%%%%%%%%%% cv4\_NV

L0: # Cvičení 4 - Náhodný vektor

L1: ## Martina Litschmannová, Adéla Vrtková, Michal Béreš

L2: # Příklad ze sbírky

L3: Náhodný vektor $Z =(Y;X)^T$ má pravděpodobnostní funkci zadanou tabulkou ![image.png](attachment:image.png)

L4: ## a) Určete chybějící hodnotu sdružené pravděpodobnostní funkce,

L5: případně byrow = ...

L6: nespoštějte tuto buňku 2krát, jinak si hodnotu zase nastavíte na 0,

L7: víte proč?

L8: ## b) Určete distribuční funkci

L9: \*\*Pozor! Vektor Z je $(Y,X)^T$ tedy první parametr je hodnota Y a druhý hodnota X.\*\*

L10: F(2.8; 7.1)

L11: = P(Y<2.8,X<7.1)

L12: projdeme řádky a sloupce, vždy si vezmeme jednu hodnotu

L13: z příslušného řádku nebo sloupce

L14: ## c) Určete marginální rozdělení

L15: ## d) Podmíněné pravděpodobnosti a podmíněné pravděpodobnostní funkce $P(x|y), P(y|x)$

L16: P(Y > 2.1 | X < 5.3)

L17: = P(Y > 2.1 ∧ X < 5.3) / P(X < 5.3)

L18: P(X = 5|Y = 1)

L19: = P(X = 5 ∧ Y = 1) / P(Y = 1)

L20: \*\*$P(x|y)=\frac{P(X=x,Y=y)}{P\_Y(y)}$\*\*

L21: má to stejnou velikost, tak si ukradneme formátování

L22: \*\*$P(y|x)$\*\*

L23: má to stejnou velikost, tak si ukradneme formátování

L24: ## e) zákládní charakteristiky náhodných veličin X a Y

L25: ## f) podmíněná střední hodnota E(X|Y = 2)

L26: P(x|Y=2)

L27: ## g) kovariance a korelace

L28: matice kde v každé kolonce je hodnota x\*y

L29: střední hodnota E(X\*Y)

L30: kovariance

L31: korelace

%%%%%%%%%% cv12-extra

L0: # Cvičení 12. Vícevýběrové testy - Extra pro zájemce ## Michal Béreš

L1: ## Načteme si testovací data a vyrobíme post-hoc + efekty pro ANOVU a KW test

L2: data jsou ve standardním dtovém formátu

L3: POST-HOC ANOVA

L4: počítání efektů ANOVA

L5: celkový průměr

L6: průměry ve skupinách

L7: efekty

L8: vypsat setřízené

L9: POST-HOC KW

L10: post hoc - jiná funkce s výstupem, který se nám více hodí

L11: číselně odpovídá té použité na cvičení

L12: install.packages("FSA")

L13: library FSA

L14: počítání efektů KW

L15: celkový průměr

L16: průměry ve skupinách

L17: efekty

L18: vypsat setřízené

L19: # Pro zájemce (nepovinné) - vytvoření setřízené tabulky p-hodnot/pisménkové schéma automatizovaně

L20: install.packages("stringi")

L21: toto je knihovna pro hledání v textu

L22: budeme hledat jména šmoulů v párových post-hoc testech

L23: inicializujeme si matici (pro pěkný výsledek jako textovou)

L24: 7x7 protože máme 7 šmoulů

L25: pojmenujeme její sloupce i řádky dle setřízených šmoulů

L26: smyčka přes všechny testy v post-hoc (řes názvy sloupců)

L27: kteří trpaslíci jsou přítomni v tomto párovém testu?

L28: jaké jsou indexy těchto trpaslíků

L29: indexy pro zápis do matice - vždy 2 hodnoty

L30: zapíšu do matice (první index je menší -> automaticky do

L31: horního trojuhelníku)

L32: zapisujeme text (pokud je matice textová, čísla se automaticky

L33: převedou na text), hodnoty na tisíciny

L34: ### Funkce pro aoutomatizované znaménkové schéma (ručně napsané a z balíčku)

L35: #### Ručně napsané funkce (to co bychom dělali na papír)

L36: tabulka p-hodnot

L37: pocet skupin

L38: pojmenujeme její sloupce i řádky dle setřízených typů

L39: smyčka přes všechny testy v post-hoc (řes názvy sloupců)

L40: kteří trpaslíci jsou přítomni v tomto párovém testu?

L41: jaké jsou indexy těchto trpaslíků

L42: indexy pro zápis do matice - vždy 2 hodnoty

L43: zapíšu do matice (první index je menší -> automaticky do

L44: horního trojuhelníku)

L45: písmenkové schéma z tabulky

L46: jak je velká matice

L47: inicilizace matice

L48: názvy řádků - kopie od vstupu

L49: nastavení diagonály na 1 - je v dané skupině

L50: cyklus přes všechny sloupce kde můžeme něco vyplňovat

L51: cyklus přes všechny řádky v sloupci kde sledujeme pval

L52: je-li pval > alpha tak přidáme do hom. skupiny

L53: #### Jak použít ručně napsané funkce pro ANOVU a KW test?

L54: Jak to udělat z POST-HOC ANOVY:

L55: vyrobíme vstupní data

L56: vyrobíme setřízenou tabulku phodnot

L57: vykreslíme zaokrouhlené na tisíciny

L58: z tabulky phodnot vyrobíme písmenkové schéma

L59: Jak to udělat z POST-HOC KW:

L60: vyrobíme vstupní data

L61: vyrobíme setřízenou tabulku phodnot

L62: vykreslíme zaokrouhlené na tisíciny

L63: z tabulky phodnot vyrobíme písmenkové schéma

L64: ## Písmenkové schéma pomocí vestavěné Rkové funkce

L65: Balíček rcompanion, funkce cldList

L66: v případě ANOVY

L67: nejprve vyrobíme dataframe se sloupci dvojic a phodnot

L68: písmenkové schéma, library rcompanion

L69: install.packages("rcompanion")

L70: v případě KW

L71: nejprve vyrobíme dataframe se sloupci dvojic a phodnot

L72: písmenkové schéma, library rcompanion

L73: install.packages("rcompanion")

%%%%%%%%%% cv9

L0: # Cvičení 9. Intervalové odhady (jednoho výběru) ## Michal Béreš, Martina Litschmannová, Veronika Kubíčková

L1: # Demonstrace na úvod - co je to intervalový odhad?

L2: Uvažujme náhodnou veličinu z normálního rozdělení se střední hodnotou $\mu$ a směrodatnou odchylkou $\sigma$. Budeme pracovat s výběry z této náhodné veličiny a pomoci jich se budeme snažit odhadnout střední hodnutu rozdělení (zde známe její skutečnou hodnotu, ale v praxi je její hodnota neznámá).

L3: velikost výběru

L4: střední hodnota

L5: směr. odchylka

L6: simulace náhodného výběru ze zadané náhodné veličiny

L7: výběrový průměr jako bodový odhad

L8: výběrová směr. odch.

L9: Pro přehlednost si můžeme výběr vizualizovat.

L10: šířka grafů v Jupyteru

L11: matice grafů 1x2

L12: ### Samotné sestrojení Intervalového odhadu pomocí výběrové charakteristiky

L13: Použijeme tuto výběrovou charakteristiku: (předpokládáme, že neznáme žádné skutečné parametry rozdělení, pouze to, že je normální)<br> $Y=\frac{\bar X - \mu}{S}\sqrt{n} \sim t\_{n-1}$ <br> Jelikož, známe rozdělení Y jsme schopni napočítat $a$ a $b$ v následujícím výrazu:<br> $P(a<Y<b)\geq 1 - \alpha$ <br> - $\alpha$ nazýváme hladinou významnosti (pravděpodobnost, že hledaná hodnota leží mimo náš interval) - $1-\alpha$ nazýváme spolehlivost intervalového odhadu

L14: $a$ a $b$ zvolíme tak aby byly v pravděpodobnosti symetrické, tzn.: - $P(Y<a)\leq \alpha / 2 \rightarrow a=t\_{\alpha / 2;n-1}$ - $P(b<Y)\leq \alpha / 2 \rightarrow P(Y\leq b)\geq 1 - \alpha / 2 \rightarrow b=t\_{1-\alpha / 2;n-1}$

L15: maximální pravděpodobnost s jakou připouštíme aby

L16: skutečná st. hod. ležela mimo sestrojený interval

L17: příslušné kvantily studentova rozdělení

L18: Dále jen doplníme do výrazu a upravíme:<br> $P(t\_{\alpha / 2;n-1}<\frac{\bar X - \mu}{S}\sqrt{n}<t\_{1-\alpha / 2;n-1})\geq 1 - \alpha$ <br> $P(\bar X - t\_{1-\alpha / 2;n-1}\frac{S}{\sqrt{n}}<\mu<\bar X - t\_{\alpha / 2;n-1}\frac{S}{\sqrt{n}})\geq 1 - \alpha$ <br>

L19: Tento konkrétní odhad, můžeme dostat také pomocí Rkovské funkce t.test:

L20: ### Otestování intervalového odhadu na více výběrech

L21: počet výběrů

L22: velikost výběru

L23: střední hodnota

L24: směrodat. odchyl.

L25: hladina významnosti

L26: příslušné kvantily studentova rozdělení

L27: vykreslení skutečné střední hodnoty

L28: cyklus přes jednotlivé výběry

L29: zvolíme barvu vykreslení, podle toho zda IO obsahuje stř. hod.

L30: vykreslíme IO jako vertikální čáru

L31: vrátím šířku na standardní velikost

L32: šířka grafů v Jupyteru

L33: # Typy intervalových odhadů

L34: (Ukázky na odhadu střední hodnoty dat z normálního rozdělení.)

L35: ## Dolní/Levostranný IO

L36: - $P(M\_D^\* < \mu) = 1-\alpha$ - v Rku \*\*alternative="greater"\*\*

L37: ## Horní/Pravostranný IO

L38: - $P(\mu < M\_H^\*) = 1-\alpha$ - v Rku \*\*alternative="less"\*\*

L39: ## Oboustranný IO

L40: - $P(M\_D < \mu < M\_H) = 1-\alpha$ - v Rku \*\*alternative="two.sided"\*\*

L41: # Přehled parametrů výběru a jejich bodových/intervalových odhadů

L42: Běžně máme k dispozici více konstrukcí IO (funkcí v Rku které to za nás udělají), ale každá konstrukce má jiné požadavky na data a vytváří různě "kvalitní" (ve smyslu velikosti IO) odhady. My budeme vždy vybírat "nejkvalitnější" IO, který \*\*má splněny\*\* předpoklady použití.<br> Pořadí různých IO níže bude vždy od "nejlepšího" po nejrobustnější.

L43: ## Míry polohy jednoho výběru

L44: Mírami polohy rozumíme údaj určující polohu dat, nehledě na tom jak jsou rozptýlená. Pro data z normálního rozdělení můžeme odhadovat střední hodnotu, pro ostatní medián.

L45: #### a) studentův t-test IO

L46: - odhadujeme střední hodnotu - bodový odhad je výběrový průměr - data musejí pocházet z normálního rozdělení - exploračně: šikmost a špičatost leží v (-2,2) - exploračně: QQ graf má body přibližně na čáře - exaktně: pomocí statistického testu, např. Shapiro-Wilk test (shapiro.test(data))

L47: exploračně test normality

L48: library(moments) - tomuto se můžeme vyhnout voláním moments::

L49: je to bezpečnější - máme jistotu, že voláme funkci z tohoto balíčku

L50: exaktně test normality dat

L51: vysledná p-hodnota musím být větší než hl. výz. (př. 0.05)

L52: bodový odhad

L53: IO

L54: #### b) Wilcoxnův test IO

L55: - odhadujeme medián - bodový odhad je výběrový medián - data musejí pocházet ze symetrického rozdělení - exploračně: šikmost leží v (-2,2) - exploračně: histogram vypadá přibližně symetricky - exaktně: pomocí statistického testu, např. balíček "lawstat", funkce "symmetry.test(data,boot=FALSE)" - funkce v Rku vyžaduje dodatečný parametr (conf.int = TRUE)

L56: exploračně

L57: exaktně: test symetrie

L58: install.packages("lawstat")

L59: vysledná p-hodnota musím být větší než hl. výz. (př. 0.05)

L60: bodový odhad

L61: IO

L62: #### c) znaménkový test test IO

L63: - odhadujeme medián - bodový odhad je výběrový medián - výběr většího rozsahu (>10) - funkce v Rku vyžaduje dodatečný parametr (conf.int = TRUE) - vyžaduje knihovnu "BSDA" - jakožto nejrobustnější test, se dá použít i na nespojitá data - např. pořadí v nějakém seznamu

L64: skutečný medián

L65: bodový odhad

L66: quantile(vyber, probs = 0.5)

L67: IO

L68: install.packages("BSDA")

L69: ## Míry variability jednoho výběru

L70: Mírami variability rozumíme údaj určující rozptýlenost/variabilitu dat, nehledě na celkových hodnotách. Pro data z normálního rozdělení můžeme odhadovat směrodatnou odchylku.

L71: #### IO směrodatné odchylky

L72: - odhadujeme směrodatnou odchylku - bodovým odhadem je výběrový směrodatná odchylka - data musejí pocházet z normálního rozdělení - exploračně: šikmost a špičatost leží v (-2,2) - exploračně: QQ graf má body přibližně na čáře - exaktně: pomocí statistického testu, např. Shapiro-Wilk test (shapiro.test(data)) - vyžaduje balíček "EnvStats" - funkce v Rku, dává výpočet rozptylu - nutná odmocnina výsledku

L73: exploračně test normality

L74: exaktně test normality dat

L75: vysledná p-hodnota musím být větší než hl. výz. (př. 0.05)

L76: bodový odhad

L77: IO

L78: install.packages("EnvStats")

L79: Přidáme si ruční výpočet: - vycházíme ze statistiky: $\frac{S^2}{\sigma^2}(n-1) \sim \chi^2\_{n-1}$ - Horní mez: - $P(\frac{S^2}{\sigma^2}(n-1) < \chi^2\_{\alpha /2, n-1}) = \alpha /2$ - $P(\frac{S^2}{\chi^2\_{\alpha /2, n-1}}(n-1) < \sigma^2 ) = \alpha /2$ - Dolní mez: - $P(\frac{S^2}{\sigma^2}(n-1) > \chi^2\_{1-\alpha /2, n-1}) = \alpha /2$ - $P(\frac{S^2}{\chi^2\_{1-\alpha /2, n-1}}(n-1) > \sigma^2 ) = \alpha /2$ - Dohromady: $P(\frac{S^2}{\chi^2\_{1-\alpha /2, n-1}}(n-1) < \sigma^2 <\frac{S^2}{\chi^2\_{\alpha /2, n-1}}(n-1)) = 1 - \alpha$

L80: ruční výpočet

L81: ## Pravděpodobnost výskytu u jednoho výběru

L82: #### IO pravděpodobnosti

L83: - odhadujeme pravděpodobnost - bodový odhad je relativní četnost - vyžadujeme dostatečný počet dat: $n>\frac{9}{p(1-p)}$ - Clopperův - Pearsonův odhad (binom.test) - jako parametr nebere data, ale počet úspěchů a počet pozorování - Waldův - z výběrových charakteristik

L84: ověření předpokladů

L85: bodový odhad

L86: intervalový odhad Clopperův - Pearsonův

L87: Intervalový odhad Waldův

L88: dolní mez IO

L89: horní mez IO

L90: Výpočet 11 nejčastěji používaných intervalů spolehlivosti param. bin. rozdělení

L91: pomocí balíčku binom

L92: install.packages("binom")

L93: # Příklady

L94: ## Příklad 1.

L95: Při kontrolních zkouškách 16 žárovek byl stanoven odhad střední hodnoty $\bar x$ = 3 000 hodin a směrodatné odchylky s = 20 hodin jejich životnosti. Za předpokladu,že životnost žárovky má normální rozdělení, určete 90% intervalový odhad pro parametry µ a σ

L96: Odhadujeme stř.hodnotu a směr.odchylku životnosti žárovek

L97: Součástí zadání je informace o normalitě dat

L98: rozsah souboru

L99: hodin.... průměr (bodový odhad střední hodnoty)

L100: hodin.... výběrová směrodatná odchylka (bodový odhad sm. odchylky)

L101: hladina významnosti (spolehlivost 1-alpha = 0.9)

L102: Oboustranný intervalový odhad střední hodnoty

L103: dolní mez IO

L104: horní mez IO

L105: Oboustranný intervalový odhad směrodatné odchylky

L106: dolní mez IO

L107: horní mez IO

L108: ## Příklad 2.

L109: Hloubka moře se měří přístrojem, jehož systematická chyba je rovna nule a náhodné chyby mají normální rozdělení se směrodatnou odchylkou 20 m. Kolik nezávislých měření je třeba provést,aby s pravděpodobností 95 % stanovila hloubku s chybou menší než 10 m?

L110: Určujeme odhad potřebného rozsahu výběru (počtu potřebnych měření)

L111: Předpokládáme normalitu dat, se známým rozpylem (dle zadání)

L112: metrů .... známá směrodatná odchylka

L113: hladina významnosti (spolehlivost 1-alpha = 0.95)

L114: metrů ... přípustná chyba měření

L115: Odhad rozsahu výběru

L116: Y = delta/sigma\*sqrt(n) ~ N(0,1), delta = X-mu

L117: P(Y > Z\_(1-alpha/2)) = alpha/2

L118: ## Příklad 3.

L119: Úkolem je určit průměrnou hladinu cholesterolu v séru v určité populaci mužů. V náhodném výběru (pocházejícím z normálního rozdělení ) 25 mužů je výběrový průměr 6,3 mmol/l a výběrová směrodatná odchylka 1,3 mmol/l.

L120: Odhadujeme střední hladinu cholesterolu v séru

L121: Předpokládáme normalitu dat (dle zadání)

L122: rozsah souboru

L123: mmol/l .... průměr (bodový odhad střední hodnoty)

L124: mmol/l .... výběrová směr. odchylka (bodový odhad sm. odchylky)

L125: hladina významnosti (spolehlivost 1-alpha = 0.95)

L126: Oboustranný intervalový odhad střední hodnoty

L127: dolní mez IO

L128: horní mez IO

L129: ## Příklad 4.

L130: Předpokládejme, že v náhodném výběru 200 mladých mužů má 120 z nich vyšší než doporučenou hladinu cholesterolu v séru. Určete 95% interval spolehlivosti pro procento mladých mužů s vyšší hladinou cholesterolu v populaci.

L131: Odhadujeme podíl mužů s vyšší hladinou cholesterolu v celé populaci,

L132: tj. pravděpodobnost,že náhodně vybraný muž bude mít vyšší hladinu cholesterolu

L133: rozsah souboru

L134: počet "úspěchů"

L135: relativní četnost (bodový odhad pravděpodobnosti)

L136: hladina významnosti (spolehlivost 1-alpha = 0.95)

L137: Ověření předpokladů

L138: Oboustranný Clopperův - Pearsonův (exaktní) int.ý odhad param. binom. rozdělení

L139: # Waldův (asymptotický) odhad (z-statistika) - aprox. normálním rozdělením dle CLV

L140: dolní mez IO

L141: horní mez IO

L142: ## Příklad 5.

