Zkouška nanečisto - řešení

18. 5. 2020

Víme, že cena mléka závisí na ceně hovězího (x1) a ceně pozemků na pastvu (x2). Kolik bude očekávaná cena mléka, pro lineární regresní parametry b0=2; b1=0,03; b2=0.5, kdy víme, že hovězí stojí 200 Kč a cena pozemku je 20 Kč. Výsledek zaokrouhlete na 2 desetinná místa.

Důležité je si určit, o jakou regresi jedná. Zde nic nebylo zadáno, protože to lze poznat. Mám 2 závisle proměnné X → vícenásobná regrese

Předpis pro vícenásobnou regresi je:

$$y = b_0 + b_1 x_1 + b_2 x_2$$

tedy po dosazení známých údajů získáme

$$y_p = 2 + 0{,}03 * 200 + 0{,}5 * 20 = 2 + 6 + 10 = 18$$

V případě obdobného zadání, ale pouze s jednou nezávisle proměnou:

- Normalita dat = lineární regrese
- Není normalita dat - Sleduji počet parametrů b
  - 2 (b_0, b_1) = exponenciální regrese
  - 3 a více ($b_0, b_1, b_2, \dots, b_n$) = polynomiální regrese

Testujeme střední hodnotu normálních dat s neznámým rozptylem souboru, jaké rozdělení bude mít statistika?

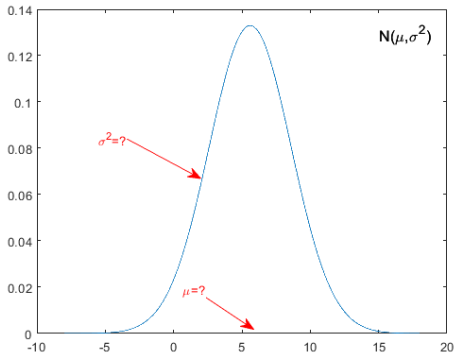
- speciálního rozdělení
- normálního rozdělení
- normovaného normálního rozdělení
- Studentovo t-rozdělení
- Chí-kvadrát rozdělení
- kategorického rozdělení

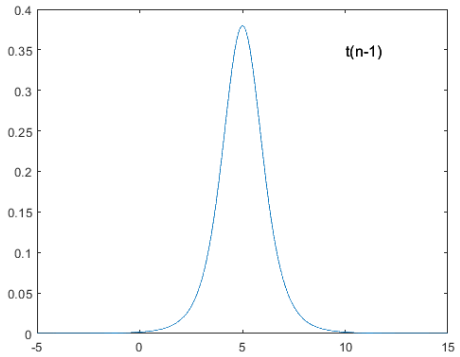
Máme normální data, proto se jeví jako možné normální rozdělení nebo rozdělení z něj vycházející.

- Normální rozdělení
- Normované normální rozdělení
- Studentovo t-rozdělení

Neznáme rozptyl souboru, proto nelze použít normální rozdělení $N(\mu,\sigma^2)$

Použijeme tedy Studentovo t-rozdělení





V jakém případě použiji dvoufaktorovou ANOVA

1. Potřebujeme mít zaručenu normalitu u všech výběrů, tedy p>0,05
2. Potřebuji mít potvrzeno, že rozptyly nejsou různé pro oba faktory (po řádcích i sloupcích) , tedy p>0,05

Test na normalitu vyšel u všech výběrů p>0.05 a Bartlettův test vyšel alespoň jednou p>0.05

Test na normalitu vyšel u všech výběrů p>0.05 a Bartlettův test vyšel po řádcích i sloupcích p>0.05

Test na normalitu vyšel u všech výběrů p<0.05 a Bartlettův test vyšel alespoň jednou p<0.05

Test na normalitu vyšel u všech výběrů p>0.05 a Bartlettův test vyšel alespoň jednou p<0.05

Test na normalitu vyšel u všech výběrů p<0.05 a Bartlettův test vyšel po řádcích i sloupcích p<0.05

