, znaménkový pouze pořadí. Z důvodu rozdílů od střední hodnoty je nutný předpoklad symetričnosti dat]

8.1.7 Test relativní četnosti

Př. 15: Při průzkumu bylo zjištěno, že 82 lidí z 1000 by volilo Stranu mírného pokroku v mezích zákona. Strana vyhlašuje, že by jí volilo 15 % lidí. Lze na hladině významnosti 5 % její tvrzení potvrdit?

[H1, testovací veličina T=-6.022]

8.1.8 Testování parametrů spojitych (nenormálních) rozdelení

Příklady vychází z příkladu 16a a 16b v kapitole 7.

Př. 15a: Doba do poruchy nedegradujícího výrobku je popsána exponenciálním rozdelením. Prováděla se zkouška 50 výrobků po dobu 1000 hodin. Po poruše nebyly výrobky nahrazovány. Bylo zjištěno 10 poruch, u ostatních 40 výrobků byla zkouška ukončena v čase 1000 hodin.

Tporuch=[80,160,240,320,400,560,720,800,900,960]

Určete na hladině významnosti 5 %, zda střední doba do poruchy je rovna 3000 h.

[stredni hodnota je 4514 h, 95% intervalovy odhad je <2642,9413> h. Protože interval obsahuje čas 3000 h, nezamítame na hladině významnosti 5 % hypotézu H0, že střední doba do poruchy je 3000 h.]

Př. 15b: Rozšířené zadání příkladu z kap. 5 př. 21 a z kap. 7 př. 16b.

Data o poruchách výrobků jsou popsány v souboru P0815b.mat. Doba do poruchy je popsána Weibullovým rozdelením. Otestujte na hladině významnosti 5 %, zda parametr degradace $b = 2$.

[95% intervalový odhad parametru $b=<5.22,5.72>$. Testujeme na hladině významnosti 5 %, zda parametr $b=2$. Protože je mimo interval, hypotézu H0 zamítame.]

8.2 Dvouvýběrový test

8.2.1 Dvouvýběrový test - test shody dvou rozptylů

Př. 16: Balicí zařízení je seřízeno na začátku ranní směny a následně kontrolováno u odpolední směny. Byly zjištěny následující hodnoty hmotnosti výrobků:

Ráno=[98.5, 98.6, 98.7, 98.7, 98.7, 98.8, 98.9, 99.2, 99.3, 99.3] g

Odpoledne=[98.1,98.2, 98.3, 98.4, 98.6, 98.7, 98.8, 98.9, 99.0, 99.0] g

Otestujte na hladině významnosti 5%, zda je shodné seřízení stroje, tj. zda rozptyl hmotnosti výrobku je shodný.

[Nezamítáme hypotézu H_0 na hladině významnosti 5 % o shodě rozptylů, pvalue=0.7187]

Př. 17: Otestujte, zda je u následujících dvou vektorů shodný rozptyl.

x=[3,4,5,6,8,9,9,10,11,12,13,13,14,15,15,15,16,16,17]

y=[5,5,6,6,6,7,7,8,8,9,9,10,10,11,13,15]

[H_0 , pvalue=0.1006]

Př. 18. Mějme 20 dat z vektoru A, který má rozptyl 1.35. Mějme 10 dat z vektoru B, který má rozptyl 0.32. Otestujte na hladině významnosti 5 %, zda podíl rozptylů $\frac{A}{B} = 2$.

[H_0 , pval=0.2516]

8.2.2 Dvouvýběrový test – test shody dvou středních hodnot

[H,P,CI,STATS] =ttest2(X,Y,ALPHA,TAILOVARTYPE)

VARTYPE – ‘equal’, ‘unequal’ – rozptyly jsou (nejsou) shodné, test rozptylu předchází testu shody středních hodnot

Př. 19: Jak by dopadl výsledek testování vlivu tepelné úpravy z párového testu (př. 7), jestliže bychom neznali informaci, že testování proběhlo na stejných kusech. Data jsou z normálního rozdělení. Zjistěte na hladině významosti 5 %, zda je shodná hodnota parametru ($H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$).

x=[35.0,36.0,36.3,36.8,37.2,37.6,38.3,39.1,39.3,39.6,39.8;

37.2,38.1,38.2,37.9,37.6,38.3,39.2,39.4,39.7,39.9,39.9];

[H_0 , pval=0.112]

Př. 20: Denní přírůstky váhy selat byly při krmení směsí A: 62, 54, 55, 60, 53, 58 dkg. U směsi B: 52, 56, 50, 49, 51 dkg. Je mezi krmnými směsmi na hladině významnosti 5 % rozdíl?

[H_0 zamítáme, pvalue=0.0217]

Př. 21: U 8 aut byla zjištěna velikost vzorku u předních pneumatik v mm:

Levá pneumatika: 2.8 2.0 3.2 1.9 2.5 2.6 1.7 4.1

Pravá pneumatika: 2.5 2.1 3.0 2.1 2.4 2.4 1.9 3.8

Zjistěte pomocí párového testu, zda dochází k opotřebování pneumatik stejně. Jak se změní výsledky, pokud bychom aplikovali (chybně) test o shodě dvou středních hodnot.