L143: V rámci výzkumné studie pracujeme s náhodným výběrem 70 žen z české populace. U každé z žen byl změřen hemoglobin s přesností 0,1 g/100 ml. Naměřené hodnoty jsou v uvedeny v souboru Hemoglobin.xls. Nalezněte 95% intervalové odhady směrodatné odchylky a střední hodnoty hemoglobinu v populaci českých žen. (Normalitu ověřte na základě exploračních grafů.)

L144: # Odhadujeme střední hodnotu a směrodatnou odchylku hemoglobinu v séru

L145: # Načtení dat z xlsx souboru (pomoci balíčku readxl)

L146: # Explorační analýza

L147: Data neobsahují odlehlá pozorování.

L148: Ověření normality - exploračně

L149: Šikmost i špičatost odpovídá norm. rozdělení.

L150: ověření normality: exaktně - test normality.

L151: Známe-li testování hypotéz, ověříme Shapirovým . Wilkovým testem.

L152: Na hl. významnosti 0.05

L153: 95% oboustranný intervalový odhad střední hodnoty

L154: # 95% oboustranný intervalový odhad směrodatné odchylky

L155: ## Příklad 6.

L156: Jaký musí být počet pozorování, jestliže chceme s pravděpodobností 0,95 stanovit průměrnou hodnotu hemoglobinu u novorozenců s chybou nejvýše 1,0 $g/l$. Populační rozptyl hodnot se odhaduje hodnotou 46,0 $g^2/l^2$.

L157: Určujeme odhad potřebného rozsahu výb. (počtu novorozenců, které musíme testovat)

L158: Předpokládáme normalitu dat, bez tohoto předpokladu je příklad neřešitelný

L159: g/l .... známá směrodatná odchylka

L160: hladina významnosti (spolehlivost 1-alpha = 0.95)

L161: g/l ... přípustná chyba měření

L162: Odhad rozsahu výběru

L163: Y = delta/sigma\*sqrt(n) ~ N(0,1), delta = X-mu

L164: P(Y > Z\_(1-alpha/2)) = alpha/2

L165: ## Příklad 7.

L166: V datovém souboru pr7.xlsx naleznete měření hluku způsobeného větrákem počítače [dB]. Spočtěte 95% intervalový odhad průměrného hluku a 95% intervalový odhad variability hluku.

L167: načtení dat

L168: vizualizace

L169: odstranění OP

L170: test normality dat exploračně

L171: test normality exaktně

L172: bodový a intervalový odhad střední hodnoty

L173: bodový a intervalový odhad směrodatné odchylky

L174: ## Příklad 8.

L175: V datovém souboru pr8.xlsx naleznete měření doby do poruchy elektrické součástky [h]. Spočtěte 99% intervalový odhad průměrné životnosti daného typu součastky.

L176: načtení dat

L177: vizualizace a ověření OP

L178: test normality dat exploračně

L179: test normality exaktně

L180: test symetrie exploračně

L181: exaktně: test symetrie

L182: install.packages("lawstat")

L183: vysledná p-hodnota musím být větší než hl. výz. (př. 0.05)

L184: bodový a intervalový odhad mediánu

L185: IO

L186: install.packages("BSDA")

%%%%%%%%%% cv6

L0: # Cvičení 6 - Vybraná rozdělení spojité náhodné veličiny

L1: ## Martina Litschmannová, Adéla Vrtková, Michal Béreš

L2: # Přehled rozdělení a jejich funkcí

L3: ## Úvod: Hustota pravděpodobnosti, Distribuční funkce a Kvantilová funkce

L4: ### Hustota pravděpodobnosti

L5: - začíná písmenkem \*\*d\*\*: p = d...(x, ...)

L6: ### Distribuční funkce

L7: - začíná písmenkem \*\*p\*\*: $p = P(X < x)$: p = p...(x, ...)

L8: ### Kvantilová funkce

L9: - začíná písmenkem \*\*q\*\*: najdi x pro zadané p: $p = F(x) \rightarrow x = F^{-1}(p)$: x = q...(p, ...)

L10: ## Rovnoměrné rozdělení: $X \sim Ro(a, b)$

L11: - náhodná veličina nabývá pouze hodnot větších než a a menších než b - všechny hodnoty mají stejnou hustotu výskytu -> hustota pravděpodobnosti je konstantní mezi a a b, jinde nulová

L12: Hustota pravděpodobnosti f(x)

L13: odkud

L14: kam

L15: vykreslíme si Hustotu pravděpodobnosti

L16: cex je velikost markerů

L17: Distribuční funkce F(x) = P(X < x)

L18: odkud

L19: kam

L20: vykreslíme si Distribuční funkci

L21: kvantilová funkce F^(-1)(p) = x: P(X<x)=p

L22: odkud

L23: kam

L24: vykreslení - kvantilová funkce F^(-1)(p) = x

L25: ## Exponenciální rozdělení: $X \sim Exp(\lambda)$

L26: - doba do 1. události, doba mezi událostmi (pouze v období stabilního života - Poissonův proces) - parametr $\lambda$ je tentýž co v Poissonově rozdělení - střední hodnota je: $E(X)=1 / \lambda$

L27: Hustota pravděpodobnosti f(x)

L28: vykreslíme si Hustotu pravděpodobnosti

L29: Distribuční funkce F(x) = P(X < x)

L30: vykreslíme si Distribuční funkci

L31: kvantilová funkce F^(-1)(p) = x: P(X<x)=p

L32: vykreslení - kvantilová funkce F^(-1)(p) = x

L33: ## Weibullovo rozdělení: $X \sim W(\theta,\beta)$

L34: - doba do 1. události (poruchy)(vhodná volba β umožuje použití v libovolném období intenzity poruch) - rozšíření exponenciálního rozdělení Exp(λ) = W(Θ=1/λ, β=1)

L35: Hustota pravděpodobnosti f(x)

L36: ekvivalent 1/lambda u exp. rozdělení

L37: beta = 1 -> exponenciální rozdělení

L38: vykreslíme si Hustotu pravděpodobnosti

L39: Distribuční funkce F(x) = P(X < x)

L40: ekvivalent 1/lambda u exp. rozdělení

L41: beta = 1 -> exponenciální rozdělení

L42: vykreslíme si Distribuční funkci

L43: kvantilová funkce F^(-1)(p) = x: P(X<x)=p

L44: ekvivalent 1/lambda u exp. rozdělení

L45: beta = 1 -> exponenciální rozdělení

L46: vykreslení - kvantilová funkce F^(-1)(p) = x

L47: ## Normální rozdělení: $X \sim N(\mu,\sigma^2)$

L48: - rozdělení modelující např. chyby měření, chování součtu/průměru mnoha jiných náhodných veličin - viz. Centrální limitní věta - $\mu$ je přímo střední hodnota rozdělení: $E(X)=\mu$ - $\sigma$ je přímo směrodatná odchyla rozdělení: $D(X)=\sigma^2$ - s parametry $\mu=0,\sigma=1$ se nazývá normované Normální rozdělení

L49: Hustota pravděpodobnosti f(x)

L50: vykreslíme si Hustotu pravděpodobnosti

L51: Distribuční funkce F(x) = P(X < x)

L52: vykreslíme si Distribuční funkci

L53: kvantilová funkce F^(-1)(p) = x: P(X<x)=p

L54: vykreslení - kvantilová funkce F^(-1)(p) = x

L55: # Příklady

L56: ## Příklad 1.

L57: Výška v populaci chlapců ve věku 3,5-4 roky má normální rozdělení se střední hodnotou 102 cm a směrodatnou odchylkou 4,5 cm. Určete, jaké procento chlapců v uvedeném věku má výšku menší nebo rovnou 93 cm.

L58: X ... výška chlapců ve věku 3.5 až 4 roky (cm)

L59: X ~ N(mu = 102, sd = 4.5)

L60: P(X<=93)=F(93)

L61: ## Příklad 2.

L62: Průměrná životnost strojní součástky je 30 000 hodin. Předpokládejme, že součástka je v období stabilního života. Určete:

L63: X ... životnost součástky (h)

L64: X ~ Exp(lambda), kde E(X)=1/lambda

L65: ### a)

L66: pravděpodobnost, že součástka nevydrží více než 2 000 hodin,

L67: a) P(X<2000)=F(2000)

L68: ### b)

L69: pravděpodobnost, že součástka vydrží více než 35 000 hodin,

L70: b) P(X>35000)=1-F(35000)

L71: ### c)

L72: dobu, do níž se porouchá 95 % součástek.

L73: c) P(X<t)=0,95 -> F(t)=0,95 -> t… 95% kvantil

L74: ## Příklad 3.

L75: Výrobní zařízení má poruchu v průměru jednou za 2000 hodin. Veličina Y představující dobu čekání na poruchu má exponenciální rozdělení. Určete dobu T0 tak, aby pravděpodobnost, že přístroj bude pracovat delší dobu než T0, byla 0,99.

L76: X ... doba čekání na poruchu (h)

L77: X ~ Exp(lambda), kde E(X)=1/lambda

L78: P(X>t)=0,99 -> 1-F(t)=0,99 -> F(t)=0,01 -> t… 1% kvant.

L79: ## Příklad 4.

L80: Výsledky měření jsou zatíženy jen normálně rozdělenou chybou s nulovou střední hodnotou a se směrodatnou odchylkou 3 mm. Jaká je pravděpodobnost, že při 3 měřeních bude alespoň jednou chyba v intervalu (0 mm; 2,4mm)?

L81: Y… velikost chyby měření (mm)

L82: Y ~ N(mu = 0,sigma = 3)

L83: pp… pravd., že chyba měření bude v int. 0,0-2,4mm

L84: X … počet chyb měření v int. 0 mm -2,4 mm ve 3 měř.

L85: X ~ Bi(n = 3,p = pp)

L86: P(X>=1)=1-P(X=0)

L87: ## Příklad 5.

L88: Ve velké počítačové síti se průměrně přihlašuje 25 uživatelů za hodinu. Určete pravděpodobnost, že:

L89: ### a)

L90: se nikdo nepřihlásí během 14:30 - 14:36,

L91: X … počet uživatelů přihlášených za 6 minut

L92: X ~ Po(lt = 2.5)

L93: P(X=0)

L94: ### b)

L95: do dalšího přihlášení uběhnou 2-3 minuty.

L96: Y … doba do dalšího přihlášení

L97: Y ~ Exp(lambda = 25/60), kde E(X)=1/lambda

L98: P(2<Y<3)=F(3)-F(2)

L99: ### c)

L100: Určete maximální délku časového intervalu tak, aby pravděpodobnost, že se nikdo nepřihlásí byla alespoň 0,90.

L101: P(Y>t)=0,90 -> 1-F(t)=0,90 -> F(t)=0,10 -> t…10% kv.

L102: ## Příklad 6.

L103: Náhodná veličina X má normální rozdělení N(µ; σ). Určete:

L104: ### a)

L105: P(µ − 2σ < X < µ + 2σ),

L106: P(µ − 2σ < X < µ + 2σ) = F(µ + 2σ) - F(µ - 2σ)

L107: X~N(µ,σ)

L108: je jedno jaké hodnoty zvolíme

L109: ### b)

L110: nejmenší k ∈ Z, tak, aby P(µ − kσ < X < µ + kσ) > 0,99.

L111: normální rozdělení je symetrické

L112: P(µ − kσ < X < µ + kσ) =

L113: = 1 - (P(X < µ − kσ ) + P(X > µ + kσ)) =

L114: = 1 - 2\*P(X > µ + kσ) = 0.99 -> P(X > µ + kσ) = 0.005

L115: -> P(X < µ + kσ) = 0.995

L116: x = µ + kσ

L117: ## Příklad 7.

L118: Na prohlídce výstavy je promítán doprovodný film o životě autora vystavovaných děl. Jeho projekce začíná každých 20 minut. Určete pravděpodobnost, že pokud náhodně přijdete do promítacího sálu,

L119: Y … doba do začátku další projekce

L120: Y ~ Ro(a=0, b=20)

L121: ### a)

L122: nebudete na začátek filmu čekat víc než 5 minut,

L123: P(X<5)

L124: ### b)

L125: budete čekat mezi 5 a 10 minutami,

L126: P(5<X<10)

L127: ### c)

L128: střední hodnotu a směrodatnou odchylku doby čekání na začátek filmu.

L129: ## Příklad 8.

L130: Při kontrole jakosti přebíráme součástku pouze tehdy, jestliže se její rozměr pohybuje v mezích 26-27 mm. Rozměry součástek mají normální rozdělení se střední hodnotou 26,4 mm a směrodatnou odchylkou 0,2 mm. Jaká je pravděpodobnost, že rozměr součástky náhodně vybrané ke kontrole bude v požadovaných mezích?

L131: X ... rozměr součástky (mm)

L132: X ~ N(mu = 26.4,sigma = 0.2)

L133: P(26<X<27)=F(27)-F(26)

L134: ## Příklad 9.

L135: Délka skoků sportovce Jakuba měřená v cm má normální rozdělení N(µ1; σ1), kde µ1 = 690 a σ1 = 10. Délka skoků sportovce Aleše měřená v cm má také normální rozdělení N(µ2; σ2), kde µ2 = 705 a σ2 = 15. Na závody se kvalifikuje ten, kdo ze dvou skoků alespoň jednou skočí více než 700 cm.

L136: SJ ... délka skoku Jakuba

L137: SJ ~ N(mu = 690,sigma = 10)

L138: SA … délka skoku Aleše

L139: SA ~ N(mu = 705,sigma = 15)

L140: J...Jakubův skok je úspěšný (delší než 700 cm)

L141: A...Alešův skok je úspěšný (delší než 700 cm)

L142: P(J)=P(SJ>700)=1-F(700)

L143: P(A)=P(SA>700)=1-F(700)

L144: KJ … Jakub se kvalifikuje na závody,

L145: P(KJ) = 1-(1-P(J))(1-P(J))

L146: KA … Aleš se kvalifikuje na závody,

L147: P(KA) = 1-(1-P(A))(1-P(A))

L148: ### a)

L149: S jakou pravděpodobností se oba dva kvalifikují na závody?

L150: ada)

L151: ### b)

L152: S jakou pravděpodobností se kvalifikuje Aleš, ale Jakub ne?

L153: adb)

%%%%%%%%%% cv1

L0: # Cvičení 1 - Stručný úvod do R, Kombinatorika

L1: ## Adéla Vrtková, Michal Béreš, Martina Litschmannová

L2: Nejprve si projdeme některé základy Rka a naimplementujeme si některé kombinatorické funkce.

L3: # Krátký úvod do R

L4: jednoduché početní operace

L5: POZOR na závorky! Pro počítání se používají pouze kulaté!

L6: Hranaté a složené mají v R jinou funkci!

L7: kombinační číslo, faktoriály

L8: datové typy -> numeric, character, logical, (complex)

L9: funkcí class zjišťujeme typ objektu

L10: ## datové struktury v R

L11: - vector (rozumíme sloupcový vektor) - factor (speciální případ vektoru) - matrix (matice s rozměry n x m) - data.frame (datový rámec s rozměry n x p) - list (seznam)

L12: definování vektoru (sloupcový=column)

L13: další možnosti

L14: vytvoří vektor se čtyřmi jedničkami

L15: posloupnost od 1 do 10 s krokem 2

L16: posloupnost od 1 do 10 s krokem 1

L17: předefinování objektu na jiný typ - např. as.vector, as.matrix, as.factor,...

L18: práce s vektory - slučování podle sloupců/řádků

L19: # definování matice

L20: diagonální matice

L21: # operace s maticemi - pozor na maticové násobení -> %\*%

L22: <hr>

L23: # Kombinatorika

L24: ## Variace

L25: V(n,k) - variace bez opakování, první argument bude celkový počet entit, druhý argument velikost výběru

L26: funkce se vytváří příkazem fucntion, je to objekt jehož jméno je dáno až proměnnou

L27: do které tento objekt přiřadím

L28: zde zadávám počet parametrů a jejich jména

L29: celé tělo funkce je uzavřeno mezi závorkami {...}

L30: faktoriál v originálním Rku existuje tak jej použijeme

L31: to co funkce vrátí se dává do příkazu return(...)

L32: V\*(n,k) - variace s opakováním

L33: ## Permutace

L34: P(n)=V(n,n) - permutace

L35: P\*(n1,n2,n3,....,nk) - permutace s opakováním, vstup bude vektor s jednotlivými počty unikátních entit

L36: vec\_n je vektro počtů hodnot př.: vec\_n = c(2,2,2,4,3)

L37: spočteme kolik máme hodnot celkem

L38: jejich faktoriál = hodnota v čitateli

L39: jednoduchý cyklus začíná příkazem for, pak v závorkách následuje název iterátoru a z

L40: jakého seznamu bude brán

L41: pocet je iterátor a postupně bude nabývat hodnot z vektoru vec\_n

L42: postupně dělíme faktoriálem každého počtu unikátních entit

L43: ## Kombinace

L44: C(n,k) - kombinace

L45: funkce for kombinace už existuje v Rku a jmenuje se choose

L46: C\*(n,k) - kombinace s opakováním

L47: použijeme známý vzorec

L48: # Úlohy na cvičení

L49: ## Příklad 1.

L50: Které heslo je bezpečnější? \* Heslo o délce osm znaků složené pouze z číslic. \* Heslo o délce pět znaků složené pouze z písmen anglické abecedy.

L51: heslo 1

L52: heslo 2

L53: ## Příklad 2.

L54: Jak dlouho by trvalo vyřešení problému obchodního cestujícího pro n = 10 měst hrubou silou, jestliže vyhodnocení délky každé z možných cest trvá 1 µs?

L55: ## Příklad 3.

L56: Jak rozdělit kořist mezi 2 loupežníky, aby dostali oba věci ve stejné hodnotě (případně co nejbližší hodnotě). Tj. lze rozdělit N zadaných čísel do dvou skupin tak, aby součet čísel v obou skupinách byl stejný? \*\*Kolik možností by bylo třeba vyzkoušet, pokud bychom úlohu řešili hrubou silou?\*\*

L57: ## Příklad 4.

L58: Kolik anagramů slova "AUTO" můžeme vytvořit? Kolik anagramů slova "AUTOMOBILKA" můžeme vytvořit? Kolik z nich začína na "K"?

L59: ## Příklad 5.

L60: V obchodě mají 6 druhů barevných hrníčků. - Kolika různými způsoby můžeme koupit 4 různě-barevné hrníčky? - Kolika různými možnostmi můžeme nakoupit 5 hrníčků (pokud nám nevadí více od stejné barvy)? - Jak se situace změní, pokud budou mít od každého pouze 4 kusy (a nám nevadí více stejné barvy)?

L61: ## Příklad 6. (sbírka kap. 1, př. 7,8)

L62: Z urny se třemi koulemi, dvěma červenými a jednou bílou, budou současně vybrány dvě koule. Student a učitel uzavřou sázku. Pokud budou obě koule stejné barvy, vyhraje student. Pokud budou mít koule různou barvu, vyhraje učitel. - Je hra férová? Jaké jsou pravděpodobnosti výhry učitele a studenta? - Jakou kouli je třeba přidat, aby hra byla férová?

L63: ## Příklad 7.