Test na normalitu vyšel u všech výběrů p<0.05 a Bartlettův test vyšel alespoň jednou p>0.05

Charakteristiky variability jsou

Charakteristika:
1. Polohy → místo
2. Variability → „rozsah“

- modus, průměr, medián, směrodatná odchylka
- dolní kvartil, horní kvartil, modus, směrodatná odchylka
- rozpětí, rozptyl, mezikvartilové rozpětí, směrodatná odchylka
- horní kvartil, dolní kvartil, medián
- kvantily, rozptyl, rozpětí, průměr, směrodatná odchylka
- medián, dolní kvartil, horní kvartil, rozptyl

Modus, průměr, medián – ch. polohy

Dolní kvartil, horní kvartil, modus – ch. polohy

Vše charakteristiky variability

Vše charakteristiky polohy

Kvantil, průměr – charakteristiky polohy

Medián, dolní kvartil, horní kvartil – ch. polohy

Jaké minimální hodnoty nabývá Pearsonův korelační koeficient

Korelační koeficient nabývá hodnot v intervalu $r \in < -1,1 >$ → minimální hodnota: -1

Hodnoty korelačního koeficientu s vysvětlením (vysvětlení se musí brát s rezervou, jedná se o orientační hodnoty pro představu)

PŘÍMÁ ÚMĚRNOST

- 1 – absolutní závislost
- $< 0{,}85; 1)$ – silná závislost – lze rozumně predikovat
- $< 0{,}55; 0{,}85)$ – závislost (predikce velmi omezená)
- $< 0{,}25; 0{,}55)$ – slabá závislost
- $< 0; 0{,}25)$ - nezávislost

NEPŘÍMÁ ÚMĚRNOST

- -1 – absolutní závislost
- $< -0{,}85; -1)$– silná závislost – lze rozumně predikovat
- $< -0{,}55; -0{,}85)$ – závislost (predikce velmi omezená)
- $< -0{,}25; -0{,}55)$ – slabá závislost
- $< 0; -0{,}25)$ - nezávislost

Střední hodnotu diskrétní náhodné veličiny spočítám takto:

$$E[X] = \frac{1}{n}\sum x_i$$

$$E[X] = \sum p_i x_i$$

- vynásobím každou hodnotu její pravděpodobností, všechno sečtu a od výsledku odečtu průměr
- vynásobím každou hodnotu její pravděpodobností, všechno sečtu a vydělím počtem dat
- vynásobím každou hodnotu její pravděpodobností, všechno sečtu, vydělím počtem dat a od výsledku odečtu průměr
- vynásobím každou hodnotu její pravděpodobností a vydělím počtem dat
- vynásobím každou hodnotu její pravděpodobností, všechno sečtu a od výsledku odečtu rozptyl
- vynásobím každou hodnotu její pravděpodobností a všechno sečtu

Hustota pravděpodobnosti spojité náhodné veličiny

- je derivací pravděpodobnostní funkce
- je derivací distribuční funkce
- je integrálem rozptylu
- je integrálem pravděpodobnostní funkce
- je integrálem střední hodnoty
- je integrálem distribuční funkce

Jedná se o spojitou náhodnou veličinu → musíme nasčítávat pod funkcí, tedy se bude jednat o derivaci nebo integrál