[párový test: H0,pvalue=0.7486 ttest: H0, pvalue=0.8345]

Př. 22: Mějme naměřená data ve vektorech x a y. Otestujte na hladině významnosti shodu středních hodnot. Testu obvykle předchází test shody rozptylů, který také aplikujete.

x=[24,26,27,28,28,28,29,31,32,33];

y=[-21,-5,3,8,14,17,19,21,29,38,46,52,68];

[data jsou z normálního rozdělení, rozptyly nejsou shodné, střední hodnoty H0, pval=0.368]

8.2.3 Dvouvýběrový test - Mannův- Whitneyův test

Př. 23: Mějme naměřená data ve vektorech x a y. Otestujte na hladině významnosti 5 %, zda median je shodný. Testu obvykle předchází testování shody rozptylů. Předpokládáme, že typ rozdělení je shodný (testování se naučíme v kapitole 9).

x=[12,14,16,18,19,19,21,23,25,27,31,35,39,42]

y=[15,18,21,24,27,29,32,35]

[medián je shodný pval=0.707]

8.2.4 Testování relativních četností

Př. 24 Bylo zjištováno, zda ve městě a na vesnici je při odpovědi na určitou otázku shoda v relativní četnosti. Ve městě bylo dotážáno 1240 respondentů a odpovědělo ano 325. Na vesnici bylo dotážáno 741 respondentů a odpovědělo ano 287. Otestujte na hladině významnosti 5 % shodu názoru ve městě a na vesnici.

[Na hladině významnosti 5% se zamítá hypotéza shodnosti názoru ve městě a na vesnici, pval=9.10⁻⁹]

8.3 Vícevýběrové testy

8.3.1 Test shody rozptylů

Př. 25: Výrobní stroj se seřídí začátkem směny. Kontrola výrobků probíhá vždy po jedné hodině a chceme zjistit, zda nedochází k většímu rozptylu určitého rozměru. Otestujte na hladině významnosti 5 %, zda rozptyly dat jsou shodné. Použijte Bartlettův i Leveneův test.

x1=[18,19,19,19,20,21,21,22,22,23,23,24,24,24,25,25,25,26,26,26,27,28];

x2=[17,18,18,19,19,20,21,21,22,22,22,23,23,23,24,24,24,25,25,26,26,27,28,29] ;

x3=[16,17,18,18,18,19,20,20,20,21,21,21,22,23,23,23,24,25,25,26,27,27,28,28,29,31];

x4=[14,15,16,16,17,18,19,20,22,22,22,23,24,25,25,27,27,27,28,28,28,31,31,33,34];

[rozptyly nejsou shodné, a) pval=0.0041, b) pval=0.000216]

8.3.2 ANOVA

Př. 26: Mějme naměřená data z 5 skupin (každá o 100 prvcích), která jsou uložena v souboru P0826.mat. Ověřte předpoklady a zjistěte, zda střední hodnota je u všech výběrů shodná.

[data jsou z normálního rozdělení, shoda rozptylů pval=0.0713, shoda středních hodnot pval=0.5909]

8.3.3 Kruskall Wallisův test

Př. 27: Mějme shodná data jako v příkladě 25. Otestujte pomocí Kruskall Wallisova testu, zda mediány jsou shodné. (Nelze použít anovu, protože shoda rozptylů nebyla prokázána.)

[shoda středních hodnot byla prokázána pval=0.7952]

8.3.4 Mnohonásobné porovnávání

Př. 28: Mějme naměřená data z 5 skupin (každá o 100 prvcích), která jsou uložena v souboru P0828.mat. Ověřte předpoklady a zjistěte, zda střední hodnota je u všech výběrů shodná. Pokud ne, porovnejte skupiny mezi sebou.

(co lze očekávat: data pravděpodobně z normálního rozdělení, shodné rozptyly, rozdílné střední hodnoty)

[boxplot ukazuje na normální rozdělení, shoda rozptylů prokázána pval=0.0791, shoda středních hodnot neprokázána pval=0.0118

Porovnání viz tabulka, shodné hodnoty nemá 3. a 5. výběr.

	1	2	3	4	5
1	OK	OK	OK	OK	OK
2	OK	OK	OK	OK	OK
3	OK	OK	OK	OK	KO
4	OK	OK	OK	OK	OK
5	OK	OK	KO	OK	OK

]

Př. 29: Mějme naměřená data z 5 skupin (každá o 100 prvcích), která jsou uložena v souboru P0829.mat. Ověřte předpoklady a zjistěte, zda střední hodnota je u všech výběrů shodná. Pokud ne, porovnejte skupiny mezi sebou.

(co lze očekávat: data pravděpodobně z normálního rozdělení, neshodné rozptyly, rozdílné střední hodnoty)

[boxplot ukazuje na normální rozdělení, shoda rozptylů zamítnuta pval=8.64E-6, shoda středních hodnot neprokázána pval=6.84E-4

Porovnání viz tabulka, třetí výběr má vyšší střední hodnotu než všechny ostatní.

	1	2	3	4	5
1	OK	OK	KO	OK	OK
2	OK	OK	OK	OK	OK
3	KO	OK	OK	KO	KO
4	OK	OK	KO	OK	OK
5	OK	OK	KO	OK	OK

]

Př. 30: Mějme naměřená data z 5 skupin (každá o 100 prvcích), která jsou uložena v souboru P0830.mat. Ověřte předpoklady a zjistěte, zda střední hodnota je u všech výběrů shodná. Pokud ne, porovnejte skupiny mezi sebou.