L64: V balíčku je 5 různých párů ponožek (levá a pravá ponožka je vždy stejná). - Kolik různých dvojic ponožek lze vybrat? - Kolika různými způsoby se mohu obout? (tj. záleží na tom co je na které noze)

L65: ## Příklad 8.

L66: Mám 12 závaží o hmotnostech 1,2,...,12 kg. - Kolika způsoby je mohu rozdělit na 2 hromádky? - Kolika způsoby je mohu rozdělit na 3 hromádky? - Kolika způsoby je mohu rozdělit na 3 hromádky, má-li na všech být stejný počet závaží? - Kolika způsoby je mohu rozdělit na 3 hromádky o stejném počtu závaží, pokud hmotnost žádné z nich nesmí překročit 40 kg?

L67: ## Příklad 9.

L68: Mám 20 semínek od každého ze tří druhů zeleniny (mrkev, ředkvička, celer). Bohužel se pomíchala. - Do truhlíku zasadím 5 náhodných semínek. Jaká je pravděpodobnost, že mezi nimi budou alespoň tři ředkvičky? - Do truhlíku zasadím 5 náhodných semínek. Jaká je pravděpodobnost, že mezi nimi bude více mrkví než celerů?

%%%%%%%%%% cv3

L0: # Cvičení 3 - Diskrétní náhodná veličina

L1: ## Martina Litschmannová, Adéla Vrtková, Michal Béreš

L2: # Příklady

L3: ## Příklad 1.

L4: Majitel servisního střediska nabídl prodejně automobilů, která si zřídila autopůjčovnu své služby. Za každý automobil zapůjčený jeho prostřednictvím obdrží od autopůjčovny 500,- Kč. Zároveň se však zavázal, že každý den investuje do údržby zapůjčených automobilů 800,- Kč. Počet automobilů zapůjčených prostřednictvím servisního střediska za 1 den je popsán následující pravděpodobnostní funkcí: ![image-3.png](attachment:image-3.png)

L5: ### 1. a)

L6: Hodnota pravděpodobnostní funkce pro 5 automobilů byla špatně čitelná. Určete ji:

L7: počítačová aritmetika zde může zlobit

L8: zaokrouhlíme na setiny

L9: zápis pro x=5 je 6-tá pozice

L10: Začínáme s grafikou -> graf dále upravujeme - používáme další parametry

L11: parametr pro rozsah osy y

L12: parametr pro popis osy y

L13: parametr pro název grafu

L14: určuje o jaký typ grafu se jedná (p -> points, bodový)

L15: parametr pro vzhled zobrazovaných bodů

L16: barvy

L17: parametr pro upravení velikosti celého grafu (ve smyslu zvětší 2x)

L18: parametr zvlášť pro velikost názvů os

L19: parametr zvlášť pro velikost hodnot na osách

L20: parametr pro velikost názvu grafu

L21: Pravděpodobnostní funkce

L22: plná kolečka - v skutečných hodnotách

L23: horizontální grid

L24: vertikální grid

L25: že chceme kreslit do jednoho grafu

L26: prázdná kolečka - tam kde je definovaná nenulová hodnota

L27: nastavení hodnot na X

L28: a Y

L29: Poznámky k úvodu do grafiky základem jsou tzv. high-level funkce, které vytvoří graf (tj. otevřou grafické okno a vykreslí dle zadaných parametrů) na ně navazují tzv. low-level funkce, které něco do aktviního grafického okna přidají, samy o sobě neotevřou nové výše použitá funkce "text" je low-level funkce - přidá text do stávajícího aktivního grafického okna další low-level funkce - např. abline, points, lines, legend, title, axis ... které přidají přímku, body, legendu... tzn. před použitím "low-level" funkce je potřeba, volat "high-level" funkci (např. plot, boxplot, hist, barplot, pie,...) další grafické parametry naleznete v nápovědě nebo např. zde http://www.statmethods.net/advgraphs/parameters.html nebo zde https://flowingdata.com/2015/03/17/r-cheat-sheet-for-graphical-parameters/ nebo http://bcb.dfci.harvard.edu/~aedin/courses/BiocDec2011/2.Plotting.pdf

L30: ### 1. b)

L31: Určete a zakreslete distribuční funkci náhodné veličiny X, která je definována jako počet zapůjčených automobilů.

L32: zjednodušený graf distribuční funkce

L33: Funkce pro výpočet a vykreslení distribuční funkce

L34: natáhneme F o 0 na začátku

L35: a x z obou stran

L36: prazdná kolečka

L37: že chceme kreslit do jednoho grafu

L38: plná kolečka

L39: horizontální grid

L40: vertikální grid

L41: graf - čáry

L42: nastavení hodnot na X

L43: a Y

L44: ### 1. c)

L45: Určete střední hodnotu, rozptyl, směrodatnou odchylku a modus počtu zapůjčených automobilů během jednoho dne.

L46: Střední hodnota

L47: Rozptyl

L48: druhý obecný moment

L49: Směrodatná odchylka

L50: Funkce pro výpočet základních číselných charakteristik

L51: zápis výsledků do tabulky

L52: ### 1. d)

L53: Určete pravděpodobnostní funkci a distribuční funkci náhodné veličiny Y, která je definována jako denní příjem majitele servisu.

L54: Distribuční funkce

L55: ### 1. e)

L56: Určete střední hodnotu, směrodatnou odchylku a modus příjmu majitele servisu ze zapůjčených automobilů během jednoho dne.

L57: ### 1. f)

L58: Určete pravděpodobnost, že příjem majitele servisu (náhodná veličina Y) z půjčování automobilů převýší jeho výdaje.

L59: zisk

L60: příjem převýší výdaje, když je zisk kladný

L61: ### 1. g)

L62: Určete střední hodnotu, směrodatnou odchylku a modus náhodné veličiny Z, která je definována jako zisk majitele servisu ze zapůjčených automobilů během jednoho dne.

L63: ## Příklad č. 2

L64: Pro distribuční funkci náhodné veličiny X platí: $F(x)=\begin{cases} 0 & x \leq -1 \\ 0.3 & -1 < x \leq 0 \\ 0.7 & 0 < x \leq 1 \\ 1 & -1 < x \end{cases}$

L65: ### 2. a)

L66: Určete pravděpodobnostní funkci náhodné veličiny X, její střední hodnotu a směrodatnou odchylku.

L67: ### 2. b)

L68: Náhodná veličina Y = 1 − 3X, určete P(y), F(y), E(Y), D(Y).

L69: Nesmyslný výstup - čím je to způsobeno?

L70: funkce order vrátí indexy setřízeného pořadí

L71: ### 2. c)

L72: Náhodná veličina W = $3X^2$, určete P(w), F(w), E(W), D(W).

L73: inicializace pole o stejné velikosti

L74: ## Příklad 3.

L75: V dílně jsou dva stroje pracující nezávisle na sobě. Pravděpodobnost poruchy prvního stroje je 0,2, pravděpodobnost poruchy druhého stroje je 0,3. Náhodná veličina X je definována jako počet současně porouchaných strojů. Určete:

L76: ### 3. a)

L77: pravděpodobnostní funkci náhodné veličiny X,

L78: spočteme jednotlivé pravděpodobnosti počtu porouchaných strojů

L79: 0 porouchaných tedy oba v provozu

L80: 2 tedy porouchané oba

L81: právě jeden - buď první nebo druhý

L82: ### 3. b)

L83: distribuční funkci náhodné veličiny X,

L84: ### 3. c)

L85: střední hodnotu a rozptyl náhodné veličiny X.

%%%%%%%%%% cv8

L0: # Cvičení 8. Výběrové charakteristiky ## Michal Béreš

L1: # Další vybraná spojitá rozdělení

L2: ## $\chi^2$ - Chí-kvadrát rozdělení (Pearsnovo rozdělení)

L3: - Použítí: při odhadu směrodatné odchylky (za použití výběrové) - Má jediný parametr - počet stupňů volnosti - $\frac{S^2}{\sigma^2}(n-1) \sim \chi^2\_{n-1}$ - $S$ je výběrová směrodatná odchylka

L4: počet stupňů volnosti

L5: osa x

L6: hustota pravděpodobnosti chí-kvad. rozdělení

L7: distrib. fce. chí-kvad. rozdělení

L8: ## $t$ - Studentovo rozdělení

L9: - Použití: při odhadu střední hodnoty bez přesné znalosti rozptylu (pouze výběrového rozptylu) - $\frac{\bar X - \mu}{S}\sqrt{n} \sim t\_{n-1} $ - $\bar X$ je výběrový průměr - $S$ je výběrová směrodatná odchylka - s rostoucím počtem stupňů volnosti konverguje k normovanému normálnímu rozdělení

L10: počet stupňů volnosti

L11: osa x

L12: hustota pravděpodobnosti studentova rozdělení

L13: hodnoty norm. normálního roz.

L14: do posledního grafu

L15: hustota pravděpodobnosti studentova rozdělení

L16: hodnoty norm. normálního roz.

L17: do posledního grafu

L18: ## $F$ - Fisher-Snedecorovo rozdělení

L19: - Používá se k testování schody rozptylů - $\frac{S\_1^2/\sigma\_1^2}{S\_2^2/\sigma\_2^2} \sim F\_{n\_1 - 1, n\_2 - 1}$

L20: počet stupňů volnosti výběr. 1

L21: počet stupňů volnosti výběr. 2

L22: osa x

L23: hustota pravděpodobnosti chí-kvad. rozdělení

L24: hustota pravděpodobnosti chí-kvad. rozdělení

L25: # Jak se chová průměr hodnot z normálního rozdělení?

L26: Funkce \*\*rnorm(n, mean, sd)\*\* generuje \*\*n\*\* hodnot z normálního rozdělení se střední hodnotou \*\*mean\*\* a směrodatnou odchylkou \*\*sd\*\*.

L27: ### Náhodná veličina: průměr hodnot

L28: numeric vyrobi vektor 0

L29: # Jak se chová průměr hodnot z uniformního rozdělení?

L30: Fukce \*\*runif(n, min, max)\*\* generuje \*\*n\*\* hodnot z uniformního rozdělení U(\*\*min,max\*\*).

L31: nah\_vyber

L32: ### Náhodná veličina: průměr hodnot

L33: # Příklady

L34: ## Příklad 1.

L35: Zatížení letadla s 64 místy nemá překročit 6 000 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?

L36: X...hmotnost 64 cestujících

L37: X ~ N(64\*90; 64\*100)

L38: P(X > 6000) = 1 - F(6000)

L39: ## Příklad 2.

L40: Zásilka obsahuje 300 výrobků určitého typu. Je známo, že pravděpodobnost zhotovení vadného výrobku tohoto typu je 0,04.

L41: ### a)

L42: Odhadněte pravděpodobnost, že absolutní odchylka podílu vadných výrobků v zásilce od pravděpodobnosti vyrobení vadného výrobku bude menší než 1 %.

L43: X = (p − π)/sqrt(π\*(1 − π))\*sqrt(n) ∼ N(0, 1)

L44: P(-0.01/sqrt(π\*(1 − π))\*sqrt(n) < X < 0.01/sqrt(π\*(1 − π))\*sqrt(n))

L45: ### b)

L46: Jak se změní výsledek, jestliže zásilka bude obsahovat 3 000 výrobků?

L47: ## Příklad 3.

L48: Cestující pravidelně jezdí do zaměstnání a zpět MHD. Je známo, že doba čekání na příjezd MHD se pohybuje v mezích od 0 do 3 minut. Jaká je pravděpodobnost, že celková doba čekání zaměstnance na příjezd MHD během 23 pracovních dnů bude kratší než 80 minut?

L49: Y...doba i-tého čekání na MHD

L50: y ~ R(0; 3)

L51: X...celková doba čekání během 23 dnů (cesta tam a zpět ⇒ 46 čekání)

L52: X ~ N(46\*EY; 46\*DY)

L53: P(X < 80)

L54: ## Příklad 4.

L55: Předpokládejme, že průměrná spotřeba elektrické energie domácností v určitém městě v lednu je 120 kWh a směrodatná odchylka spotřeby je 100 kWh. Určete pravděpodobnost, že průměrná spotřeba 100 náhodně vybraných domácností bude větší než 140 kWh.

L56: Xi...spotřeba i-té domácnosti

L57: X...průměrná spotřeba 100 domácností

L58: X ~ N(EXi;Dxi/n)

L59: P(X > 140)

L60: ## Příklad 5.

L61: Společnost Acme Battery Company vyvinula nový typ baterie mobilních telefonů. V průměru vydrží baterie 60 minut na jedno nabití. Směrodatná odchylka této doby je 4 minuty. Předpokládejme, že výrobní oddělení po 6 měsících spustí test kontroly kvality. Provedli dva náhodné výběry o rozsahu 10 baterií a v obou zjistili směrodatnou odchylku výdrže baterií větší než 6 minut. S jakou pravděpodobností takový výsledek mohli očekávat?

L62: X = (n − 1)S^2/σ^2

L63: X ∼ χ\_n-1

L64: P(S > 6) = P(X > ...)

L65: ## Příklad 6.

L66: Z úmrtnostních tabulek vyplývá pravděpodobnost 0,99, že se 35 - letý muž dožije dalšího roku. Roční pojistné této věkové skupiny činí 2 000 Kč, v případě úmrtí pojišťovna vyplatí 100 000 Kč. Jaká je pravděpodobnost, že zisk z 500 pojištěných mužů ve věku 35 let bude alespoň 500 000 Kč? (Řešte dvěma způsoby - pomocí binomického rozdělení a pomoci aproximace binomického rozdělení rozdělením normálním.)

L67: X...počet mužů z 500, kteří se nedožijí dalšího roku

L68: X ~ Bi(500; 0.01)

L69: Z = 500 · 2 000 − X · 100 000

L70: P(Z ≥ 500 000) = P(X ≤ 5)

L71: X ~ Bi(500; 0.01) ~ N(500\*0.01; 500\*0.01\*(1-0.01))

L72: P(X ≤ 5) ~ P(X < 5.5) (oprava na spojitost)

L73: ## Příklad 7.

L74: Předpokládejme, že v populaci má přibližně 60 % mladých mužů vyšší než doporučenou hladinu cholesterolu v séru. S jakou pravděpodobností bude mít v náhodném výběru 200 mladých mužů více než 120 z nich vyšší než doporučenou hladinu cholesterolu v séru?

L75: X...počet mladých mužů z 200 s vyšší než doporučenou hladinou cholesterolu v séru

L76: X ∼ Bi(200; 0.6)

L77: P(X > 120) = 1 - P(X ≤ 120)

L78: X ~ N(200\*0.6; 200\*0.6(1-0.6)), tj. X ≈ N(120; 48)

L79: 1 - P(X ≤ 120) ~ 1 - P(X < 120.5) (oprava na spojitost)

%%%%%%%%%% cv5

L0: # Cvičení 5 - Vybraná rozdělení diskrétní náhodné veličiny

L1: ## Martina Litschmannová, Adéla Vrtková, Michal Béreš

L2: # Přehled rozdělení a jejich funkcí

L3: ## Úvod: Pravděpodobnostní, Kumulativní pravděpodobnostní (Distribuční) a Kvantilová funkce

L4: ### Pravděpodobnostní funkce

L5: - začíná písmenkem \*\*d\*\*: $p = P(X = x)$: p = d...(x, ...)

L6: ### Kumulativní pravděpodobnostní (Distribuční funkce)

L7: - začíná písmenkem \*\*p\*\*: $p = P(X \leq x)$: p = p...(x, ...) - pozor Kumulativní pravděpodobnostní je s alternativní definicí $P(X \leq t)$ - pro naši distribuční funkci $F(t) = P(X<t)$: F(t) = p...(t - 1, ...)

L8: ### Kvantilová funkce

L9: - začíná písmenkem \*\*q\*\*: $p \geq P(X \leq x)$: x = q...(p, ...) - hledá nejmenší $x$ pro které je $P(X \leq x)$ větší než $p$

L10: ## Binomické (Alternativní): $X \sim Bi(n, π),X \sim A(π) = Bi(1, π)$

L11: - počet úspěchů v $n$ Bernoulliho pokusech (případně pro jeden pokus v případě Alternativní) - každý pokus má šanci na úspěch $π$

L12: Pravděpodobnostní funkce P(X = x)

L13: hodnota, pro níž hledáme p-stní funkci

L14: rozsah výběru

L15: pravděpodobnost úspěchu

L16: tímto se dají vypnout warningy

L17: tímto zase zapnout

L18: vykreslíme si pravděpodobnostní funkci

L19: minimálně 0, maximálně n má kladnou pravděpodobnost

L20: Kumulativní pravděpodobnostní funkce P(X <= x)

L21: hodnota, pro níž hledáme kumulativní p-stní funkci

L22: rozsah výběru

L23: pravděpodobnost úspěchu

L24: Distribuční funkce F(x) = P(X < x)

L25: hodnota, pro níž hledáme kumulativní p-stní funkci

L26: rozsah výběru

L27: pravděpodobnost úspěchu

L28: nebo

L29: vykreslíme si distribuční funkci

L30: minimálně 0, maximálně n má kladnou pravděpodobnost

L31: nebo

L32: minimálně 0, maximálně n

L33: zkontrolujeme korektnost na hodnotě 10

L34: minimálně 0, maximálně n

L35: najdi x pro dané q: q = P(X <= x)

L36: h

L37: rozsah výběru

L38: pravděpodobnost úspěchu

L39: Kvantilová funkce (inverzi dist. fce): q = F(x) = P(X < x)

L40: pravděpodobnost pro kterou hledáme kvantil

L41: rozsah výběru

L42: pravděpodobnost úspěchu

L43: ## Hypergeometrické: $X \sim H(N, M, n)$

L44: - počet úspěchů v $n$ závislých pokusech - závislost typu:

L45: - $N$ objektů,

L46: - z toho $M$ objektů se zadanou vlastností,

L47: - výběr velikosti $n$

L48: - \*\*při výběru nevracíme zpět - pravděpodobnost výběru objektu s danou vlastností se mění s každým dalším vybraným objektem\*\*

L49: - \*\*R funkce bere jako parametry \*hyper(k, M, N - M, n)\*\*

L50: - k je počet úspěchů pro které počítáme pravděpodobnost,

L51: - M je počet objektů se zadanou vlastností,

L52: - N-M je počet objektů bez zadané vlastnosti,

L53: - n je ceklová velikost výběru.

L54: Pravděpodobnostní funkce P(X = x)

L55: hodnota, pro níž hledáme p-stní funkci

L56: celkový počet objektů

L57: z toho se zadanou vlastností

L58: velikost výběru

L59: vykreslíme si pravděpodobnostní funkci

L60: minimálně 0, maximálně n nebo M má kladnou pravd.

L61: Distribuční funkce F(x) = P(X < x)

L62: hodnota, pro níž hledáme dist. funkci

L63: celkový počet objektů

L64: z toho se zadanou vlastností

L65: velikost výběru

L66: vykreslíme si Distribuční funkci

L67: minimálně 0, maximálně n nebo M má kladnou pravd.