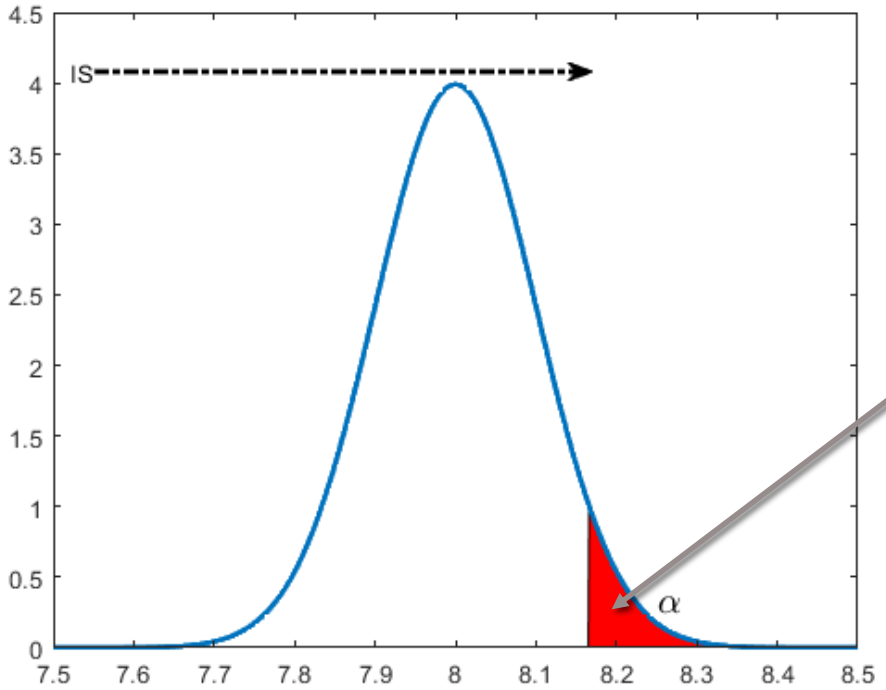
Hustota pravděpodobnosti – plocha pod křivou =1,

Distribuční funkce je součet ploch pod křivkou, tedy integrál

My známe součtovou charakteristiku → musíme derivovat

Provádíme-li jakýkoliv test, mluvíme o hladině významnosti. Pokud chceme při zamítnutí mít co nejmenší riziko omylu (s co nejmenší chybou), kterou z následujících hodnot vybereme?

- 0.01
- 0.005
- 0.1
- 0.01
- 0.05
- 0.5



Čím je menší plocha k zamítnutí, tím je jistější že zamítáme správně… vzpomeňte si na cvičení – kde byla hodně malá p-hodnota, věděli jsme, že třeba data jsou určitě vhodná k regresi, nebo zamítáme původní tvrzení.

Abychom si byli jistí, potřebujeme co nejmenší plochu, tedy co nejmenší hladinu významnosti.

Podmíněné rozdělení náhodné veličiny f(x|y) se rovná

- $f(x,y)f(y)$
- $f(x,y)/f(y|x)$
- $f(x,y)f(y)f(x)$
- $f(x)f(y)$
- $f(y)/f(x,y)$
- $f(x,y)/f(y)$

$$f(x,y) = f(x|y)f(y) = f(y|x)f(x)$$

Ptáme se na podmíněné rozdělení, tedy

$$f(x|y) = \frac{f(x,y)}{f(y)}$$

Pro konzistentní statistiku platí

- s rostoucím rozsahem výběru se rozptyl statistiky blíží k rozptylu souboru
- rozptyl statistiky se rovná rozptylu souboru
- s rostoucím rozsahem výběru se střední hodnota statistiky blíží nule
- rozptyl statistiky se rovná nule
- střední hodnota statistiky se rovná nule
- s rostoucím rozsahem výběru se rozptyl statistiky blíží k nule

Vlastnosti statistiky:

1. Nestrannost
   - Čím víc mám bodových odhadů, tak jejich průměr se blíží hledanému souborovému parametru
2. Konzistence
   - Čím mám víc dat, tím lepší je odhad a míň se liším od skutečného parametru. Každá extrémní hodnota je „vyrušena“ velkým množstvím přesných hodnot.
   - $D[\bar{X}] = \frac{\sigma^2}{n}$
3. Vydatnost
   - Vydatnější je ta statistika, která má menší rozptyl

Je dáno sdružené rozdělení dvou náhodných veličin: x={1,2} a y={1,2,3} ve tvaru tabulky [0.1 0.2 0.23; 0.11 0.01 0.35]. Čemu se rovná marginální rozdělení náhodné veličiny y?