(co lze očekávat: data pravděpodobně z normálního rozdělení, neshodné rozptyly, rozdílné střední hodnoty)

[boxplot ukazuje na nesymetričnost prvního výběru, shoda rozptylů zamítnuta pval=1E-47, shoda středních hodnot neprokázána pval=0.037

Porovnání viz tabulka, první výběr má nižší střední hodnotu než druhý a třetí.

	1	2	3	4	5
1	OK	KO	KO	OK	OK
2	KO	OK	OK	OK	OK
3	KO	OK	OK	OK	OK
4	OK	OK	OK	OK	OK
5	OK	OK	OK	OK	OK

]

9 Testy dobré shody

9.1 χ^2 -test ověření relativních četností

Př. 1: Bylo provedeno 50 hodů šestistěnnou kostkou s výsledky: 1 – 11x, 2 – 8x, 3 – 14x, 4 – 5x, 5 – 7x, 6 – 5x. Zjistěte, zda přijmete na hladině významnosti 5 % hypotézu, že pravděpodobnost padnutí každého z čísel je 1/6.

[H0, pval=0.1797]

Př. 2: Bylo provedeno 50 náhodných pokusů s výsledky: 1 – 15x, 2 – 10x, 3 – 10x, 4 – 8x, 5 – 7x. Přičemž očekávaný výskyt je: 1 – 18x, 2 – 14x, 3 – 10x, 4 – 4.5x, 5 – 3.5x. Zjistěte, zda přijmete na hladině významnosti 5 % hypotézu, že naměřená četnost výsledků koresponduje s očekávanými.

[H0, pval=0.0511]

Př. 3: Pan Novák při sezenív hospůdce se svěřil, že si dělá statistiku svého sázení Sportky. Řekl, že při 1000 násobném sázení (tahá se 6 čísel ze 49), ve 385 případech neuhodl ani jedno číslo; 431x jedno číslo, 148x dvě čísla, 29x tři čísla, 5x čtyři čísla a 1x pět čísel. Zjistěte (hladina významnosti 5 %), zda zjištěné výsledky mohou odpovídat realitě. Použijte hypergeometrické i binomické rozdělení.

[hypergeometrické H1, pval=4.14E-5; binomické H1, pval=1.024E-4]

Př. 4: V obchodě sledují počet návštěvníků za hodinu. Byly zjištěny následující výsledky: 0 – 3x, 1 – 10x, 2 – 15x, 3 – 12x, 4 – 17x, 5 – 10x, 6 – 11x, 7 – 9x, 8 – 5x, 9 – 5x, 10 – 4x, více než 10 – 5x. Určete na hladině významnosti 5 %, zda data jsou z Poissonova rozdělení.

[H1, pval=2.308E-5]

Př. 5: Vygenerujte si 100 dat z Poissonova rozdělení s parametrem $\lambda t = 10$. Otestujte těchto 100 dat, zda pocházejí z Poissonova rozdělení. Nyní k datům přidejte dalších 15 (celkem jich bude 115): [14,15,17,18,19,21,22,24,26,27,27,28,32,34,36]. Otestujte na hladině významnosti 5%, zda data pocházejí z Poissonova rozdělení.

[1. Data – pravděpodobně budou H0, 2. Data – pravděpodobně budou již H1.]

Př. 6: V osudí je 10 koulí černých a 20 bílých. Vylosujeme vždy 5 koulí, které vracíme zpět. Několikrát opakujeme stejný pokus. Máme podezření, že losující podvádí do osudí vidí preferuje černé koule. Obdrželi jsme následující výsledky počtu vytažených bílých koulí:

x=[0,1,0,1,0,2,0,1,1,0,0,1,0,1,0,2,0,1,1,0,0,0,0,1,2,1,1,2,1,0,1,2,1,2,3];

Lze předpokládat, že losující podvádí?

[H1,pval=2.6E-27]

χ^2 -test ověření shody s očekávaným rozdělením

Př. 9.7: V souboru P0907.xlsx máte data o poruchách výrobku. Zjistěte na hladině významnosti 5 %, zda data jsou z exponenciálního rozdělení. Ověřte, že data jsou z toho rozdělení i na základě Weibullová papíru (funkce wblplot)

[data nejsou z exponenciálního rozdělení, pval=0.0462; data nejsou z Weibullovova rozdělení, pval=0.0145]

Př. 9.8: Výpočet z dat v souboru P0907.xlsx budeme obměňovat.

- a) Zkuste si nadefinovat jiné hraniční body a zjistěte, jak se změní výsledek analýzy.
 - b) Jaká jsou základní podmínky pro použití testu dobré shody?
 - c) Otestujte data, jestliže budeme předpokládat exponenciální rozdělení s parametrem $\lambda = \frac{1}{mean(x)}$, **0.0009, 0.0008, 0.0007, 0.0006, 0.0005**.
 - d) Otestujte data, jestliže budete předpokládat, že jsou z normálního rozdělení.
 - e) Otestujte data, zda jsou z normálního rozdělení pomocí normálního papíru (funkce normplot)
- [ad a) výsledky pval se budou měnit, při příznivých datech může být přijata i hypotéza H0
ad b) v každé skupině musí být očekáváno minimálně 5 dat
ad c) $H1, \lambda = \frac{1}{mean(x)}$ pval=0.0462, 0.0009 pval=0.0464, 0.0008 pval=0.0174, 0.0007 pval=0.0019, 0.0006 pval=7E-6
ad d) výsledkem bude H1, pval= 1.21E-6]

Př. 9.9: V souboru P0909.mat máte data o poruchách výrobního procesu. Zjistěte, zda data jsou z normálního rozdělení. Zkuste otestovat, zda mohou být data z normálního rozdělení s parametry $\mu = 15, \sigma^2 = 25$.