L68: Kvantilová funkce (inverzi dist. fce): q = P(X < x)

L69: pravděpodobnost pro kterou hledáme kvantil

L70: celkový počet objektů

L71: z toho se zadanou vlastností

L72: velikost výběru

L73: ## Negativně binomické (Geometrické): $X \sim NB(k, π), X \sim Ge(π) = NB(1, π)$

L74: - počet pokusů do $k$. úspěchu (včetně) - každý pokus má šanci na úspěch $π$ - \*\*Negativně binomická NV je v Rku definována jako počet neúspěchů před k-tým úspěchem\*\*

L75: - proto jako první parametr budeme posílat x - k

L76: Pravděpodobnostní funkce P(X = x)

L77: počet pokusů pro který hledáme pravd. fci

L78: požadovaný počet úspěchů

L79: pravd. jednotlivých pokusů

L80: pozor první argument musí být počet neúspěchů

L81: vykreslíme si pravděpodobnostní funkci

L82: minimálně k, maximum neomezeno

L83: hodnoty 0,1,2,3,4 mají P(x)=0

L84: Distribuční funkce F(x) = P(X < x)

L85: počet pokusů pro který hledáme pravd. fci

L86: požadovaný počet úspěchů

L87: pravd. jednotlivých pokusů

L88: pozor první argument musí být počet neúspěchů

L89: vykreslíme si Distribuční funkci

L90: minimálně 0, maximálně n nebo M má kladnou pravd.

L91: Kvantilová funkce (inverzi dist. fce): q = P(X < x)

L92: pravd. pro kvantil

L93: požadovaný počet úspěchů

L94: pravd. jednotlivých pokusů

L95: ## Poissonovo: $X \sim Po(λt)$

L96: - počet událostí v Poissonově procesu v uzavřené oblasti (v čase, na ploše, v objemu) - s hustotou výskytu $λ$ - v čase/ploše/objemu velikosti $t$

L97: Pravděpodobnostní funkce P(X = x)

L98: počet pokusů pro který hledáme pravd. fci

L99: hustota výskytu

L100: pravd. jednotlivých pokusů

L101: vykreslíme si pravděpodobnostní funkci

L102: minimálně 0, maximum neomezeno

L103: Distribuční funkce F(x) = P(X < x)

L104: počet pokusů pro který hledáme pravd. fci

L105: hustota výskytu

L106: pravd. jednotlivých pokusů

L107: vykreslíme si Distribuční funkci

L108: minimálně 0, maximálně n nebo M má kladnou pravd.

L109: Kvantilová funkce (inverzi dist. fce): q = P(X < x)

L110: pravd. pro kvantil

L111: hustota výskytu

L112: pravd. jednotlivých pokusů

L113: # Příklady

L114: ## Příklad 1.

L115: Bridž se hraje s 52 bridžovými kartami, které se rozdají mezi 4 hráče. Vždy 2 hráči hrají spolu. Při rozdávání (13 karet) jste dostali do rukou 2 esa. Jaká je pravděpodobnost, že váš partner bude mít zbývající dvě esa?

L116: X ... počet es mezi 13 kartami

L117: X ~ H(N = 39, M = 2, n = 13)

L118: P(X = 2)

L119: 52-13

L120: výpočet

L121: což je dhyper(2,2,37,13)

L122: graf pravděpodobnostní funkce

L123: všechny možné realizace NV X

L124: hodnoty pravděpodobnostní funkce pro x

L125: ## Příklad 2.

L126: Pokusy se zjistilo, že radioaktivní látka vyzařuje během 7,5 s průměrně 3,87 α-částice. Určete pravděpodobnost toho, že za 1 sekundu vyzáří tato látka alespoň jednu α-částici.

L127: X ... počet vyzářených alfa částic během 1 s

L128: X ~ Po(lt = 3.87/7.5)

L129: četnost výskytu

L130: za 1 sekundu

L131: parametr Poissonova rozdělení

L132: P(X >= 1) = P(X > 0) = 1 - P(X <= 0)

L133: graf pravděpodobnostní funkce

L134: teoreticky může být vyzářeno až nekonečně mnoho částic,

L135: od jisté hodnoty je pravděpodobnost zanedbatelná

L136: hodnoty pravděpodobnostní funkce pro x

L137: ## Příklad 3.

L138: Kamarád vás pošle do sklepa, abyste donesl(a) 4 lahvová piva - z toho dvě desítky a dvě dvanáctky. Nevíte, kde rozsvítit, proto vezmete z basy poslepu 4 láhve. S jakou pravděpodobností jste vyhověl(a), víte-li, že v base bylo celkem 10 desítek a 6 dvanáctek?

L139: X ... počet 10°piv mezi 4 vybranými

L140: X ~ H(N = 16, M = 10, n = 4)

L141: P(X = 2)

L142: graf pravděpodobnostní funkce

L143: všechny možné realizace NV X

L144: hodnoty pravděpodobnostní funkce pro x

L145: ## Příklad 4.

L146: V jednom mililitru určitého dokonale rozmíchaného roztoku se v průměru nachází 15 určitých mikroorganismů. Určete pravděpodobnost, že při náhodném výběru vzorku o objemu 1/2 mililitru bude ve zkumavce méně než 5 těchto mikroorganismu.

L147: X ... počet mikroorganismů v 0.5 ml roztoku

L148: X ~ Po(lt = 15/2)

L149: parametr Poissonova rozd.

L150: P(X < 5) = P(X <= 4)

L151: nebo

L152: graf pravděpodobnostní funkce

L153: teoreticky může být v roztoku až nekonečně mnoho mikroorganismů,

L154: od jisté hodnoty je pravděpodobnost zanedbatelná

L155: hodnoty pravděpodobnostní funkce pro x

L156: ## Příklad 5.

L157: Na stůl vysypeme 15 mincí. Jaká je pravděpodobnost, že počet mincí ležících lícem nahoře, je od 8 do 15?

L158: X ... počet mincí, které padnou lícem nahoru z celkového množství 15 mincí

L159: X ~ Bi(n = 15, p = 0.5)

L160: P(8 <= X <= 15) = P(X <= 15) - P(X < 8) = P(X <= 15) - P(X <= 7)

L161: jinak: P(8<=X<=15)=P(X>7)=1-P(X<=7)

L162: graf pravděpodobnostní funkce

L163: všechny možné realizace NV X

L164: hodnoty pravděpodobnostní funkce pro x

L165: ## Příklad 6.

L166: Pravděpodobnost, že se dovoláme do studia rozhlasové stanice, která právě vyhlásila telefonickou soutěž je 0,08. Jaká je pravděpodobnost, že se dovoláme nejvýše na 4. pokus?

L167: X ... počet pokusů než se dovoláme do rozhlasového studia

L168: X ~ NB(k = 1,p = 0.08) nebo G(0.08)

L169: P(X <= 4)

L170: graf pravděpodobnostní funkce

L171: teoreticky můžeme uskutečnit až nekonečně mnoho pokusů,

L172: od jisté hodnoty je pravděpodobnost zanedbatelná

L173: hodnoty pravděpodobnostní funkce pro x

L174: ## Příklad 7.

L175: V továrně se vyrobí denně 10 % vadných součástek. Jaká je pravděpodobnost, že vybereme-li třicet součástek z denní produkce, tak nejméně dvě budou vadné?

L176: X ... počet vadných součástek ze 30 vybraných

L177: X ~ Bi(n = 30, p = 0.1)

L178: P(X >= 2) = 1 - P(X < 2) = 1 - P(X <= 1)

L179: nebo P(X >= 2) vsechno mimo 0 a 1

L180: graf pravděpodobnostní funkce

L181: všechny možné realizace NV X

L182: hodnoty pravděpodobnostní funkce pro x

L183: ## Příklad 8.

L184: Ve skladu je 200 součástek. 10 % z nich je vadných. Jaká je pravděpodobnost, že vybereme-li ze skladu třicet součástek, tak nejméně dvě budou vadné?

L185: X ... počet vadných součástek ze 30 vybraných z 200

L186: X ~ H(N = 200, M = 20, n = 30)

L187: P(X >= 2) = 1 - P(X < 2) = 1 - P(X <= 1)

L188: graf pravděpodobnostní funkce

L189: všechny možné realizace NV X

L190: hodnoty pravděpodobnostní funkce pro x

L191: ## Příklad 9.

L192: V určité firmě bylo zjištěno, že na 33 % počítačů je nainstalován nějaký nelegální software. Určete pravděpodobnostní a distribuční funkci počtu počítačů s nelegálním softwarem mezi třemi kontrolovanými počítači.

L193: X ... počet počítačů s nelegálním softwarem ze 3 kontrolovaných

L194: X ~ Bi(n = 3,p = 0.33)

L195: pravděpodobnostní funkce

L196: všechny možné realizace NV X

L197: hodnoty pravděpodobnostní funkce pro x

L198: zaokrouhlení pravděpodobností na 3 des. místa

L199: dopočet poslední hodnoty do 1

L200: vytvoření tabulky pravděpodobnostní funkce

L201: graf pravděpodobnostní funkce

L202: distribuční funkce

L203: zjednodušený výpis distribuční funkce

L204: ## Příklad 10.

L205: Sportka je loterijní hra, v níž sázející tipuje šest čísel ze čtyřiceti devíti, která očekává, že padnou při budoucím slosování. K účasti ve hře je nutné zvolit alespoň jednu kombinaci 6 čísel (vždy 6 čísel na jeden sloupec) a pomocí křížků tato čísla označit na sázence společnosti Sazka a.s. do sloupců, počínaje sloupcem prvním. Sázející vyhrává v případě, že uhodne alespoň tři čísla z tažené šestice čísel. Jaká je pravděpodobnost, že proto, aby sázející vyhrál, bude muset vyplnit:

L206: Nejprve pravděpodobnost, že vehrajeme v jednom sloupci

L207: Y ... počet uhádnutých čísel v 6 tažených ze 49

L208: Y ~ H(N = 49, M = 6, n = 6)

L209: P-st uhádnutí alespoň 3 čísel v jednom sloupci

L210: P(Y >= 3) = 1 - P(Y < 3) = 1 - P(Y <= 2)

L211: ### a)

L212: právě tři sloupce,

L213: X … počet sloupců, které bude muset sázející vyplnit, aby vyhrál

L214: X ~ NB(k = 1, p = pp)

L215: a) P(X = 3)

L216: ### b)

L217: alespoň 5 sloupců,

L218: b) P(X >= 5) = 1 - P(X < 5) = 1 - P(X <= 4)

L219: ### c)

L220: méně než 10 sloupců,

L221: c) P(X < 10) = P(X <= 9)

L222: ## d)

L223: více než 5 a nejvýše 10 sloupců?

L224: P(5 < X <= 10) = P(X <= 10) - P(X <= 5)

L225: nebo P(X < 11) - P(X < 6)

L226: ## Příklad 11.

L227: Pravděpodobnost, že hodíme 6 na 6stěnné kostce je 1/6. Hážeme tak dlouho, než hodíme šestku 10 krát.

L228: ### a)

L229: Jaká je střední hodnota počtu hodů.

L230: X … hodů kostkou než hodíme 10 šestek

L231: X ~ NB(k = 10, p = 1/6)

L232: ### b)

L233: S kolika hody nejméně musíme počítat, pokud chceme, aby pradvěpodobnost, že se nám podaří naházet 10 šestek, byla alespoň 70%.

L234: P(X <= k) >= 0.7

%%%%%%%%%% cv4\_SNV

L0: # Cvičení 4 - Spojitá náhodná veličina

L1: ## Martina Litschmannová, Adéla Vrtková, Michal Béreš

L2: \*\*Obsah tohoto skriptu je pouze jako doplňující ilustrace k cvičení, není nutno znát ke zkoušce. Důležité je to umět spočítat ručně.\*\*

L3: ## Numerická integrace v Rku

L4: Rkovská funkce \*\*integrate\*\*<br> integrate(f, a, b) = $\int\_{a}^{b}f(x)dx$ - \*\*f\*\* je Rková funkce (námi definovaná) která má jeden vstupní argument - vektor hodnot ve kterém má vrátit své hodnoty - \*\*a\*\* dolní integrační mez - \*\*b\*\* horní integrační mez

L5: x^2

L6: # Příklady

L7: ## Příklad 1.

L8: Náhodná veličina X má distribuční funkci<br> $F(x)=\begin{cases} 0 & x \leq 0 \\ cx^2 & 0 < x \leq 1 \\ 1 & 1 < x \end{cases}$<br> Jaké hodnoty může nabývat konstanta c?

L9: derovací F(x) získáme hustotu pravd. f(x)

L10: příslušná hustota pravděpodobnosti na intrvalu <0,1>

L11: f(x) = 2x

L12: c = 1, proto distribuční funkce vypadá takto:

L13: x^2

L14: 0 pro x<=0

L15: 1 pro x>1

L16: body na ose x

L17: hodnoty F(x)

L18: vykreslit jako čáru

L19: ## Příklad 2.

L20: Rozdělení náhodné veličiny X je dáno hustotou<br> $f(x)=\begin{cases} 2x+2 & x \in <-1;0> \\ 0 & x \notin <-1;0> \end{cases}$<br> Určete:

L21: ### 2. a)

L22: $F(x)$,

L23: pozor na x<-1 protože '<-' je v rku přiřazení

L24: 0 pro x<=0

L25: 1 pro x>1

L26: body na ose x

L27: hodnoty f(x)

L28: vykreslit tečky (cex je velikost)

L29: x^2+2x+1

L30: 0 pro x<=0

L31: 1 pro x>1

L32: body na ose x

L33: hodnoty f(x)

L34: vykreslit tečky (cex je velikost)

L35: ### 2. b)

L36: P(−2 ≤ X ≤ −0.5), P(−2 ≤ X ≤ −1), P(X > 0.5), P(X = 0.3)

L37: P(−2 ≤ X ≤ −0.5)

L38: P(−2 ≤ X ≤ −1)

L39: P(X > 0.5)

L40: tohle nebude vždy fungovat

L41: P(X = 0.3)

L42: je jasné že tato pravděpodobnost je 0

L43: odpovídá integrálu s a=b tedy s nulovou velikostí na ose x

L44: ### 2. c)

L45: střední hodnotu, rozptyl a směrodatnou odchylku náhodné veličiny X.

L46: E(X)

L47: integrujeme jen tam kde víme, že je f(x) nenulová

L48: E(X^2)

L49: integrujeme jen tam kde víme, že je f(x) nenulová

L50: D(X)

L51: sigma(x)

L52: ### 2. d)

L53: modus $\hat{x}$

L54: modus = 0

L55: ### 2. e)

L56: medián $x\_{0,5}$

L57: body na ose x

L58: první prvek z x pro který F(x)>=0.5

L59: ## Příklad 3.

L60: Náhodná veličina Y je definována jako: Y = 3X+1, kde X je náhodná veličina z předcházejícího příkladu. Určete:

L61: ### 3. a)

L62: $F\_Y(y)$

L63: spočteno ze vztahu FY(y) = P(Y < y) = P(3X + 1 < y) = ...

L64: body na ose x

L65: ### 3. b)

L66: $f\_Y(y)$

L67: derivace F\_Y

L68: 0 pro x<-2

L69: 1 pro x>1

L70: kontrola celkového integrálu

L71: body na ose x

L72: ### 3. c)

L73: E(Y), D(Y), σ(Y)

L74: E(Y)

L75: integrujeme jen tam kde víme, že je f(y) nenulová

L76: alternativně

L77: E(Y^2)

L78: integrujeme jen tam kde víme, že je f(y) nenulová

L79: D(Y)

L80: alternativně

L81: sigma(Y)

L82: ## Příklad 4. (není ze sbírky)

L83: Spočtěte $\omega$ takové, aby náhodná veličina X s hustotou pravděpodobnosti:<br> $f(x)=\begin{cases} 0 & x < 0 \\ 3e^{-3x} & x \geq 0 \end{cases}$ <br> byla s pravděpodobností 0.3 větší než $\omega$

L84: 0 pro x<=0

L85: body na ose x

%%%%%%%%%% cv10

L0: # Cvičení 10. Úvod do testování hypotéz, jednovýběrové testy ## Michal Béreš, Martina Litschmannová

L1: # Od intervalových odhadů k testům hypotéz

L2: ## Co je to statistický test hypotéz?

L3: Mějme následující: - náhodná veličina X (například výška mužů) - výběr z náhodné veličiny (měření výšky 30 mužů) Statistické testování hypotéz rozhoduje na základě získaných dat z náhodného výběru o platnosti: - $H\_0$ - nulové hypotézy - $H\_A$ - alternativní hypotézy Například:<br> $H\_0$: $\mu\_X = 175$<br> $H\_A$: $\mu\_X > 175$<br> Jelikož se jedná o statistické rozhodnutí, vždy bude vázáno k nějaké hladnině významnosti $\alpha$. Vždy můžeme dospět pouze k 2 různým rozhodnutím: - Zamítám $H\_0$ ve prospěch $H\_A$ - to znamená, že tvrdím, že $H\_0$ neplatí - toto rozhodnutí je s maximální chybou $\alpha$ (hladina významnosti, chyba I. druhu) - to znamená, že velikost této chyby jsme schopni ovlivnit - Nezamítám $H\_0$ - to znamená, že tvrdím, že vzhledem k získaným datům (výběr) nelze vyvrátit $H\_0$ - toto rozhodnutí je s chybou $\beta$ (chyba II. druhu), tato chyba není přímo ovlivnitelná a záleží na typu použitého testu Jak testy hypotéz souvisí s intervalovými odhady a jak do nich vstupuje hladina významnosti si ukážeme v další části.

L4: ## Intervalový odhad a hladina významnosti

L5: šířka grafů v Jupyteru

L6: matice grafů 1x2

L7: šikmost

L8: špičatost

L9: test normality

L10: Vyrobíme 95% intervalový odhad střední hodnoty pomocí t-testu:

L11: Představme si nyní, že chceme testovat hypotézu:<br> $H\_0$: $\mu = 100$<br> $H\_A$: $\mu \neq 100$<br> Jaké by bylo rozhodnutí vzhledem k spočtenému IO a tedy hladině významnosti $\alpha = 0.05$?

L12: Představme si dále, že chceme testovat hypotézu:<br> $H\_0$: $\mu = 105$<br> $H\_A$: $\mu \neq 105$<br> Jaké by bylo rozhodnutí vzhledem k spočtenému IO a tedy hladině významnosti $\alpha = 0.05$?

L13: \*\*To co jsme právě udělali se nazývá klasický test.\*\*<br> Ukážeme si ještě klasické testy pro jednostranné alternativy.<br> $H\_0$: $\mu = 105$<br> $H\_A$: $\mu > 105$<br>

L14: $H\_0$: $\mu = 105$<br> $H\_A$: $\mu < 105$<br>

L15: Všimněte si, že první z těchto jednostranných alternativ vedla k "nezamítnutí" $H\_0$. Je to z důvodu porovnávání nepravděpodobné $H\_0$ s ještě méně pravděpodobnou $H\_A$.

L16: #### Čistý test významnosti a souvislost s IO

L17: Alternativou ke klasickému testu (kde vytváříme IO - v terminologii klasických testů tzv. obor přijetí a jeho doplněk do R kritický obor) je tzv. čistý test významnosti:

L18: H\_0: mu = 105

L19: H\_A: mu <> 105

L20: Výsledkem čistého testu významnosti je p-hodnota. Na jejím základě rozhodujeme o zamítnutí či nezamítnutí $H\_0$.<br> p-hodnota se dá chápat jako nejvyšší možná hladina váznamnosti, taková aby naše rozhodnutí bylo - nezamítám. Tedy IO/obor přijetí by obsahoval zkoumanou hodnotu:

L21: H\_0: mu = 105

L22: H\_A: mu <> 105

L23: H\_0: mu = 105

L24: H\_A: mu > 105

L25: H\_0: mu = 105

L26: H\_A: mu < 105

L27: ## Přehled testů

L28: ### Míry polohy

L29: Mírami polohy rozumíme údaj určující polohu dat, nehledě na tom jak jsou rozptýlená. Pro data z normálního rozdělení můžeme odhadovat střední hodnotu, pro ostatní medián.