Máme sdružené rozdělení pro diskrétní náhodnou veličinu, kterou přepíšeme do tabulky

	Y=1 (termín 1)	Y=2 (termín 2)	Y=3 (termín 3)
X=1 (muž)	0,1	0,2	0,23
X=2 (žena)	0,11	0,01	0,35

Zajímá nás marginální r. pro y, tedy s jakou pravděpodobností člověk přijde na první, druhý či třetí termín bez ohledu na pohlaví

	Y=1 (termín 1)	Y=2 (termín 2)	Y=3 (termín 3)
X=1 (muž)	0,1	0,2	0,23
X=2 (žena)	0,11	0,01	0,35
	0,21	0,21	0,58

Zkoumali jsme oblibu jednotlivých dopravních prostředků mezi zákazníky obchodního centra. Opakovaně jsme měřili počty zákazníků, kteří vystoupili u OC z jednotlivých dopravních prostředků a zapsali jsme je: Tramvaj=[76. 77. 19. 53. 14. 25. 21. …]; Autobus=[20. 84. 112. 32. 106. …]. Na hladině významnosti 95% testujte nulovou hypotézu, že v průměru cestuje stejný počet zákazníků v autobuse a v tramvaji.

Abychom mohli určit, který test použijeme, nejdříve otestujeme normalitu dat pomocí Shapiro testu nebo AD testu s tímto výsledkem. U prvního výběru (tramvaj) nám vyšla p-hodnota p_tram=0.0021, u druhého (autobus) p_bus=0.0047.

Pro použití parametrických testů: potřeba normalita dat u všech výběrů → už při prvním zamítnutí ($p < \alpha$), není potřeba pokračovat. Nelze použít parametrický test hypotéz.

Hladina významnosti $\alpha = 0.05$, takže už pro p-hodnota pro tramvaj je menší a tedy použijeme neparametrický test.

- už pro první testovaný výběr nám p-hodnota vyšla menší než hladina významnosti, proto použijeme neparametrické testy hypotéz

Zkoumali jsme oblibu jednotlivých dopravních prostředků mezi zákazníky obchodního centra. Opakovaně jsme měřili počty zákazníků, kteří vystoupili u OC z jednotlivých dopravních prostředků a zapsali jsme je: Tramvaj=[76. 77. 19. 53. 14. 25. 21. …]; Autobus=[20. 84. 112. 32. 106. …]. Na hladině významnosti 95% testujte nulovou hypotézu, že v průměru cestuje stejný počet zákazníků v autobuse a v tramvaji.

Na základě předchozího zjištění testujte nulovou hypotézu, že v průměru cestuje stejný počet cestujících v autobuse a v tramvaji.

- Neparametrické testy hypotéz, 2 výběry, Napárová data, Střední hodnota → Mann-Whitney test

- Testovali jsme nulovou hypotézu, že v průměru cestuje stejný počet cestujících v autobuse a v tramvaji a vyšla nám p-hodnota = 0.012. Jaký je závěr?
  - H0: mediány výběrů jsou stejné
  - HA: mediány výběrů nejsou stejné

  - $(p < \alpha)$ → Zamítám hypotézu, že v průměru cestuje stejný počet cestujících v autobuse a v tramvaji.

Po semestru distančního studia jsme se zeptali, jakou formou komunikovali a jak byli spokojeni. výsledky jsme zapsali do tabulky:

	Spokojen	Tak napůl	Nespokojen
E-mail	5	12	20
MS Tams	12	45	12
Jiný kanál	6	22	44
Vůbec	??	24	55

Je dobré si říci, že tu máme diskrétní data
Test nezávislosti pro diskrétní data:

- Fisherův exact test
- Chí-kvadrát test nezávislosti
  - n>2, v 80% n>5 → výsledek: 3

Chí-kvadrát test nezávislosti

Nezamítám, když p>α… tzn. p>0,05
Maximální p-hodnota je 1, minimální je 0

Odpověď: 1

	Spokojen	Tak napůl	Nespokojen
E-mail	5	12	20
MS Tams	12	45	12
Jiný kanál	6	22	44
Vůbec	41	24	55