[data jsou z normálního rozdělení s parametry: $\mu = 13.15, \sigma = 4.58$, H0, pval=0.6397

data nejsou z normálního rozdělení s parametry $\mu = 15, \sigma = 5$, H1, pval=2.71E-5]

Př. 9.10: V souboru P0910.mat máte vygenerováno 100 dat z normálního rozdělení s parametry $\mu = 20, \sigma^2 = 100$. Zbylých 10 vstupních dat jsou odlehlé hodnoty. Zjistěte na hladině významnosti 5 %, zda všechna data pocházejí z normálního rozdělení s parametry $\mu = 20, \sigma^2 = 100$. Pokud nikoliv, odhalte odlehlé hodnoty a testujte znova na normální rozdělení

[H1, pval=5.90E-7; H1, pval=0.0071]

Př. 9.11: V souboru P0911.mat máte vygenerováno 100 dat, o kterých se domníváme, že mohou být z logaritmicko-normálního rozdělení. Ověřte.

[bez použití hraničních bodů bude mít statistika pouze 0 stupňů volnosti. Test nelze použít.

Při použití hraničních bodů [0,10,50,100,200,500,1000,10000] je z lognorm rozdělení H0, pval=0.676]

Př. 9.12: Vygenerujte si 10, 100, 1000 a 10000 náhodných dat z normovaného normálního rozdělení. Ověřte testy na hladině významnosti 5 %, zda skutečně vygenerovaná data jsou z normálního rozdělení.

[10 dat nelze určit pro málo dat, pro větší počet dat malý počet intervalů, doporučil bych udělat další parametr hranice]

Př. 9.13: Vygenerujte 50 dat z normálního rozdělení s parametry $\mu = 20, \sigma^2 = 100$ a dalších 50 dat z normálního rozdělení s parametry $\mu = 30, \sigma^2 = 100$. Otestujte na hladině významnosti 5 %, zda výsledný vektor bude opět z normálního rozdělení.

[pravděpodobně H0]

Př. 9.14: V souboru P0914.mat máte vygenerováno 100 dat o poruchách. Ověřte, že data jsou z Weibullovova rozdělení (rozdělení se používá pro popis doby do poruchy degradujících výrobků). Ověřte, že data jsou z toho rozdělení i na základě Weibullovova papíru (funkce wblplot)

[H0, pval=0.5317, parametry Weibullovova rozdělení: a=990.45, b=1.5124]

9.2 Kolmogorov – Smirnovův jednovýběrový test rozdělení

Př. 9.15: Studenti ve třídě mají následující výšku. Ověřte na hladině významnosti 5 %, zda výška studentů splňuje normální rozdělení. Pokud data splňují normální rozdělení, zjistěte optimální parametry normálního rozdělení.

vyska=[162,167,170,171,172,175,178,179,180,181,182,184,185,187,191,195].

[Kolmogorov test H0,pval=0.9962; Lillieforsuv test H0, pval>0.5]

Př. 9.16: Byla zaznamenána doba do poruchy zařízení. Zjistěte na hladině významnosti 5 %, zda data splňují exponenciální rozdělení. Pokud ano, zjistěte jeho parametry. Pokud nikoliv použijte Weibullovovo rozdělení. Pro optické posouzení použijte Weibullův papír (funkce wblplot).

```
t=[37, 48, 54, 75, 81, 104, 123, 141, 156, 187, 195, 213, 241, 254, 271, 289, 312, 345, 395, 4  
12, 461, 512, 651, 731];  
[exponenciální: H0, pval=0.7231; Weibullovovo: H0, pval=0.842]
```

Př. 9.16a: Máte data o poruchách uvedená v Př. 9.16. Otestujte na hladině významnosti 5 %, zda data jsou z exponenciálního rozdělení s parametrem $\lambda = 0.01, 0.005, 0.00333, 0.002, 0.001$. Nelze použít Lillieforsův test, protože ten používá pouze pro optimální parametry rozdělení.

[0.01 H1 pval=2.1E-5; 0.005 H0 pval=0.127; 0.00333 H0 pval=0.8666, 0.002 H1 pval=0.0418, 0.001 H1 pval=1.8E-6]

Př. 9.17: Vygenerujte postupně 3, 5, 10, 20, 50 a 100 náhodných čísel. Zjistěte, zda vygenerovaná data jsou skutečně z rovnoměrného rozdělení v rozmezí $\langle 0,1 \rangle$. Následně otestujte, zda vygenerovaná data mohou být z rovnoměrného rozdělení v rozmezí $\langle 0.2,1.2 \rangle$. Jaký maximální rozdíl distribučních funkcí (naměřené a teoretické) může být při určitém rozsahu výběru.

[a) Jestliže porovnáváme s daty z rovnoměrného rozdělení $\langle 0,1 \rangle$, tak prakticky vždy vyjde H0, pval je přibližně konstantní

b) autorovi textu vyšlo pval pro 3 data 0.2222, pro 5 dat 0.3332, pro 10 dat 0.1354, pro 20 dat 0.0045, pro 50 dat 0.0015, pro 100 dat 0.000 072]

Př. 9.18: V souboru P0918.mat máte vygenerována data. Ověřte, zda jsou z normálního rozdělení s parametry $\mu = 10$, $\sigma^2 = 9$. Pro optické posouzení použijte normální papír (funkce normplot).