L30: #### a) studentův t-test

L31: - testujeme střední hodnotu - data musejí pocházet z normálního rozdělení - exploračně: šikmost a špičatost leží v (-2,2) - exploračně: QQ graf má body přibližně na čáře - exaktně: pomocí statistického testu, např. Shapiro-Wilk test (shapiro.test(data))

L32: H\_0: mu = 100

L33: H\_A: mu <> 100

L34: H\_0: mu = 100

L35: H\_A: mu > 100

L36: H\_0: mu = 100

L37: H\_A: mu < 100

L38: #### b) Wilcoxnův test

L39: - testujeme medián - data musejí pocházet ze symetrického rozdělení - exploračně: šikmost leží v (-2,2) - exploračně: histogram vypadá přibližně symetricky - exaktně: pomocí statistického testu, např. balíček "lawstat", funkce "symmetry.test(data,boot=FALSE)"

L40: H\_0: X\_0.5 = 100

L41: H\_A: X\_0.5 <> 100

L42: H\_0: X\_0.5 = 100

L43: H\_A: X\_0.5 > 100

L44: H\_0: X\_0.5 = 100

L45: H\_A: X\_0.5 < 100

L46: #### c) znaménkový test test

L47: - testujeme medián - výběr většího rozsahu (>10) - vyžaduje knihovnu "BSDA" - jakožto nejrobustnější test, se dá použít i na nespojitá data - např. pořadí v nějakém seznamu

L48: H\_0: X\_0.5 = 100

L49: H\_A: X\_0.5 <> 100

L50: H\_0: X\_0.5 = 100

L51: H\_A: X\_0.5 > 100

L52: H\_0: X\_0.5 = 100

L53: H\_A: X\_0.5 < 100

L54: ### Míry variability

L55: Mírami variability rozumíme údaj určující rozptýlenost/variabilitu dat, nehledě na celkových hodnotách. Pro data z normálního rozdělení můžeme odhadovat směrodatnou odchylku.

L56: #### test směrodatné odchylky

L57: - testujeme směrodatnou odchylku - data musejí pocházet z normálního rozdělení - exploračně: šikmost a špičatost leží v (-2,2) - exploračně: QQ graf má body přibližně na čáře - exaktně: pomocí statistického testu, např. Shapiro-Wilk test (shapiro.test(data)) - vyžaduje balíček "EnvStats" - funkce v Rku, porovnává rozptyl!!!

L58: H\_0: sigma = 10

L59: H\_A: sigma <> 10

L60: H\_0: sigma = 10

L61: H\_A: sigma > 10

L62: H\_0: sigma = 10

L63: H\_A: sigma < 10

L64: ## Pravděpodobnost výskytu u jednoho výběru

L65: #### IO pravděpodobnosti

L66: - testujeme pravděpodobnost - vyžadujeme dostatečný počet dat: $n>\frac{9}{p(1-p)}$ - Clopperův - Pearsonův odhad (binom.test) - jako parametr nebere data, ale počet úspěchů a počet pozorování

L67: H\_0: pi = 0.2

L68: H\_A: pi <> 0.2

L69: H\_0: pi = 0.2

L70: H\_A: pi > 0.2

L71: H\_0: pi = 0.2

L72: H\_A: pi < 0.2

L73: # Příklady

L74: ## Příklad 1.

L75: Máme výběr 216 pacientů a změřili jsme jejich bílkovinné sérum (soubor testy\_jednovyberove.xlsx list bilk\_serum). Ověřte, zda se průměrné bílkovinné sérum (Albumin) všech pacientů tohoto typu (populační průměr µ) statisticky významně liší od hodnoty 35 g/l.

L76: Načtení dat z xlsx souboru (pomoci balíčku readxl)

L77: Explorační analýza

L78: sd zaokrouhlujeme na 3 platné cifry

L79: sd a míry polohy zaokrouhlujeme na tisíciny

L80: \*\*Test na míru polohy\*\*

L81: Ověření normality - exploračně

L82: šikmost

L83: špičatost

L84: šířka grafů v Jupyteru

L85: matice grafů 1x2

L86: Pro konečné rozhodnutí o normalitě dat použijeme test normality.

L87: Předpoklad normality ověříme Shapirovovým - Wilkovovým testem.

L88: H0: Data jsou výběrem z normálního rozdělení.

L89: Ha: Data nejsou výběrem z normálního rozdělení.

L90: p-value > 0.05 -> Na hl. významnosti 0,05 nelze předpoklad normality zamít.

L91: normalita OK -> t.test

L92: H0: mu = 35 g/l

L93: Ha: mu <> 35 g/l

L94: p-value < 0.05 -> Na hl. významnosti 0,05 zamítáme nulovou hypotézu

L95: ve prospěch hypotézy alternativní

L96: Střední hodnota albuminu se statisticky významně liší od 35 g/l.

L97: ## Příklad 2.

L98: V souboru testy\_jednovyberove.xlsx list preziti jsou uvedeny doby přežití pro 100 pacientů s rakovinou plic léčených novým lékem. Z předchozích studií je známo, že průměrné přežití takových pacientů bez podávání nového léku je 22,2 měsíce. Lze na základě těchto dat usoudit, že nový lék prodlužuje přežití?

L99: Načtení dat z xlsx souboru (pomoci balíčku readxl)

L100: # Explorační analýza

L101: matice grafů 1x2

L102: \*\*Data obsahují OP -> můžeme je odstranit. Nebo si také všimnout, že se pravděpdobně jedná o exponenciální rozdělení a OP tam ve skutečnosti nejsou (rozdělení se tak prostě chová).\*\*

L103: Data obsahují odlehlá pozorování. Pomoci f-ce boxplot je umíme vypsat.

L104: rozhodli-li jsme se pro odstranění odlehlých hodnot, pak

L105: doporučujeme nepřepisovat původní data

L106: # Explorační analýza pro data bez odlehlých pozorování

L107: sd zaokrouhlujeme na 3 platné cifry

L108: sd a míry polohy zaokr. na desetiny

L109: \*\*Test o míře polohy (střední hodnotě / mediánu)\*\*

L110: Ověření normality - exploračně

L111: matice grafů 1x2

L112: QQ - graf i hist. ukazují, že výběr pravd. není výběrem z norm. rozdělení.

L113: Šikmost i špičatost odpovídá norm. rozdělení.

L114: použijeme test normality.

L115: Předpoklad normality ověříme Shapirovovým . Wilkovovým testem.

L116: p-value < 0.05 -> Na hl. významnosti 0.05 zamítáme předpoklad normality

L117: explorační posouzení symetrie - výše hist. a šikmost

L118: Předpoklad symetrie - ověření testem

L119: H0: data pocházejí ze symetrického rozdělení

L120: HA: ~H0

L121: p-value < 0.05 -> Na hl. významnosti 0.05 zamítáme předpoklad symetrie

L122: normalita zamítnuta -> symetrie zamítnuta -> Sign. test

L123: H0: median = 22,2 měsíců

L124: Ha: median > 22,2 měsíců

L125: p-value > 0.05 -> Na hl. významnosti 0,05 nelze zamítnout nulovou hypotézu

L126: Medián doby přežití není statisticky významně větší než 22,2 měsíců.

L127: H0: median = 22,2 měsíců

L128: Ha: median < 22,2 měsíců

L129: ## Příklad 3.

L130: Automat vyrábí pístové kroužky o daném průměru. Výrobce udává, že směrodatná odchylka průměru kroužku je 0,05 mm. K ověření této informace bylo náhodně vybráno 80 kroužků a vypočtena směrodatná odchylka jejich průměru 0,04 mm. Lze tento rozdíl považovat za statisticky významný ve smyslu zlepšení kvality produkce? Ověřte čistým testem významnosti. Předpokládejte, že průměr pístových kroužků má normální rozdělení.

L131: Test o směrodatné odchylce

L132: Předpokládáme normalitu dat (dle zadání)

L133: rozsah souboru

L134: mm .... výběrová směrodatná odchylka (bodový odhad sm. odchylky)

L135: H0: sigma = 0.05 mm

L136: Ha: sigma < 0.05 mm

L137: p.hodnota < 0.05 -> Na hladině významnosti 0,05 zamítáme nulovou hypotézu

L138: ve prospěch alternativní hypotézy

L139: Směr. odchylka průměru kroužku je statisticky významně menší než 0,05 mm.

L140: ## Příklad 4.

L141: Automat vyrábí pístové kroužky o daném průměru. Výrobce udává, že směrodatná odchylka průměru kroužku je 0,05 mm. K ověření této informace bylo náhodně vybráno 80 kroužků a byl změřen jejich průměr (soubor testy\_jednovyberove.xlsx list krouzky). Lze zjištěné výsledky považovat za statisticky významné ve smyslu zlepšení kvality produkce? Ověřte čistým testem významnosti.

L142: Načtení dat z xlsx souboru (pomoci balíčku readxl)

L143: # Explorační analýza

L144: Data obsahují odlehlá pozorování. Pomoci f-ce boxplot je umíme vypsat.

L145: rozhodli-li jsme se pro odstranění odlehlých hodnot, pak

L146: Explorační analýza pro data bez odlehlých pozorování

L147: sd zaokrouhlujeme na 3 platné cifry

L148: sd a míry polohy zaokr. na tisíciny

L149: Ověření normality - exploračně

L150: matice grafů 1x2

L151: Šikmost i špičatost odpovídá norm. rozdělení.

L152: Pro konečné rozhodnutí o normalitě dat použijeme

L153: test normality.

L154: Předpoklad normality ověříme Shapirovovým . Wilkovovým testem.

L155: p-value > 0.05 -> Na hl. významnosti 0,05 nelze předpoklad norm. zamítnout

L156: test na míru variability -> test o rozptylu

L157: H0: sigma = 0,05 mm

L158: Ha: sigma < 0,05 mm

L159: p-value < 0.05 -> Na hladině významnosti 0,05 zamítáme H0 ve prospěch Ha

L160: Jak najít 95% intervalový odhad směrodatné odchylky?

L161: ## Příklad 5.

L162: Firma TT udává, že 1% jejich rezistorů nesplňuje požadovaná kritéria. V testované dodávce 1000 ks bylo nalezeno 15 nevyhovujících rezistorů. Potvrzuje tento výsledek tvrzení TT? Ověřte čistým testem významnosti.

L163: rozsah výběru

L164: počet "úspěchů"

L165: relativní četnost (bodový odhad pravděpodobnosti)

L166: Ověření předpokladů

L167: Dále předpokládáme n/N < 0.05, tj. že daná populace (rezistorů) má rozsah

L168: alespoň 1000/0.05 = 1000\*20 = 20 000 rezistorů

L169: # Clopperův - Pearsonův (exaktní) test

L170: # H0: pi = 0.01

L171: # Ha: pi <> 0.01

L172: # Clopperův - Pearsonův (exaktní) test

L173: # H0: pi = 0.01

L174: # Ha: pi > 0.01

L175: Na hladině významnosti 0,05 nezamítáme H0

L176: Nelze očekávat, že podíl vadných rezistorů ve výrobě statisticky významně

L177: převyšuje 1 %.

%%%%%%%%%% cv7

L0: # Cvičení 7. Preprocesing dat a explorační analýza ## Adéla Vrtková, Martina Litschmannová, Michal Béreš

L1: # 1. Rozšiřující balíčky funkcí - instalace a načítání

L2: Instalování balíčků nutné pouze jednou (pokud je již nemáte)

L3: install.packages("readxl")

L4: install.packages("dplyr")

L5: install.packages("openxlsx")

L6: Načtení balíčku (nutno opakovat při každém novém spuštění Rka, vhodné mít na začátku skriptu)

L7: obsahuje upozornění na přepsané funkce případně na starší verzi balíčku

L8: # 2. Pracovního adresář (working directory) - odkud načítáme a kam ukládáme data

L9: - Pozor aktuální otebvřená složka v Rstudiu, případně umístění Rskriptu není automaticky pracovní adresář

L10: Výpis pracovního adresáře

L11: Nastavení pracovního adresáře -> do uvozovek, celou cestu (relativní nebo absolutní)

L12: kde jsme teď

L13: zase zpátky

L14: kontrola

L15: # 3. Načtení datového souboru

L16: ## Ze souboru CSV

L17: Základní funkce - read.table, read.csv, read.csv2, ... Záleží hlavně na formátu souboru (.txt, .csv), na tzv. oddělovači jednotlivých hodnot, desetinné čárce/tečce

L18: Načtení a uložení datového souboru ve formátu csv2 z pracovního adresáře

L19: Načtení a uložení datového souboru ve formátu csv2 z lokálního disku do datového rámce data

L20: Načtení a uložení datového souboru ve formátu csv2 z internetu do datového rámce data

L21: ## Z Excelu (souboru xlsx)

L22: Načtení a uložení datového souboru ve formátu xlsx z lokálního disku do datového rámce data Používáme funkci z balíčku readxl, který jsme v úvodu rozbalili

L23: specifikace listu v xlsx souboru

L24: řádky, které se přeskočí

L25: ## Odstranění nepotřebných řádků/sloupců a pojmenování řádků/sloupců pro snadnější adresování dat

L26: indexování se zápornými indexy vrátí vše kromě hodnoty indexů

L27: nemíchat záporné a kladné indexy!

L28: odstraníme první sloupec s indexy

L29: Přejmenování sloupců - je-li nutné

L30: #### Poznámka (kterou je dobré dočíst až do konce....)

L31: (v Rstudiu) je možné importovat pomocí "Import Dataset" z okna Environment bez nutnosti psát kód V tom případě ale nesmí být v "cestě" k souboru žádné speciální znaky (háčky, čárky). Jinak se objeví error. Objekt importovaný touto cestou bude v novém RStudiu jako typ "tibble". Jedná se o modernější "data.frame" a v některých funkcích může dělat problémy a házet errory! Jednoduše lze tento objekt převést na typ data.frame pomocí \*\*as.data.frame()\*\* Pokud budete mít problém, s tím, že nějaká funkce nebude brát sloupec z "tibble" jakožto non-numeric output, můžete to napravit příkazem pull: data[,1] nahradit pull(data,1)

L32: # 4. Pre-processing dat + knihovna Dplyr

L33: ### Přehled funkcí knihovny Dplyr

L34: - \*\*%>%\*\* je takzvaný pipe operátor, typické využití je "res = data %>% operace", kde výsledkem je operace opalikovaná na data - \*\*select(...)\*\* je jednou z operací kterou můžeme vložit do "pipe" operátoru - slouží k výběru dat

L35: - select(1) - vybere první sloupec

L36: - select(A5) - vybere sloupec se jménem A5

L37: - select(1,3,5) - vybere sloupce 1,3,5

L38: - \*\*mutate(novy\_sloupec=...)\*\* je operace, které vyrobí v datovém rámci nový datový sloupec pomocí zadaného výpočtu nad aktuálními sloupci

L39: - data %>% mutate(C=A-B) vyrobí v datovém rámci "data" nový sloupec s názvem "C" jako rozdíl hodnot ve stávajícím sloupci "A" a "B"

L40: - \*\*filter(...)\*\* vyfiltruje z dat hodnoty splňující zadané požadavky

L41: - data %>% filter(vyrobce=="A" | vyrobce=="B") vrátí datový soubor, který má ve sloupci "vyrobce" pouze hodnoty "A" nebo "B"

L42: - data %>% filter(vyrobce=="A", hodnoty>1000) pokud požadavky píšeme za sebou (oddělené čárkou) chápeme to jako a zároveň

L43: - \*\*summarise(...)\*\* vypočte předepsanéčíslené charakteristiky v rámci zadaných sloupců (vhodné pro kombinaci s group.by)

L44: - data %>% summarise(prum=mean(kap5),median=median(kap5))

L45: - \*\*arrange(...)\*\* vzestupné, případně sestupné seřazení řádků

L46: - data %>% arrange(pokles) vzestupně

L47: - data %>% arrange(desc(pokles)) sestupně

L48: - \*\*group\_by(...)\*\* seskupení dat dle unikátních hodnot v zadaném sloupci

L49: - data %>% group\_by(vyrobce)

L50: Velice užitečný Dplyr "cheat sheet" naleznete zde: https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf

L51: ### Výběry sloupců/řádků

L52: Výpis datového souboru

L53: Zobrazení prvních šesti řádků

L54: Zobrazení posledních šesti řádků

L55: Zobrazení 10. řádku

L56: Zobrazení 3. sloupce - několik způsobů

L57: nebo (víme-li, jak se jmenuje proměnná zapsána ve 3. sloupci)

L58: nebo pomocí funkce select balíčku dplyr, která vybere zvolené sloupce

L59: <hr>

L60: Uložení prvního a pátého sloupce dat. rámce data do dat. rámce pokus

L61: nebo pomocí funkce z dplyr

L62: nebo pomocí názvů

L63: <hr> Vylučování dat ze souboru.

L64: Vyloučení prvního a pátého sloupce z dat. rámce data a uložení do dat. rámce pokus

L65: nebo pomocí dplyr

L66: nebo pomocí názvů

L67: <hr> Úprava dat do několika menších logických celků s různou strukturou Pozn. při ukládání dat mysleme na přehlednost v názvech

L68: ### Základní převod jednoduché datové matice do standardního datového formátu - stack(...)