[pro přesný test Kolmogorov Smirnov H1, pval asi 10^{-275}]

Při použití Lillieforsova testu se testuje obecné normální rozdělení, pval je menší než 0.001]

Př. 9.19: Máte vektor vstupních hodnot: $t=[37,54,81,123,156,213,254,289,345,512,731]$. Otestujte, zda data pochází z normálního rozdělení. Vykreslete distribuční funkci teoretickou i zjištěnou z dat.

[H_0 , pval>0.5]

Př. 9.20: Máte vektor vstupních hodnot: $x=[24,35,61,87,120,151,187,214,341,541,653,1213,2421]$. Zkuste zjistit zda data pochází z logaritmicko-normálního rozdělení.

[Lillietest H_0 pval>0.5 KS test Lognormal H_0 pval=0.9966 KS test normal H_0 pval=0.9966]

Př. 9.21: Zkuste vysvětlit princip testu dobré shody a Kolmogorova-Smirnovova testu. Jaký je rozdíl mezi Kolmogorov-Smirnovovým testem a Lillieforsovým testem a proč má normální rozdělení méně přísné požadavky.

9.3 Kolmogorov – Smirnovův dvouvýběrový test shody rozdělení

Př. 9.22: Máte následující data: $x=[3,5,9,12,15,17,21,24]$ a $y=[5,8,12,12,15,17,19,24,25,28]$. Ověřte, zda data jsou ze shodného rozdělení. Ověřte, zda data z vektoru x a y jsou z normálního rozdělení.

[H_0 , pval=0.9854]

Př. 9.23: Máte následující data: $x=[31,36,42,48,52,57]$ a $y=[15,18,22,27,29,34,35,38,43,49,52]$. Ověřte, zda data jsou ze shodného rozdělení.

[H_0 , pval=0.2615]

Př. 9.24: Vygenerujte 20 dat z normálního rozdělení s parametry $\mu = 0.5$, $\sigma = 0.25$, obdobně vygenerujte 20 dat z rovnoměrného rozdělení $\langle 0,1 \rangle$. Ověřte, zda data mohou být ze stejného rozdělení.

Obdobně příklad vypočítejte pro 200 dat. Interpretujte rozdíl výsledků.

[pro 20 dat – H_0 , pval mne výšlo 0.4973; pro 200 dat – H_1 , pval mne výšlo 0.012]

Př. 9.25: Ověřte, zda data ze souboru P0925.xlsx (ve sloupci A a B) pochází ze stejného rozdělení.

[H_1 , pval=4.1E-11]

Př. 9.26: Máte plat za měsíc leden u několika zaměstnanců v Praze a v Liberci. Ověřte Mann-Whitneyovým testem, že v Praze je vyšší mediánový plat (údaje v tis. Kč).

Zároveň ověřte H_0 , že pracovníci v Praze mají „nižší distribuční funkci“ než v Liberci (to znamená, že v Liberci je nižší plat, doporučuji vykreslit obě distribuční funkce).

Praha=[16,18,18,18,21,23,25,28,31,34,37,41,45,48,48,61]

Liberec=[13,14,14,15,15,16,17,18,23,28,34,36]

[MW test H1 pval=0.003 Na hladině významnosti 5 % zamítáme hypotézu, že v Praze je menší nebo roven medián platu jako v Liberci (neboli zjednodušeně prokázali jsme na hladině významnosti 5 %, že v Praze je vyšší mediánový plat)]

[H1, pval= 0.0151, Na hladině významnosti 5 % zamítáme hypotézu, že v Praze je větší nebo rovna distribuce jako v Liberci (neboli zjednodušeně prokázali jsme na hladině významnosti 5 %, že v Praze je menší distribuce platu než v Liberci)].

Př. 9.27: Pro data:

x=[16,18,18,18,19,19,19,20,20,21,21,21,23,26]

y=[13,14,14,15,15,16,16,16,17,17,18,19,20,21,23,23,23,24,25,28]

ověřte, zdajou ze shodného rozdělení a vyneste jejich distribuční funkce. Následně otestujte, zda každé z rozdělení lze popsat normálním rozdělením.

[H0, pval=0.0689, jsou ze stejného rozdělení Obě jsou z normálního rozdělení pval=0.1862 a 0.1454]

10 Analýza závislostí

Př. 1: V tabulce je zaznamenáno dosažené vzdělání 100 párů snoubenců v den uzavření sňatku. Ověřte na hladině významnosti 5 %, zda existuje závislost mezi vzděláním nevěsty a ženicha.

	Nevěsta základní	Nevěsta středoškolské	Nevěsta vysokoškolské
Ženich základní	24	12	3
Ženich středoškolské	7	24	3
Ženich vysokoškolské	3	9	15

[H1, pval=9.32E-9, Na hladině významnosti 5 % zamítáme hypotézu, že vzdělání nevěsty a ženicha jsou vzájemně nezávislé veličiny.]

Př. 2: Zeptali jsme se 234 středoškoláků, aby uvedli nejoblíbenější sport, který skutečně provozují a který nejradiji sledují v televizi. Otestujte na hladině významnosti 5 %, zda oblíbenost sledování sporů v televizi závisí na oblíbenosti vlastního sportování.

	Sport – fotbal a hokej	Sport - atletika	Sport - gymnastika	-	Sport - plavání
TV – fotbal a hokej	133	6	2	4	
TV – atletika	15	10	4	3	
TV – gymnastika	4	1	25	0	
TV - plavání	9	0	1	17	

[H1, pval=5E-53, Na hladině významnosti 5 % zamítáme hypotézu, že obliba dělaného sportu je nezávislá na oblibě sportu sledovaného v televizi.]

Př. 3: Bylo zjištováno, zda konzumace alkoholu má vliv na odvykání kouření. Ověřte na hladině významnosti 5 %, zda ano, či nikoliv.

	Přestal kouřit	Nepřestal kouřit
Konzumace alkoholu – ano	58	112
Konzumace alkoholu – ne	24	19

[Tmin=0.2076, T=0.41, Tmax=0.8095, Zamítáme hypotézu H0, že konzumace alkoholu nemá vliv na odvykání kouření.]

Př. 4: Vypočtěte příklad 3 pomocí kontingenčních tabulek.

[H1, pval=0.009]

Př. 5: Otestujte na hladině významnosti 5 %, zda je zájem o sport u dětí nezávislý na pohlaví.

	Sportuje	Nesportuje
Chlapci	30	36
Dívky	21	43

[Tmin=0.8372, T=1.7063, Tmax=3.4778, H0]

Př. 6: Máme k dispozici výsledky prvního a druhého zápočtového testu deseti studentů. Vypočtěte korelační koeficient.

1. test 7, 8, 10, 4, 14, 9, 6, 2, 13, 5

2. test 9, 7, 12, 6, 15, 6, 8, 4, 11, 8

[korelace je $r=0.8452$]

Př. 7: Z dat z příkladu 6 otestujte na hladině významnosti 5 %, zda veličiny mohou být nezávislé. Podle výsledků testu normality dat vyberte vhodný test.

[data jsou z normálního rozdělení, zjištěno lillietestem, pval > 0.5

zamítáme hypotézu H_0 o nezávislosti dat, $r=0.8452$, pval=0.021, 95% intervalový odhad korelace je mezi 0.4608 a 0.9626]

Př. 8: V souboru P1008.mat máte uloženy data o průměrných ročních srážkách ze srážkoměrných stanic. V proměnné x za období 1970 až 1980, v proměnné y za roky 2010 až 2020. Otestujte na hladině významnosti 5 %, zda výsledky jsou vzájemně nezávislé.

[data jsou z normálního rozdělení, $r=0.9414$, pval=1E-95, 95% intervalový odhad korelace je <0.9233,0.9554>.]

Př. 9: Máte 1000 přípravků. Na každý z nich se usadí komponenta A a B. Otestujte na hladině významnosti 5 %, zda životnost komponenty A a B jsou vzájemně nezávislé.

Data o životnosti jsou uvedeny v souboru P1009.mat.

[data komponenty A nejsou normálně rozdělené pval<0.001, data komponenty B jsou normálně rozdělené, použiji Spearmanův korelační koeficient; $r=0.0336$, H_0 , pval=0.2887]

11 Úvod do korelační a regresní analýzy

11.1 Lineární regrese

Př. 1: Nalezněte parametry lineární regrese ($y = \alpha x + b$) pro následující data: $x=[3,5,8,11,12,14,15]$; $y=[6,11,15,22,25,27,30]$. Určete parametry a zjistěte, zda parametr b se může rovnat 0. Vysvětlete výslednou tabulku.

[$y = 0.3935 + 1.9595x$; parameter b může být 0, protože $t_{stat} = 0.474$ a $pval=0.65546$]

Př. 2: Nalezněte parametry lineární regrese pro následující data: $x=[2,5,8,11,5,10,6]$; $y=[6,11,15,22,25,27,30]$. Vykreslete data. Je daný model lineární regrese vhodný? Vysvětlete.

[$y = 9.1635 + 1.5288x$; Model vhodný není, protože F -test je $F=2.05$ ($pval=0.212$) a také koeficient determinace je $r^2 = 0.291$]

Př. 3: Nalezněte parametry lineární regrese pro data uložená v souboru P1103.xlsx. V prvním sloupci jsou hodnoty pro vektor x , ve druhém sloupci pro vektor y .

[$y = 0.59591 + 3.0005x$]

Př. 4: Jaké všechny proměnné obsahuje struktura výsledků počítající lineární regresi? Vysvětlete na výsledcích z předchozího příkladu.

Př. 5: Nalezněte parametry kvadratické regrese pro data uložená v souboru P1105.xlsx (obsahuje stejná data jako v příkladě 3). V prvním sloupci jsou hodnoty pro vektor x , ve druhém sloupci pro vektor y .

Je nutný kvadratický člen? Proveďte diskuzi nutnosti každého z parametrů.

[$y = 0.6276 + 2.9931x + 0.0002928x^2$, kvadratický člen není třeba $pval=0.847$]

Př. 6: Nalezněte parametry regresního modelu: $z = \alpha + bx + cy$ pro naměřená data: $x=[2,4,5,6,7,8,9,10]'$; $y=[1,2,3,4,5,6,7,8]'$; $z=[6,11,14,15,18,23,26,31]'$. Vstupní data musejí být sloupcové vektory.