L69: z dat vybereme ty sloupce, které odpovídají měřením po 5 cyklech

L70: přejmenujeme sloupce

L71: a převedeme do st. datového formátu

L72: a ještě jednou upravíme názvy sloupců

L73: Totéž provedeme pro měření provedené po 100 cyklech

L74: z dat vybereme ty sloupce, které odpovídají měřením po 100 cyklech

L75: přejmenujeme sloupce

L76: a převedeme do st. datového formátu

L77: a ještě jednou upravíme názvy sloupců

L78: Nakonec si ještě vytvoříme datový soubor ve st. datovém formátu se všemi údaji

L79: sloučení "podle sloupců"

L80: vynecháme nadbytečný druhý sloupec

L81: vynecháme řádky s NA hodnotami

L82: \*\*!!! S funkci na.omit zacházejte extrémně opatrně, aby jste nechtěně nepřišli o data !!!\*\*

L83: <hr>

L84: ### Definování nových sloupců v datovém rámci

L85: Definování nové proměnné pokles

L86: nebo pomocí funkce z balíčku dplyr

L87: ### Vybírání dat ze standardního datového formátu

L88: Může se hodit - vytvoření samostatných proměnných

L89: Třída (typ) numeric

L90: takto s výsledkem typu data frame

L91: vyfiltruje řádky odpovídající výrobci A

L92: Vybere pouze hodnoty ve sloupci kap5,

L93: Ostatní samostatné proměnné (uveden pouze jeden způsob)

L94: ### Podrobnější okénko do funkcí knihovny Dplyr - práce nad daty ve standardním datovém formátu

L95: Je nutné aplikovat na data ve st. datovém formátu !!! Operátor pipe %>% - pomáhá při řetězení funkcí - v novém RStudiu klávesová zkratka Ctrl+Shift+M

L96: #### filter - aplikuje filtr na daný sloupec

L97: filter - vybere / vyfiltruje řádky na základě daných podmínek

L98: Výběr výrobků od výrobce A

L99: Výběr výrobků od výrobce A nebo B

L100: | oddělující podmínky odpovídá logickému "nebo"

L101: Výběr všech výrobků s poklesem o 200 mAh a větším od výrobce C

L102: čárka oddělující podmínky odpovídá logickému "a zároveň"

L103: #### mutate - vyrobí nový sloupec

L104: mutate - přidá novou proměnnou nebo transformuje existující

L105: Vytvoření nového sloupce pokles\_Ah, který údává pokles kapacit v Ah (původní data v mAh, 1 Ah = 1000 mAh)

L106: pozor! pokud výsledek s nový sloupcem nikam neuložíme, tak se pouze vypíše a zmizí

L107: #### summarise - generuje souhrnné charakteristiky různých proměnných

L108: Výpočet průměru a mediánu všech hodnot proměnné kap5

L109: #### arrange - seřadí řádky podle zvolené proměnné

L110: Vzestupné a sestupné seřazení řádků podle hodnoty poklesu

L111: #### group\_by - seskupí hodnoty do skupin podle zvolené proměnné

L112: tabulka je "virtuálně" rozdělená na skupiny pro pozdější zpracování např. summarise

L113: Ideální pro spočítání sumárních charakteristik pro každého výrobce zvlášť, např. průměru

L114: \*\*Závěrečná poznámka k dplyr (kterou je dobré dočíst až do konce...) Některé operace mohou vyhodit objekt typu "tibble". Jedná se o modernější data.frame, nicméně v některých funkcích může dělat problémy a způsobovat chybová hlášení! Jednoduše lze tento "tibble" objekt převést na typ data.frame pomocí as.data.frame().\*\*

L115: # 5. Převod dat do standardního datového formátu (u dvou nejčastějších formátu dat)

L116: ## Z dat ve formátu Datová matice

L117: ### Funkce reshape

L118: Její parametry: - \*\*data\*\* - data k převedení musí být fe formátu data.frame (as.data.frame(data)) - \*\*direction\*\* - kterým směrem chceme transformaci udělat - "long" - do standardního formátu - "wide" - zpátky do datové matice - \*\*varying\*\* - názvy sloupců, které označují stejná data pro různé kategorie - je to list vektorů - každá položka listu je jedno měření - každý vektor je pak seznam sloupců - \*\*v.names\*\* - názvy sloupců ve st. dat. formátu - počet názvů musí sedět na počet vektorů v varying - \*\*times\*\* - názvy jednotlivých kategorií - POZOR!! musí být ve stejném pořadí jako u proměné varying - \*\*timevar\*\* - název sloupce s kategoriemi

L119: a pokud bychom chtěli, můžeme převést data zpět

L120: ## Z datového souboru, kde jsou kategorie v jednotlivých listech excelu

L121: # 6. Explorační analýza a vizualizace kategoriální proměnné

L122: ### Poznámky ke grafice v R

L123: základem jsou tzv. high-level funkce, které vytvoří graf (tj. otevřou grafické oknou a vykreslí dle zadaných parametrů) na ně navazují tzv. low-level funkce, které něco do aktviního grafického okna přidají, samy o sobě neotevřou nové př. low-level funkcí - např. abline, points, lines, legend, title, axis ... které přidají přímku, body, legendu... tzn. před použitím "low-level" funkce je potřeba, volat "high-level" funkci (např. plot, boxplot, hist, barplot, pie,...) Další grafické parametry naleznete v nápovědě nebo např. zde http://www.statmethods.net/advgraphs/parameters.html nebo zde https://flowingdata.com/2015/03/17/r-cheat-sheet-for-graphical-parameters/ nebo http://bcb.dfci.harvard.edu/~aedin/courses/BiocDec2011/2.Plotting.pdf Barvy v R http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/colorPaletteCheatsheet.pdf Ukládání grafů lze např. pomocí funkce dev.print, jpeg, pdf a dalších. Jednodušeji pak v okně Plots -> Export

L124: Tabulka absolutních četností kategoriální proměnné výrobce...

L125: výpis - objekt typu "table" - většinou vhodnější, ale těžší převedení do typu data.frame

L126: ...a pomocí funkcí z dplyr (složitější)

L127: počet výrobků pro každého výrobce

L128: výpis - objekt typu "tibble" - hodí se, když potřebujeme jednoduše převést na typ data.frame

L129: ### Tabulka relativních četností

L130: Přímým výpočtem

L131: výpis

L132: nebo pomocí funkce prop.table

L133: výpis

L134: nebo pomocí funkcí dplyr, kde budou zahrnuty i absolutní četnosti

L135: výpis

L136: U všech tabulek je potřeba pohlídat zaokrouhlení a s ním spojené riziko zaokrouhlovací chyby.

L137: Postup pro rel.cetnosti a rel.cetnosti2 je stejný.

L138: zaokrouhlení na 1 desetinné místo

L139: ohlídání zaokrouhlovací chyby

L140: Postup pro tabulka\_abs\_rel je jiný, a to kvůli jinému formátu (tibble)

L141: #### Vytvoření tabulky s absolutními i rel. četnostmi (bez dplyr). Máme:

L142: sloučení tabulek

L143: změna názvů sloupců

L144: #### Uložení tabulky do csv souboru

L145: Kde je tabulka uložena? Bez uvedení kompletní cesty v předchozím příkazu je uložena v pracovním adresáři.

L146: <hr>

L147: ### Vizualizace pomocí grafů

L148: Sloupcový graf

L149: Základní (tzn. nevyžadující žádný balíček) sloupcový graf vychází z tabulky četností, kterou máme nachystanou

L150: jednoduché rozdělení grafického okna - 1 řádek, 1 sloupec

L151: okraje kolem každého z grafů v počtech řádků - - c(dole, vlevo, nahoře, vpravo)

L152: vnější okraje v počtech řádků - c(dole, vlevo, nahoře, vpravo)

L153: Změna barev, přidání názvu

L154: alt. může být volen vektor konkrétních barev, např. c("blue","yellow,"red","green")

L155: nebo jiné škály (heat.colors, topo.colors, terrain.colors a mnoho dalších)

L156: parametr space vytvoří mezeru mezi sloupci

L157: Přidání dalších popisků a legendy

L158: horizontální orientace grafu

L159: nevykresluje čáru kolem sloupečků

L160: Funkce paste0 umožňuje sloučit textové řetězce a hodnoty proměnných, symbol "\n" tvoří nový řádek v textu

L161: umístění legendy u sloupcového grafu je velmi ošemetné

L162: mnohem snadněji se v tomto případě pracuje s ggplot2

L163: Přidání absolutních a relativních četností k odpovídajícím sloupcům

L164: parametr pos udává, kde bude text uveden vzhledem k dané pozici (1 = pod, 2 = vlevo, 3 = nad, 4 = vpravo)

L165: Zkuste využít předešlého kódu a vytvořit si sloupcový graf pro proměnnou Výrobce podle sebe.

L166: # 7. Explorační analýza a vizualizace kvantitativní proměnné

L167: Popisná statistika

L168: Výpočet průměru jedné proměnné

L169: Pozor na chybějící hodnoty

L170: Výpočet mediánu jedné proměnné

L171: Určení rozsahu

L172: #### Další charakteristiky -> var(), sd(), min(), max(),...

L173: Pozor! Funkce pro výpočet šikmosti (skewness) a špičatosti (kurtosis) nejsou součástí základního R, najdete je v balíčku moments Normálnímu rozdělení odpovídá špičatost 3, resp. špčatost v intervalu (1,5) Pro standardizaci špičatosti je nutno od vypočtené hodnoty odečíst 3. Napíšete-li před název funkce název balíčku a "::", zajistíte tím, že bude použita funkce z daného balíčku Nutno ohlídat, když jsou v různých balíčcích definovány různé funkce pod stejným jménem

L174: install.packages("moments")

L175: Chceme-li spočítat danou charakteristiku pro proměnnou kapacita po 5 cyklech

L176: podle výrobců, můžeme použít funkci tapply

L177: nebo pomocí dplyr - zde pozor na automatické (ne vždy správné zaokrouhlení)

L178: Pro zjednodušení práce můžeme využít funkce dplyr a všechny charakteristiky si nasázet do jedné tabulky

L179: bez použití group\_by pro celou proměnnou kap5

L180: preventivní na.rm=T

L181: variační koeficient v procentech

L182: preventivní specifikace balíčku moments

L183: Nezapoměňte na správné zaokrouhlení!

L184: Použijeme group\_by a dostaneme charakteristiky pro kapacitu po 5 cyklech podle výrobců

L185: Vzhledem k neúplnému výpisu je vhodné si výstup uložit a prohlédnout si jej v novém okně

L186: variační koeficient v procentech

L187: ### Krabicový graf

L188: \*\*Vykreslujeme pro originální data, můžeme doplnit i vykreslení pro data bez OP.\*\*

L189: Jednoduché a rychlé vykreslení pomocí základní funkce pouze pro výrobce C

L190: Další úprava grafu, využití funkce points pro zobrazení průměru

L191: do stávajícího grafu doplní bod znázorňující průměr

L192: Horizontální orientace, změna šířky krabice

L193: při horizontální orientaci je třeba si ohlídat opačné nastavení popisků

L194: změní šířku krabice na 1/2

L195: Využijte předešlého kódu a vytvořte si krabicový graf podle sebe.

L196: A ještě vykreslení vícenásobného krabicového grafu

L197: grafické parametry lze nastavit obdobně jako u předchozích

L198: ### Histogram

L199: \*\*Vykreslujeme vždy pro data bez odlehlých pozorování!!\*\*

L200: Jednoduché a rychlé vykreslení

L201: Co dělají různé hodnoty parametru breaks s grafem?

L202: Již tradičně lze nastavit popisky, barvy a další parametry

L203: barva výplně

L204: barva ohraničení sloupců

L205: přidá absolutní četnosti daných kategorií ve formě popisků

L206: Změna měřítka osy y, kvůli vykreslení odhadu hustoty pravděpodobnosti

L207: změna měřítka na ose y --> f(x)

L208: připojí graf odhadu hustoty pravděpodobnosti

L209: Generování hustoty normálního rozdělení a přidání k histogramu

L210: generování hodnot pro osu x

L211: generování hodnot pro osu y

L212: do posledního grafu přidání křivky na základě výše vygenerovaných hodnot

L213: Takto kombinovaný graf může posloužit k vizuálnímu posouzení normality.

L214: \*\*Využijte předešlého kódu a vytvořte si histogram podle sebe.\*\*

L215: ### QQ-graf

L216: \*\*Vykreslujeme vždy pro data bez odlehlých pozorování!!\*\*

L217: Jednoduché a velmi rychlé vykreslení...

L218: ... s úpravou popisků os...

L219: Pro pokročilé a zájemce - automatizace, využití for-cyklu, více grafů do jednoho obrázku Využíváme-li základní funkce (barplot, boxplot, histogram), pak se využívá funkce par() nebo layout() V těchto funkcích specifikujeme strukturu - jak chceme více obrázků vykreslit

L220: Např. chceme vykreslit histogram i boxplot pro kapacitu po 5 cyklech akumulátorů od výrobce A

L221: vytvoření struktury

L222: nastavení velikosti okrajů

L223: Pomocí for-cyklu histogramy a boxploty pro všechny výrobce

L224: Kombinace histogramu a QQ-plotu

L225: # 8. Vnitřní hradby a identifikace odlehlých pozorování

L226: ## Ruční odstranění pomocí napočítání vnitřních hradeb

L227: separace datového sloupce pro výrobce A

L228: mezikvartilové rozpěti

L229: výpočet dolní mezi vnitřních hradeb

L230: výpočet horní mezi vnitřních hradeb

L231: nastavíme hodnoty které jsou mimo meze na NA

L232: můžeme hondoty NA smazat

L233: ## Automatické odstranění dle box-plotu

L234: ### Jak to udělat pro data ve standardním formátu s více kategoriemi?

L235: \*\*Pozor, je to třeba udělat po jednotlivých kategoriích, jinak riskujeme smazání dat, které do dat. souboru patří!!!\*\*

L236: c("A","B",..)

L237: \*\*Analytik může vždy říct, že odlehlá pozorování odstraňovat nebude, ale tuto informaci musí do zápisu o analýze uvést!\*\*

L238: # 9. pravidlo 3 $\sigma$ a Čebyševova nerovnost

L239: ## Empirické ověření normality

L240: \*\*Vycházíme z dat po odstranění odlehlých pozorování:\*\*

L241: použijeme data z ukázky odstranění op

L242: Vykreslíme QQ graf a spočteme šikmost a špičatost:

L243: jiná definice posunutá o 3

L244: - tečky v QQ grafu musí ležet přibližně na čáře - tzn. kvantily odpovídají přibližně kvantilům normálního rozdělení - šikmost (skewness) musí ležet v intervalu <-2, 2> - špočatost (kurtosis) musí ležet v intervalu <-2,2>

L245: - pozor výsledek Rkové funkce musíme ponížit o 3

L246: \*\*Je-li splněna normalita dat -> pravidlo 3σ\*\*<br> σ: P(µ − σ < X < µ + σ) = 0,6827<br> 2σ: P(µ − 2σ < X < µ + 2σ) = 0,9545<br> 3σ: P(µ − 3σ < X < µ + 3σ) = 0,9973<br> <br> \*\*Není-li splněna normalita dat -> Čebyševova nerovnost\*\*<br> σ: P(µ − σ < X < µ + σ) = 0<br> 2σ: P(µ − 2σ < X < µ + 2σ) = 0,75<br> 3σ: P(µ − 3σ < X < µ + 3σ) = 0,8889<br>

L247: # 10. Zaokrouhlování

L248: vše potřebné k zaokrouhlování naleznete na LMS v dokumento zaokrouhlování. https://lms.vsb.cz/pluginfile.php/1298954/mod\_folder/content/0/Leg%C3%A1ln%C3%AD%20tah%C3%A1ky/zaokrouhlovani.pdf <br> To nejdůležitější: - směrodatnou odchylku zaokrouhlujeme na předepsaný počet cifer nahoru (ceiling)

L249: - velikost datového souboru = <2,10> -> 1 platná cifra

L250: - velikost datového souboru = (10,30> -> 2 platné cifry

L251: - velikost datového souboru = (30,2000> -> 3 platné cifry

L252: - míry polohy (průměry, kvantily, ...) pak zaokrouhlujeme klasicky (round) na stejnou platnou cifru jako směrodatnou odchylku

%%%%%%%%%% cv2

L0: # Cvičení 2 - Pravděpodobnost

L1: ## Adéla Vrtková, Michal Béreš, Martina Litschmannová

L2: V tomto cvičení projdeme úvod do pravděpodobnosti. Předpokládáme znalosti z přednášky, především pojmy: \*\*definice pravděpodobnosti, podmíněná pravděpodobnost, věta o úplné pravděpodobnosti, Bayesova věta\*\*.

L3: # Pomocné funkce

L4: ## Úplná pravděpodobnost

L5: $P(A)=\sum\_{i=1}^{n}P(B\_i)P(A|B\_i)$

L6: spočítání pravděpodobnosti P(A) - věta o úplné pravděpodobnosti

L7: uvažujeme P\_B jako vektor hodnot P(B\_i) a P\_BA jako vektor hodnot P(A|B\_i)

L8: ## Bayesova věta

L9: $P(B\_k|A)=\frac{P(B\_k)P(A|B\_k)}{\sum\_{i=1}^{n}P(B\_i)P(A|B\_i)}$

L10: spočítání podmíněné pravděpodobnosti P(B\_k|A) - Bayesova věta

L11: uvažujeme P\_B jako vektor hodnot P(B\_i), P\_BA jako vektor hodnot P(A|B\_i)

L12: \*\*Přidáme funkce z munulého cvičení pro počítání kombinatorických výběrů, jsou v skriptu kombinatorika.R\*\*

L13: # Příklady

L14: ## Příklad 1.

L15: Určete pravděpodobnost, že při hodu 20stěnnou spravedlivou (férovou) kostkou padne číslo větší než 14.

L16: pravděpodobnost jako podíl příznivých ku všem

L17: ## Příklad 2.

L18: Určete pravděpodobnost, že při hodu 20stěnnou kostkou padne číslo větší než 14, víte-li, že sudá čísla padají 2x častěji než lichá.

L19: pravdepodobnost je

L20: ## Příklad 3.

L21: Určete pravděpodobnost, že ve sportce uhodnete 4 čísla. (Ve sportce se losuje 6 čísel ze 49.)

L22: ## Příklad 4.

L23: Z abecedního seznamu studentů zapsaných na dané cvičení vybere učitel prvních 12 a nabídne jim sázku: „Pokud se každý z Vás narodil v jiném znamení zvěrokruhu, dám každému z Vás 100 Kč. Pokud jsou však mezi Vámi alespoň dva studenti, kteří se narodili ve stejném znamení, dá mi každý z Vás 100 Kč.“ Vyplatí se studentům sázku přijmout? S jakou pravděpodobností studenti vyhrají?

L24: ## Příklad 5.

L25: Spočtěte pravděpodobnost toho, že z bodu 1 do bodu 2 bude protékat elektrický proud, je-li část el. obvodu včetně pravděpodobnosti poruch jednotlivých součástek vyznačen na následujícím obrázku. (Poruchy jednotlivých součástek jsou na sobě nezávislé.) ![image.png](attachment:image.png)

L26: rozdělíme na bloky I=(A,B) a II=(C,D,E)

L27: výsledek

L28: ## Příklad 6.

L29: Ohrada má obdélníkový tvar, východní a západní stěna mají délku 40 m, jižní a severní pak 100 m. V této ohradě běhá kůň. Jaká je pravděpodobnost, že je k jižní stěně blíž než ke zbývajícím třem?

L30: geometrická pravděpodobnost

L31: blize k jihu

L32: pravdepodobnosti

L33: ## Příklad 7.

L34: U pacienta je podezření na jednu ze čtyř vzájemně se vylučujících nemocí - N1, N2, N3, N4 s pravděpodobností výskytu P(N1)=0,1; P(N2)=0,2; P(N3)=0,4; P(N4)=0,3. Laboratorní zkouška A je pozitivní v případě první nemoci v 50 % případů, u druhé nemoci v 75 % případů, u třetí nemoci v 15 % případů a u čtvrté v 20 % případů. Jaká je pravděpodobnost, že výsledek laboratorní zkoušky bude pozitivní?

L35: věta o úplné pravděpodobnosti

L36: P(N1), P(N2), ...

L37: P(P|N1), P(P|N2), ...

L38: P(P)

L39: ## Příklad 8.

L40: Telegrafické znaky se skládají ze signálů „tečka“, „čárka“. Je statisticky zjištěno, že se zkomolí 25 % sdělení „tečka“ a 20 % signálů „čárka“. Dále je známo, že signály se používají v poměru 3:2. Určete pravděpodobnost, že byl přijat správně signál, jestliže byl přijat signál „tečka“.