[$z = 2.1429 + 0.57143x + 2.7143y$, nejvyšší $pval$ má proměnná x , proto lze zkusit ji vypustit]

Př. 7: Nalezněte parametry regresního modelu $y = \alpha + bx + cx^2$, pro data uložená v souboru P1107.xlsx. V prvním sloupci jsou hodnoty pro vektor x, ve druhém pro vektor y.

$$[y = 17.398 + 4.4198x + 1.9426x^2]$$

Př. 8: Mějte data: $x=[1,2,3,4,5,6,6.5]; y=[3,5.1,6.9,8.8,10.9,13.3,14.1]$; Proložte data kvadratickým modelem. Určete 95% interval spolehlivosti pro člen αx^2 . V případě, že některý člen není třeba, vypusťte ho, a nahraďte jiným modelem.

[$y = 0.0322x^2 + 1.7803x + 1.2634$, intervalový odhad kvadratického členu je: $(-0.0177, 0.0821)$, není třeba kvadratický člen. Po odstranění kvadratického členu je $y = 2.0261x + 0.9119$.]

Př. 9: Mějme data: $x=[1,2,3,4,5,6,7,8,9,10]; y=[1,2,3,1,2,3,1,2,3,1]; z=[3,9,17,10,16,26,14,25,38,23]$; Proložte data modelem $z = \alpha + bx + cy$ (linear) a dále modelem $z = \alpha + bx + cy + dxy$ (interactions). Co si myslíte o výsledcích modelu?

[$z = -8.533 + 2.5778x + 6.556y; z = -1.9434 + 1.434x + 2.5912y + 0.6824xy$; U druhého (možná i u prvního) modelu je možno absolutní člen vynechat.]

Př. 10: Mějte data z příkladu 9: $x=[1,2,3,4,5,6,7,8,9,10]; y=[1,2,3,1,2,3,1,2,3,1]; z=[3,9,17,10,16,26,14,25,38,23]$; a proložte je modelem:

- a) $z = \alpha + bx + cy + dxy + ex^2 + fy^2$ (quadratic)
- b) $z = \alpha + bx + cy + dx^2 + ey^2$ (purequadratic)

Jsou všechny parametry modelu nutné? Učiňte doporučení.

$$[z = 1.16 + 0.63x + 0.46y + 0.64xy + 0.078x^2 + 0.40y^2]$$

$$z = -3.74 + 1.43x + 3.22y + 0.102x^2 + 0.91y^2$$

Oba modely jsou překombinované, protože pval jsou u všech parametrů vysoké. Je málo vstupních dat, proto doporučuji ještě zjednodušit model. Například odstraněním kvadratických členů.]

Př. 11: Mějme data: $x=[1,2,3,5,7,4,1,2,3,4,1,2,2,2,3]'; y=[2,1,1,4,1,2,5,2,1,2,1,2,1,2,3]'; z=[8,11,14,18,12,18,15,15,15,12,1,11,8,7,5]'$; Proložte je modelem $z = \alpha + bx + cy$ (linear) a dále modelem $z = \alpha + bx + cy + dxy$ (interactions). Co si myslíte o výsledcích modelů?

[$z = 4.8674 + 1.322x + 1.3822y$; Ftest má pval=0.116 – model špatně určen není vhodné použít regresi.

$z = 4.4877 + 1.4711x + 1.5765y - 0.078xy$; Ftest má pval=0.248 – model špatně určen není vhodné použít regresi.]

Př. 11a:

Mějme data: $x=[1,2,3,4,5,6,7,8,9,10]'$; $y=[2,5,8,10,12,15,18,19,21,24]'$. Proložte je lineárním modelem $y = \alpha x + b$ a otestujte na hladině významnosti 5%, zda parametra může být roven 2.
[$y = 0.333 + 2.3758x$; H1 pval=2.53E-4]

11.2 Nelineární regrese

Př. 12: Máte naměřená data v souboru P1112.xlsx. Určete výsledky regresního modelu: a) $y = \alpha$, b)
 $y = \frac{\alpha}{x} + b$. V prvním sloupci jsou hodnoty pro vektor x, ve druhém pro vektor y. Vysvětlete výsledky.

[a) $y=5.174$, b) $y=1.4735/x+4.9344$]

Př. 13: Máte data v souboru P1113.xlsx. V prvním sloupci jsou hodnoty pro vektor x, ve druhém pro vektor y.

- Určete výsledky regresního modelu $y = c * \sin(\alpha x + b)$. Počáteční vektor uvažujte $c=5$, $a=1/3$, $b=1$. Zjistěte koeficient determinace a určete vhodnost modelu.
- Uvažujme stejné zadání jako v a). Ale počáteční vektor uvažujte $c=1$, $a=1$, $b=0$. Zdůvodněte, proč jsou odlišné od předchozích?
- Vykreslete data a proložení do grafu.

[a) $y=4.9581*\sin(0.34398x+0.89692)$, $r^2=0.998$, F pval=1E-61]

[b) $y=1.1501*\sin(0.8986x-0.38972)$, $r^2=0.0387$, F pval=0.521, našel lokální extrém funkcionálu, nikoliv globální]

Př. 15: Máte naměřená data, která jsou uložena v souboru P1115.xlsx. V prvním sloupci jsou hodnoty pro vektor x, ve druhém pro vektor y. Vykreslete data. Zkuste regresní model $y = \alpha x^7 + b x^6 + \dots$. Pokud nejvyšší řád polynomu není třeba, proveděte výpočet bez něj.