L41: Bayesova věta

L42: P(O.), P(O-)

L43: P(P.|O.), P(P.|O-)

L44: k = 1 protože správně = O.

L45: ## Příklad 9.

L46: V jednom městě jezdí 85 % zelených taxíků a 15 % modrých. Svědek dopravní nehody vypověděl, že nehodu zavinil řidič modrého taxíku, který pak ujel. Testy provedené za obdobných světelných podmínek ukázaly, že svědek dobře identifikuje barvu taxíku v 80 % případů a ve 20 % případů se mýlí.

L47: - Jaká je pravděpodobnost, že viník nehody skutečně řídil modrý taxík?

L48: - Následně byl nalezen další nezávislý svědek, který rovněž tvrdí, že taxík byl modrý. Jaká je nyní pravděpodobnost, že viník nehody skutečně řídil modrý taxík?

L49: - Ovlivní pravděpodobnost, že viník nehody skutečně řídil modrý taxík to, zda dva výše zmínění svědci vypovídali postupně nebo najednou?

L50: a) opět Bayesova věta

L51: P(Z), P(M)

L52: P(SM|Z), P(SM|M)

L53: modrý je druhý

L54: b) první možnost - druhý průchod Bayesem

L55: P(Z), P(M)

L56: P(S2M|Z), P(S2M|M)

L57: c) nebo odpověděli najednou

L58: P(Z), P(M)

L59: P(S1M&S2M|Z), P(S1M&S2M|M)

L60: ## Příklad 10.

L61: Potřebujeme zjistit odpověď na určitou citlivou otázku. Jak odhadnout, kolik procent dotazovaných na otázku odpoví ANO a přitom všem respondentům zaručit naprostou anonymitu? Jedním z řešení je tzv. dvojitě anonymní anketa:<br> Necháme respondenty hodit korunou a dvojkorunou a ti, kterým padl na koruně líc napíšou na lísteček odpověď (ANO/NE) na citlivou otázku. Ostatní respondenti napíší, zda jim padl na dvojkoruně líc (ANO/NE). Jakým způsobem určíme podíl studentů, kteří na citlivou otázku odpověděli ANO?<br> Předpokládejme, že respondenti byli dotazování, zda podváděli u zkoušky. Z anketních lístků se zjistilo, že „ANO“ odpovědělo 120 respondentů a „NE“ odpovědělo 200 respondentů. Kolik procent studentů podvádělo u zkoušky?

L62: věta o úplné pravděpodobnosti

L63: P(A) = P(K\_lic)\*P(A|K\_lic)+P(K\_rub)\*P(D\_lic|K\_rub)

L64: rovnice 120/320=0.5\*x+0.5\*0.5

L65: ## Bonus - Monty Hall Problem

L66: Začneme s vygenerováním n instancí soutěže - cena bude náhodný index dveří (1,2,3) za kterými se může nacházet cena

L67: počet pokusů

L68: náhdný výběr dveří

L69: head vykresli prvních 6 prvků/řádků

L70: Totéž pro naši původní volbu - náhodný index dveří.

L71: původní volba

L72: V prvním kole moderátor jedny prázdné dveře otevře, takto se to dá nasimulovat:

L73: inicializace vektoru

L74: pomocná proměnná - identifikátory dveří

L75: inicializace

L76: nesmíme otevřít dveře s cenou

L77: ani naše vybrané dveře

L78: ve zbytku jsou buď 2 (pokud jsme se trefili) nebo 1 dveře (pokud ne)

L79: pokud jedny otevřeme je

L80: pokud 2 tak jedny náhodně vybereme a otevřeme je

L81: Naše nová volba pokud se tak rozhodneme - součet indexů je 1+2+3=6 takže pokud my máme vybraný nějaký index, dále nějaký index se otevře, tak do zbytku 6 jsou ty třetí = naše nová volba.

L82: Úspěšnost při originální volbě:

L83: Úspěšnost při výměně:

%%%%%%%%%% cv12

L0: # Cvičení 12. Vícevýběrové testy ## Michal Béreš, Martina Litschmannová, Adéla Vrtková

L1: ## Testovací data pro ukázku volání funkcí

L2: vyrobím data framy o jednom sloupci z náh. dat

L3: přejmenuju název sloupce

L4: doplním typ pro všechny data framy

L5: slepím po řádcích dohromady

L6: převedu typ na typ factor

L7: pokud jsou k dispozici nějaká OP, budu je igrorovat

L8: (vím, že data jsou z normálního rozdělení!)

L9: (vím i že mají stejný rozptyl)

L10: # Přehled funkcí

L11: ## Míry variability

L12: ### Bartlettův test

L13: - ověřuje shodu rozptylů - $H\_0: \sigma^2\_1 = \sigma^2\_2 = \sigma^2\_3 = \ldots$ - $H\_A: \neg H\_0$ - předpokladem je normalita dat (a samozřejmě nezávislost a spojitost)

L14: ### Leveneův test

L15: - ověřuje shodu rozptylů - $H\_0: \sigma^2\_1 = \sigma^2\_2 = \sigma^2\_3 = \ldots$ - $H\_A: \neg H\_0$ - předpokladem je pouze nezávislost a spojitost

L16: ### Cochranův a Hartleyův test

L17: - také ověřují shodu rozptylů - požadují normalitu dat a tzv. vyvážené třízení - vyvážené třízení znamená, že máme přibližně stejné množství dat v každé skupině - nebudeme je používat

L18: ## Míry polohy

L19: ### ANOVA (analýza rozptylu)

L20: - ověřuje shodu polohy (středních hodnot) - $H\_0: \mu\_1 = \mu\_2 = \mu\_3 = \ldots$ - $H\_A: \neg H\_0$ - předpoklady: - normalita dat - homoskedasticita (shodné rozptyly) - (a samozřejmě nezávislost a spojitost) - pokud zamítáme $H\_0$ je vyžadována Post-Hoc analýza - pomocí TukeyHSD testu

L21: Základní ANOVA

L22: H0: mu1 = mu2 = mu3 = mu4

L23: HA: ~H0 (negace H0)

L24: Post-Hoc analýza

L25: počítání efektů

L26: celkový průměr

L27: průměry ve skupinách

L28: efekty

L29: vypsat setřízené

L30: ### Kruskalův - Wallisův test

L31: - ověřuje shodu polohy (mediánů) - $H\_0: X\_{0.5,1} = X\_{0.5,2} = X\_{0.5,3} = \ldots$ - $H\_A: \neg H\_0$ - předpoklady: - symetrie dat - (a samozřejmě nezávislost a spojitost) - pokud zamítáme $H\_0$ je vyžadována Post-Hoc analýza - pomocí Dunnové testu/metody

L32: Základní KW test

L33: H0: X0.5,1 = X0.5,2 = X0.5,3 = X0.5,4

L34: HA: ~H0 (negace H0)

L35: Post-Hoc analýza

L36: altP = T nastavuje p-hodnotu tak,aby se při rozhodování

L37: o statistické významnosti srovnávala s alfa

L38: (defaultně: altp = FALSE, pak srovnáváme s alfa/2)

L39: install.packages("dunn.test")

L40: počítání efektů

L41: celkový median

L42: mediany ve skupinách

L43: efekty

L44: vypsat setřízené

L45: # Příklady

L46: ## Příklad 1.

L47: Testujeme nulovou hypotézu µ1 = µ2 = µ3. Bylo zjištěno, že data, která máme k dispozici jsou výběry z normálního rozdělení splňující předpoklad homoskedasticity (shody rozptylů). Na základě údajů získaných explorační analýzou doplňte tabulku ANOVA a vyplývající závěry.<br> ![image.png](attachment:4429db71-6c58-4bf9-b61f-fd9ec4015d20.png)

L48: rozsahy výběrů

L49: průměry v jednotlivých skupinách / třídách

L50: směr. odchylky v jednotlivých skup. / třídách

L51: celkový rozsah výběrů

L52: počet tříd

L53: počet stupňů volnosti - meziskupinový

L54: počet stupňů volnosti - reziduální

L55: celkový průměr (pomocí váženého průměru)

L56: meziskupinový součet čtverců

L57: reziduální součet čtverců

L58: celkový součet čtverců

L59: rozptyl mezi skupinami / třídami

L60: rozptyl uvnitř skupin / tříd

L61: F-poměr

L62: p-hodnota

L63: Na hladině významnosti 0,05 zamítáme hypotézu o shodě středních hodnot

L64: tj. střední hodnoty alespoň jedné dvojice skupin se stat. významně liší.

L65: odhady skupinových efektů

L66: Oproti celkovému průměru vykazuje nejvíce podprůměrné výsledky skupina 2

L67: (o cca 10 jednotek nižší než celkový průměr). Naopak průměr skupiny 3 je

L68: o cca 10 jednotek vyšší než celkový průměr. Průměrné výsledky skupiny 1

L69: odpovídají celkovému průměru.

L70: ## Příklad 2.

L71: 122 pacientů, kteří podstoupili operaci srdce, bylo náhodně rozděleno do tří skupin.<br> \*\*Skupina 1:\*\* Pacienti dostali 50 % oxidu dusného a 50 % kyslíkové směsi nepřetržitě po dobu 24 hodin.<br> \*\*Skupina 2:\*\* Pacienti dostali 50 % oxidu dusného a 50 % kyslíkové směsi pouze během operace.<br> \*\*Skupina 3:\*\* Pacienti nedostali žádný oxid dusný, ale dostali 35-50 % kyslíku po dobu 24 hodin.<br> Data v souboru kyselina listova.xls odpovídají koncentracím soli kyseliny listové v červených krvinkách ve všech třech skupinách po uplynutí 24 hodin ventilace. Ověřte, zda pozorované rozdíly mezi koncentracemi soli kyseliny listové jsou statisticky významné, tj. zda existuje vliv složení směsi na sledovaný parametr.

L72: přejmenování sloupců

L73: převod do standardního datového formátu

L74: Data neobsahují odlehlá pozorování.

L75: otestujeme normalitu pomocí S.-W. testu

L76: Informace potřebné pro nastavení zaokrouhlování

L77: sd zaokrouhlujeme na 3 platné cifry

L78: sd a míry polohy zaokrouhlujeme na desetiny

L79: Ověření shody rozptylů

L80: výběrové rozptyly

L81: Dle krabicového grafu a informace o poměru největšího a nejmenšího

L82: rozptylů (<2) nepředpokládáme, že se rozptyly statisticky významně liší

L83: Předpoklad normality nebyl zamítnut -> Bartlettův test

L84: Na hladině významnosti 0,05 nelze zamítnout předpoklad o shodě rozptylů

L85: (Bartlettův test, x\_OBS = 0,878, df = 2, p-hodnota = 0,645).

L86: Chceme srovnávat stř. hodnoty nezávislých výběrů z normálních rozdělení

L87: se stejnými rozptyly -> ANOVA

L88: příkaz aov() vyžaduje data ve standardním datovém formátu

L89: Na hladině významnosti 0,05 zamítáme hypotézu o shodě středních hodnot

L90: (ANOVA, p-hodnota<<0,001) -> mnohonásobné porovnávání

L91: post-hoc analýza

L92: počítání efektů

L93: celkový průměr

L94: průměry ve skupinách

L95: efekty

L96: vypsat setřízené

L97: Považujeme-li vysoký obsah kyseliny listové za pozitivní, pak statisticky

L98: významně nejlepších výsledků dosáhli pacienti ze skupiny 1 (průměrný obsah

L99: kys. listové o cca 27 jednotek vyšší než prům. obsah kys. listové v krvi

L100: všech testovaných pacientů) a statisticky významně nejhorších výsledků

L101: dosáhli pacienti ze skup. 2 (průměrný obsah kys. listové o cca 26 jednotek

L102: nižší než průměrný obsah kys. listové v krvi všech testovaných pacientů).

L103: Obsah kys. listové v krvi pacientů ze skupiny 3

L104: odpovídá celkovému průměru. Všechny tři skupiny pacientů jsou navzájem dle

L105: obsahu kys. listové v krvi statisticky významně odlišné.

L106: ## Příklad 3.

L107: Na farmě jsou chována tři plemena králíků. Byl proveden pokus kralici.xls, jehož cílem bylo zjistit, zda i když chováme a vykrmujeme všechny králíky po stejnou dobu a za stejných podmínek, existuje statisticky významný (průkazný) rozdíl mezi plemeny v hmotnostech králíků. Ověřte.

L108: přejmenování sloupců

L109: převod do standardního datového formátu

L110: data obsahojí OP

L111: Odstranění odlehlého pozorování

L112: Krabicový graf

L113: Na hladině významnosti 0,05 nezamítáme předpoklad normality.

L114: Informace potřebné pro nastavení zaokrouhlování

L115: sd zaokrouhlujeme na 2 platné cifry

L116: sd a míry polohy zaok. na setiny (sjednocení napříč druhy králíků)

L117: Ověření shody rozptylů

L118: Dle krabicového grafu a informace o poměru největšího a nejmenšího rozpt.

L119: (blízký 2, avšak rozsah výběrů < 30) je těžší odhadnout, zda lze

L120: předpokládat shodu rozptylů. Rozhodnout nám pomůže test.

L121: Předpoklad normality nebyl zamítnut -> Bartlettův test

L122: Na hladině významnosti 0,05 nelze zamítnout předpoklad o shodě rozptylů

L123: (Bartlettův test, x\_OBS = 3,1, df = 2, p-hodnota = 0,217).

L124: Chceme srovnávat stř. hodnoty nezávislých výběrů z normálních

L125: rozdělení se stejnými rozptyly -> ANOVA

L126: příkaz aov() vyžaduje data ve standardním datovém formátu

L127: Na hladině významnosti 0,05 zamítáme hypotézu o shodě středních hodnot

L128: (p-hodnota<<0,001, ANOVA) -> mnohonásobné porovnávání

L129: post-hoc analýza

L130: počítání efektů

L131: celkový průměr

L132: průměry ve skupinách

L133: efekty

L134: vypsat setřízené

L135: ## Příklad 4.

L136: Soutěž o nejlepší jakost výrobků obeslali čtyři výrobci A, B, C, D celkem 66 výrobky. Porota sestavila pořadí (uvedeno pouze pořadí výrobku od nejlepšího k nejhoršímu), jež je uvedené v souboru jakost.xls. Na základě uvedených údajů posuďte, zda původ výrobků má vliv na jeho jakost.

L137: přejmenování sloupců

L138: data již jsou v standardním formátu

L139: Ověření normality nemá smysl provádět - z povahy jde o diskr. data-pořadí

L140: Informace potřebné pro nastavení zaokrouhlování

L141: sd zaokrouhlujeme na 2 platné cifry

L142: sd a míry polohy zaokrouhlujeme na celá čísla

L143: Ověření shody rozptylů

L144: Dle krabicového grafu a informace o poměru největšího a nejmenšího

L145: rozptylů (<2) ze předpokládat shodu rozptylů.

L146: (Kruskalův - Wallisův test má větší sílu testu, jsou-li data homosk.)

L147: Jde o "pořadová" data, nemá smysl uvažovat o předpokladu norm.

L148: -> Leveneův test

L149: Na hladině významnosti 0,05 nelze zamítnout předpoklad o shodě rozptylů

L150: (Leveneho test, x\_OBS = 0,4, df\_num = 3, df\_denom = 62, p-hodnota = 0,750)

L151: Ověření symetrie

L152: Chceme srovnávat mediány nezávislých výběrů -> Kruskalův-Wallisův test

L153: Na hladině významnosti 0,05 nelze zamítnout hypotézu o shodě mediánů

L154: (Kruskalův-Wallisův test, x\_OBS = 3,7, df = 3, p-hodnota=0,295).

L155: Tj. statisticky významné rozdíly mezi výrobci (z hlediska pořadí

L156: výrobků v soutěži) neexistují.

L157: ## Příklad 5.

L158: Byl sledován vliv tří preparátů na srážlivost krve. Kromě jiných ukazatelů byl zjišťován tzv. trombinový čas. Údaje o 42 sledovaných osobách jsou zaznamenány v souboru trombin.xls. Závisí velikost trombinového času na tom, jaký byl použit preparát?

L159: přejmenování sloupců

L160: data jsou již ve standardním formátu

L161: explorační analýza - ověření OP

L162: neobsahuje OP

L163: ověření normality

L164: Na hladině významnosti 0,05 zamítáme předpoklad normality (u Ačka)

L165: Informace potřebné pro nastavení zaokrouhlování

L166: sd zaokrouhlujeme na 2 platné cifry

L167: sd a míry polohy zaokrouhlujeme na setiny (sjednocení napříč skupinami)

L168: Ověření shody rozptylů (není nutné - stejně musíme použít KW)

L169: Dle krabicového grafu a informace o poměru největšího a nejmenšího

L170: rozptylů (>>2) nelze předpokládat shodu rozptylů.

L171: Předpoklad normality byl zamítnut -> Leveneho test

L172: trombin.s$skupina = as.factor(trombin.s$skupina)

L173: předpoklad homoskedasticity byl zamítnut

L174: Ověření symetrie

L175: nezamítáme přepoklad symetrie dat

L176: Chceme srovnávat mediány nez. výběrů, která nemají norm. rozdělení

L177: -> Kruskalův - Wallisův test

L178: Na hladině významnosti 0,05 zamítáme hypotézu o shodě mediánů

L179: Tj. trombinový čas je statisticky významně

L180: ovlivněn preparátem. -> mnohonásobné porovnávání

L181: altP = T nastavuje p-hodnotu tak,aby se při rozhodování

L182: o statistické významnosti srovnávala s alfa

L183: (defaultně: altp = FALSE, pak srovnáváme s alfa/2)

L184: počítání efektů

L185: celkový průměr

L186: průměry ve skupinách

L187: efekty

L188: vypsat setřízené

L189: ## Příklad 6. (více skupin)

L190: Co se Sněhurka dostala k sedmi trpaslíkům vycítila příležitost nemalého výdělku. Trpaslíci Sněhurce v podstatě zobou z ruky a veškeré vydolované zlato jí ihned předávají. Sněhurce však ani toto úplně nestačí a má pocit, že by mohla z trpaslíků benefitovat více. Proto si začla zaznamenávat kolik kilogramů zlata denně od každého z trpaslíků obdrží (snehurka.xlsx). Ověřte, zda se trpaslíci liší v množství vytěženého zlata, pokud ano sestave homogenní skupiny z hlediska vytěženého zlata.