[x^7 : model je špatně určen – warning na přeparametrování dat. Parametr b8 je velmi blízký 0 oproti například koeficientu u x^3 .

x^6 : model je špatně určen – warning na přeparametrování dat, pvalue parametru b7 je 0.063 – může být roven 0.

x^5 : model je špatně určen – warning na přeparametrování dat, pvalue parametru b6 je sice 0.007

x^4 : model je špatně určen pvalue b5=0.34919.

x^3 : model je OK, $y = -5.2355 + 0.9407x - 1.9974x^2 + 3x^3$; pvalue parametrů jsou max. 1E-33

Všimněte si, že celková chyba čtverců se u modelu x^3 ... oproti modelu x^7 ... prakticky nezměnila. U x^3 ... je 0.296, zatímco u x^7 ... je 0.283. Při neúměrně vysokém polynomu vznikají nežádoucí oscilace modelu.]

Př. 16: Vygenerujte následujících 100 dat, která vznikla následující rovnicí:

- Parametr i sudý: $y(i) = \left\lceil \frac{x(i)-1}{2} \right\rceil + \left(\left\lceil \frac{x(i)-1}{2} \right\rceil \right)^2$
- Parametr i lichý: $y(i) = \left\lceil \frac{x(i)}{2} \right\rceil - \left(\left\lceil \frac{x(i)}{2} \right\rceil \right)^2$

Funkci lze approximovat funkci $y = \frac{x}{2} \pm \left(\frac{x}{2} \right)^2$. Vykreslete graf. Zkuste vypočítat kvadratickou regresi této funkce a odůvodněte, proč tento model není vhodný.

[Funkce kvadraticky roste a zároveň klesá. Model není vhodný, protože Ftest má pval=0.938. U modelu se zvyšuje s rostoucím x rozptyl, proto nelze použít metodu nejmenších čtverců.]

Př. 17: V souboru P1117.xlsx máte uložena data. První sloupec x, druhý y, třetí z. Zkuste je proložit funkci $z = \frac{a}{x} + \frac{b}{y} + \frac{c}{x+y}$. Počáteční vektor stanovte (2,4,1).

$[z = \frac{-0.055}{x} + \frac{4.47}{y} + \frac{6.69}{x+y};$ první člen není třeba. Po jeho vynechání obdržíme výsledek $z = \frac{4.47}{y} + \frac{6.531}{x+y};]$

Př. 18: Mějte data uložená v souboru P1118.xlsx. První sloupec x, druhý y. Proložte data funkci $y = e^{ax+bx}$. Následně zlogaritmujte $y=\ln(y)$ a proložte lineárním modelem. ($y = e^{ax+bx}$; $\ln y = a + bx$). Uvědomte si, jaká data by ve výsledcích měla být vlivem prosté transformace shodná.

$[y = e^{2.0047+1.0002x}$. Výsledek by měl být shodný, vlivem numerického výpočtu tomu tak není.]

Př. 19: Máte naměřená následující data: $x=[1,2,3,4,5,6,7,8,9,10]; y=[2,4,6,12,15,23,29,38,45,59];$. Vypočtěte parametry regresního modelu $y = a + ax + ax^2$. Lze použít tento model?
[a=0.51715; ano Ftest má pval=2.2E-14, tstat má pval=2.2E-14]

Př. 20: Máte naměřená následující data: $x=[1,2,3,4,5,6,7,8,9,10,1,2,3,4,5,6,7,8,9,10]; y=[1,2,3,4,5,6,7,8,9,10,1,4,9,16,25,36,49,64,81,100]$. Proložte data modelem: $y = bx; y = ax^2 + bx; y = ax^3$. Který z uvedených modelů je nejlepší, a proč. Jaký vzniká problém při proložení těchto dat?

a) $y = 4.43x; r^2 = 0.338; Ftest - pval = 5.11E - 5$

b) $y = 0.5x + 0.5x^2; r^2 = 0.387; Ftest - pval = 1.75E - 4$

c) $y = 0.560x^3; r^2 = 0.386; Ftest - pval = 2.45E - 5$

Ani jeden model není vhodný. Hlavním důvodem je, že data jsou získána ze dvou souborů, které jsou zcela odlišné. Pokud bychom měli naměřena více dat, výsledky pro dané modely budou ještě horší.]

Př. 21: Mějme naměřena data $x=[1,2,3,4,5,6,7,8,9,10]$; $y=[0,5,9,18,28,39,69,111,177,277]$; Proložte data modelem $y = \alpha x^2 + bx + c$. Zjistěte, zda všechny parametry jsou nutné.

$$[y = 5.216x^2 - 30.884x + 42.35; r^2 = 0.975; Ftest pval = 2.62E - 6]$$

Konstantní člen není třeba, protože tstat pval=0.0643, proto hledám ve tvaru $y = \alpha x^2 + bx$

$$y = 3.94x^2 - 14.81x; r^2 = 0.957; Ftest pval = 3.83E - 7]$$

Př. 22. Máte data týkající se složení vody ve vrtech uložena v souboru P1122.csv. První sloupec odpovídá x , druhý sloupec vektoru y . Proložte data modelem $y = \frac{\alpha}{x^b}$. Počáteční vektor volte $\alpha = 0.08, b = 1$. Zjistěte, zda všechny parametry jsou nutné. Proveďte změny počátečního vektoru a porovnejte výsledky.

$$[\alpha = 0.0103, b = 2.054, r^2 = 0.872, \text{všechny parametry jsou nutné, u obou parametrů je } pval=0]$$