L191: data jsou ve standardním dtovém formátu

L192: data neobsahují OP

L193: ověření normality

L194: Na hladině významnosti 0,05 ne zamítáme předpoklad normality

L195: Předpoklad normality nebyl zamítnut -> Bartlettův test

L196: Na hladině významnosti 0,05 nelze zamítnout předpoklad o shodě rozptylů

L197: ANOVA

L198: Zamítáme předpoklad o shodě

L199: -> existují stat. významné rozdíly ve středních hodnotách

L200: POST-HOC

L201: počítání efektů

L202: celkový průměr

L203: průměry ve skupinách

L204: efekty

L205: vypsat setřízené

%%%%%%%%%% cv13

L0: # Cvičení 13. Neparametrické testy, testy dobré shody ## Michal Béreš, Martina Litschmannová, Adéla Vrtková

L1: # Testování shody rozdělení pravděpodobnosti diskrétní NV (o konečném počtu hodnot) - test dobré shody

L2: - testujeme zda naměřená data (jejich relativní četnosti) souhlasí s nějakým konkrétním rozdělením (tedy jeho pravděpodobnostmi) - testujeme pomocí $\chi^2$ testu dobré shody - předpoklady testu: (POZOR týkají se předpokládaných četností - tedy těch které bychom sledovali pokud by naměřená data byla 100% dle rozdělení v hypotéze) - Očekávané četnosti ≥ 2, - alespoň 80% očekávaných četností > 5 - testová statistika (ta která má $\chi^2$ rozdělení) je $G = \sum\_{i = 1}^k (O\_i - E\_i)^2 / E\_i$ - rozdělení ma stupeň volnosti $df = k - 1 - h$ - k je počet možností - h je počet odhadovaných parametrů (toto se týká neúplně specifikovaných testů)

L3: ukázka viz příklad 1

L4: ## Test dobré shody pro spojitou náhodnou veličinu (případně diskrétní o nekonežném počtu hodnot)

L5: - musíme převést na tabulku s konečným počtem hodnot - u diskrétní (např. poison) shlukneme od určitého počtu kolonek např. 4,5,6,... na "4 a více" - u spojité vyrobíme sérii intervalů a koukáme se kolik hodnot spadne do daného intervalu - např.: (-$\infty$, 3), <3, 4), ..., <10, $\infty$) - pak musíme pro očekávané rozdělení napočítat pro každý interval kolik % dat do nich patří, čímž získáme tabulku očekávaných pravděpodobností - dále pokračujeme jako u předchozího - pro testování normality rozdělení existuje funkce pearson.test(data) z balíčku nortest

L6: ukázka příklady 2,3,4

L7: # Kontingenční tabulky

L8: - tabulky obsahující data v závislosti na dvou faktorech - jeden z faktorů je z pravidla nezávislá proměnná u které sledujeme, zda má vliv na druhý faktor (závislá proměnná) - nezávislá proměnná je zpravidla v řádcích - závislá je zpravidla ve sloupcích - pozor celé testování zkoumá korelaci, nikoliv kauzalitu! Kauzalita se dá posoudit "expertním" zhodnocením - statistický závěr: existuje statisticky významná závislost mezi nezávislou a závislou proměnnou (korelace) - expertní posouzení: nezávislá proměnná statisticky významně ovlivňuje závislou proměnnou (kauzalita)

L9: ## Vizualizace kontingenční tabulky

L10: - vizualizace např. pomocí funkce barplot - pozor na to co jsou řádky a sloupce, vždy chceme aby jednotlivé dělené sloupce byly přes nezávislé proměnné (každý sloupec pro jednu hodnotu nezávislé proměnné) - beside = T určuje zda chceme přilehlé sloupce sloučit do jednoho děleného sloupce, či ne - preferovaná vizualizace pomocí mozaikového grafu mosaicplot - totéž co u barplotu, spojené sloupce musí být přes nezávislé proměnné

L11: ukázky v příkladech 5,6,7

L12: ## Míry závislosti v koningenční tabulce

L13: - Koeficient korelace CC - Korigovaný koeficient korelace CCcor - Cramerův koeficient V - tento budeme používat především - funkce cramersV(kont.tab) z balíčku lsr

L14: ukázky v příkladech 5,6,7

L15: ## Test závislosti v kontingenční tabulce

L16: - $H\_0:$ neexistuje závislost mezi nezávislou (např. je kouřák) a závislou (např. trpí onemocněním) proměnnou - $H\_A: \neg H\_0$ - funkce chisq.test(kont.tab) - předpoklady: Očekávané četnosti ≥ 2, alespoň 80% očekávaných četností > 5 - očekávané četnosti se dají zjistit z chisq.test(kont.tab)\$expected

L17: # Asociační tabulky

L18: - jedná se o speciální případ kontingenční tabulky - má vždy právě 2 možnosti pro závislou a právě 2 možnosti pro nezávislou proměnnou

L19: ## Povinný tvar asociační tabulky

L20: - řádky označují možnosti nezávislé proměnné - první řádek je takzvaná exponovaná část populace (ta vystavena jevu který zkoumáme - např. kuřáci poku zkoumáme vliv kouření) - druhý řádek je neexponovaná část populace - sloupce označují možnosti závislé proměnné - první sloupec označuje výskyt zkoumaného jevu (např. výskyt nemoci, chyba výrobku, ...) - druhý sloupec označuje zbytek - bez výskytu zkoumaného jevu

L21: ## Relativní riziko a poměr šancí

L22: - Relativní riziko a poměr šancí podávají stejnou informaci, pouze v jiném formátu - všechny bodové IO se spočtou pomocí funkce epi.2by2(asoc.tab) z balíčku epiR - funkce bere jako vstup asociační tabulku, která musí být ve správném formátu!

L23: ### Relativní riziko

L24: - značíme $RR$ - jedná se o poměr rizika (pravděpodobnost výskytu zkoumaného jevu) u exponované a neexponované populace - pokud je roven 1 tak to značí stejné pravděpodobnosti výskytu u exponované i u neexponované - pokud je větší než 1 pak exponovaná polulace má větší pravděpodobnost výskytu - pokud je menší než 1 pak exponovaná populace má menší pravděpodobnost výskytu - bodový odhad $\hat{RR}$ se spočte jako poměr rel. četností zkoumaného jevu u exponované a neexponované populace - funkce epi.2by2 dodá intervalové odhady - pokud IO neobsahuje hodnotu 1 pak existuje statisticky významná závislost mezi závislou a nezávislou proměnnou

L25: ### Poměr šancí

L26: - značíme $OR$ - jedná se o poměr šancí (šance výskytu zkoumaného jevu) u exponované a neexponované populace - pokud je roven 1 tak to značí stejné šance výskytu u exponované i u neexponované - pokud je větší než 1 pak exponovaná polulace má větší šanci výskytu - pokud je menší než 1 pak exponovaná populace má menší šanci výskytu - bodový odhad $\hat{OR}$ se spočte jako poměr šancí (výběrových) zkoumaného jevu u exponované a neexponované populace - funkce epi.2by2 dodá intervalové odhady - pokud IO neobsahuje hodnotu 1 pak existuje statisticky významná závislost mezi závislou a nezávislou proměnnou

L27: ukázka v příkladu 7

L28: # Příklady (testy dobré shody)

L29: ## Příklad 1.

L30: Hodilo se 6 000 krát hrací kostkou a zaznamenaly se počty padlých ok.<br> Je možné na základě příslušného testu na hladině významnosti 0,05 spolehlivě tvrdit, že kostka je „falešná“, tj. že pravděpodobnosti všech čísel na kostce nejsou stejné?<br> ![image.png](attachment:64f1169e-6bc1-470a-8afb-b282230c2c9f.png)

L31: H0: Kostka je férová. (tedy všechny pravděpodobnosti jsou 1/6)

L32: Ha: Kostka není férová. (negace H0)

L33: nutno zkontrolovat předpoklady testu

L34: Všechny očekávané četnosti jsou větší než 5.

L35: Na hladině významnosti 0,05 nezamítáme HO (p-hodnota = 0,711,

L36: Chí-kvadrát test nezávislosti, df = 5).

L37: ## Příklad 2.

L38: Výrobní firma odhaduje počet poruch určitého zařízení během 100 hodin pomocí Poissonova rozdělení s parametrem 1,2. Zaměstnanci zaznamenali pro kontrolu skutečné počty poruch celkem ve 150 100hodinových intervalech (výsledky jsou uvedeny v tabulce). Ověřte čistým testem významnosti, zda má počet poruch daného zařízení během 100 hodin skutečně Poissonovo rozdělení s parametrem λt = 1,2.<br> ![image.png](attachment:65bc506a-e77d-4533-b942-ae5e7596ad9a.png)

L39: Úplně specifikovaný test

L40: H0: Počet poruch během 100 hodin provozu lze modelovat

L41: Poissonovým rozdělením s parametrem 1,2.

L42: Ha: Počet poruch během 100 hodin provozu nelze modelovat

L43: Poissonovým rozdělením s parametrem 1,2.

L44: nutno zkontrolovat předpoklady testu

L45: 4 z 5 očekávaných četnosti, tj. 80%, jsou větší než 5.

L46: Na hladině významnosti 0,05 nezamítáme HO (p-hodnota = 0,590,

L47: Chí-kvadrát test nezávislosti, df = 4).

L48: ## Příklad 3.

L49: Zaměstnanci zaznamenali pro kontrolu počty poruch celkem ve 150 100hodinových intervalech (výsledky jsou uvedeny v tabulce). Ověřte čistým testem významnosti, zda má počet poruch daného zařízení během 100 hodin skutečně Poissonovo rozdělení.<br> ![image.png](attachment:4da89057-87d1-4bcd-a488-3365237654f7.png)

L50: Neúplně specifikovaný test

L51: H0: Počet poruch během 100 hodin provozu lze modelovat

L52: Poissonovým rozdělením.

L53: Ha: Počet poruch během 100 hodin provozu nelze modelovat

L54: Poissonovým rozdělením.

L55: odhad parametru Poissonova rozdělení

L56: nutno zkontrolovat předpoklady testu

L57: 4 z 5 očekávaných četnosti, tj. 80%, jsou větší než 5.

L58: Na hladině významnosti 0,05 nezamítáme HO (p-hodnota = 0,491,

L59: Chí-kvadrát test nezávislosti, df = 3).

L60: ## Příklad 4.

L61: Na dálnici byly v průběhu několika minut měřeny časové odstupy (s) mezi průjezdy jednotlivých vozidel. Zjištěné hodnoty těchto odstupů jsou zaznamenány v souboru dalnice.xlsx. Ověřte, zda se jedná o data z normálního rozdělení (použijte test dobré shody).

L62: automatický test dobré shody ze spojitých dat

L63: generování hodnot pro osu x

L64: generování hodnot pro osu y

L65: do posledního grafu přidání křivky na základě výše vygenerovaných hodnot

L66: install.packages("nortest")

L67: H0: Rozestupy mezi vozidly lze modelovat normálním rozdělením.

L68: Ha: Rozestupy mezi vozidly nelze modelovat normálním rozdělením.

L69: Určení počtu stupňů volbnosti

L70: Na hladině významnosti 0,05 lze zamítnout HO (p-hodnota << 0,001,

L71: Chí-kvadrát test dobré shody, df = 12).

L72: test co už znáte

L73: # Příklady na kontingenční a asociační tabulky

L74: ## Příklad 5.

L75: Rozhodněte na základě datového souboru experimentovani-s-telem.xls (Dudová, J. - Experimentování s tělem (výsledky průzkumu), 2013. Dostupné online na http://experimentovani-stelem.vyplnto.cz), zda existuje souvislost mezi pohlavím respondentů a tím, zda mají tetování. Pro posouzení míry kontingence použijte Cramerovo V.

L76: Preprocessing

L77: Varianty kat. proměnných (faktorů) je nutné uspořádat a pojmenovat tak,

L78: jak mají být uspořádány a pojmenovány v kont. tabulce

L79: Explorační analýza

L80: sdružené relativní četnosti

L81: řádkové relativní četnosti

L82: sloupcové relativní četnosti

L83: Vizualizace ve standardním R

L84: Shlukový sloupcový graf

L85: srovnejte grafy, který z grafů je pro prezentaci daných dat vhodnější

L86: šířka grafů v Jupyteru

L87: matice grafů 1x2

L88: Skládaný sloupcový graf

L89: šířka grafů v Jupyteru

L90: matice grafů 1x2

L91: Mozaikový graf

L92: šířka grafů v Jupyteru

L93: otočení popisků osy y

L94: srovnejte, který z grafů je pro prezentaci daných dat vhodnější

L95: install.packages("lsr")

L96: Výpočet Cramerova V ####

L97: Test nezávislosti v kontingenční tabulce

L98: H0: Data jsou nezávislá -> to jestli je jedinec muž nebo žena

L99: neovlivní jeho pravděpodobnost, že budemít tetování

L100: HA: negace H0 (existuje závislost)

L101: Nutné pro ověření předpokladů

L102: Všechny očekávané četnosti jsou větší než 5.

L103: Na hladině významnosti 0,05 lze zamítnout HO (p-hodnota = 0,003,

L104: Chí-kvadrát test dobré shody, df = 1).

L105: Pozorovanou závislost lze hodnotit jako slabou (Cramerovo V = 0,121).

L106: ## Příklad 6.

L107: Pro diferencovaný přístup v personální politice potřebuje vedení podniku vědět, zda spokojenost v práci závisí na tom, jedná-li se o pražský závod či závody mimopražské. Výsledky šetření jsou v následující tabulce. Zobrazte data pomocí mozaikového grafu a na základě testu nezávislosti v kombinační tabulce rozhodněte o závislosti spokojenosti v zaměstnání na umístnění podniku. Pro posouzení míry kontingence použijte Cramerovo V.<br> ![image.png](attachment:ebc6062c-1020-4fab-8855-6665d65d59a7.png)

L108: Nemáme k dispozici datovou matici, tj. kont. tabulku musíme zadat "ručně"

L109: Explorační analýza ####

L110: sdružené relativní četnosti

L111: řádkové relativní četnosti

L112: sloupcové relativní četnosti

L113: Vizualizace ve standardním R

L114: Mozaikový graf

L115: otočení popisků osy y o 90

L116: Cramerovo V

L117: H0: Mezi spokojenosti v práci a umístěním podniku neexistuje souvislost.

L118: Ha: Mezi spokojenosti v práci a umístěním podniku existuje souvislost.

L119: Chí-kvadrát test nezávislosti v kontingenční tabulce ####

L120: Všechny očekávané četnosti jsou větší než 5.

L121: Na hladině významnosti 0,05 lze zamítnout HO (p-hodnota << 0,001,

L122: Chí-kvadrát test dobré shody, df = 3).

L123: Pozorovanou závislost lze hodnotit jako středně silnou (Cramerovo V = 0,296)

L124: ## Příklad 7. (Asociační tabulka)

L125: V letech 1965 až 1968 bylo v kohortové studii kardiovaskulárních onemocnění v rámci „Honolulu Heart Program“ zahájeno sledování 8 006 mužů, z nichž 7 872 nemělo při zahájení studie v anamnéze mrtvici (apoplexii). Z tohoto počtu bylo 3 435 kuřáků a 4 437 nekuřáků. Při jejich sledování po dobu 12 let dostalo mrtvici 171 mužů ve skupině kuřáků a 117 mužů ve skupině nekuřáků.

L126: #### a)

L127: Zapište zjištěné výsledky do asociační tabulky.

L128: doplnění tabulky absolutních četností

L129: doplnění tabulky relativních četností

L130: #### b)

L131: Na základě vizuálního posouzení odhadněte vliv kouření na výskyt kardiovaskulárních onemocnění.

L132: Vizualizace mozaikovým grafem v základním R

L133: Výpočet Cramerova V ####

L134: Dle mozaikového grafu a Cramerova V (0,061) lze souvislost mezi kuřáctvím

L135: a výskytem apoplexie hodnotit jako velmi slabou.

L136: #### c)

L137: Určete absolutní riziko vzniku kardiovaskulárních onemocnění u kuřáků a nekuřáků.

L138: riziko = pravděpodobnost

L139: Kuřáci

L140: Kontrola předpokladů

L141: OK (3 435 > 190,3)

L142: Výpočet bodového a 95% Clopperova-Pearsonova intervalového odhadu

L143: U kuřáku je riziko vzniku apoplexie cca 5,0 %. 95% Clopperův-Pearsonův

L144: intervalový odhad tohoto rizika je 4,2 % až 5,8 %.

L145: Nekuřáci

L146: Kontrola předpokladů

L147: OK (4 437 > 350,6)

L148: Výpočet bodového a 95% Clopperova-Pearsonova intervalového odhadu

L149: U nekuřáku je riziko vzniku apoplexie cca 2,6 %. 95% Clopperův-Pearsonův

L150: intervalový odhad tohoto rizika je 2,1 % až 3,2 %.

L151: #### d)

L152: Určete relativní riziko (včetně 95% intervalového odhadu) vzniku kardiovaskulárních onemocnění u kuřáků a nekuřáků. Vysvětlete praktický význam zjištěných výsledků.

L153: install.packages("epiR")

L154: U kuřáků je cca 1,89 krát vyšší riziko apoplexie než u nekuřáků. 95%

L155: intervalový odhad tohoto relativního rizika je 1,50 až 2,38.

L156: Dle intervalového odhadu relativního rizika je zřejmé, že na hladině

L157: významnosti 0,05 je u kuřáků statisticky významně vyšší riziko vzniku

L158: apoplexie než u nekuřáků.

L159: #### e)

L160: Určete absolutní šance vzniku kardiovaskulárních onemocnění u kuřáků a nekuřáků.

L161: U kuřáku je šance vzniku apoplexie cca 52:1 000. Tj. u 1 052 kuřáků

L162: lze očekávat cca 52 výskytů apoplexie.

L163: U nekuřáku je šance vzniku apoplexie cca 27:1 000. Tj. u 1 027 nekuřáků

L164: lze očekávat cca 27 výskytů apoplexie.

L165: #### f) Určete relativní šance vzniku kardiovaskulárních onemocnění u kuřáků.

L166: U kuřáků je cca 1,93 (= 0,0524/0,0271) krát vyšší šance apoplexie

L167: než u nekuřáků. 95% intervalový odhad tohoto poměru

L168: šancí je 1,52 až 2,46.

L169: Dle intervalového odhadu poměru šancí je zřejmé, že na hladině

L170: významnosti 0,05 je u kuřáků

L171: statisticky významně vyšší šance na vznik apoplexie než u nekuřáků.

L172: #### g)

L173: Rozhodněte na hladině významnosti 0,05 o závislosti výskytu kardiovaskulárních chorob na kouření.

L174: Pozor! Příkaz epi.2by2 nemá jako výstup očekávané četnosti pro

L175: Chí-kvadrát test nezávislosti.

L176: Není tak možno ověřit předpoklady testu!

L177: H0: Mezi kouřením a výskytem apoplexie neexistuje souvislost.

L178: Ha: Mezi kouřením a výskytem apoplexie existuje souvislost.

L179: Všechny očekávané četnosti jsou větší než 5.

L180: Na hladině významnosti 0,05 lze zamítnout HO (p-hodnota << 0,001,

L181: Chí-kvadrát test dobré shody,

L182: df = 1). Pozorovanou závislost lze hodnotit jako velmi slabou

L183: (Cramerovo V = 0,061).

%%%%%%%%%% cv1\_add

L0: # Cvičení 1 - Kombinatorika

L1: ## Adéla Vrtková, Michal Béreš, Martina Litschmannová

L2: ## Variace

L3: V(n,k) - variace bez opakování, první argument bude celkový počet entit, druhý argument velikost výběru

L4: funkce se vytváří příkazem fucntion, je to objekt jehož jméno je dáno až proměnnou

L5: do které tento objekt přiřadím

L6: zde zadávám počet parametrů a jejich jména

L7: celé tělo funkce je uzavřeno mezi závorkami {...}

L8: faktoriál v originálním Rku existuje tak jej použijeme

L9: to co funkce vrátí se dává do příkazu return(...)

L10: V\*(n,k) - variace s opakováním

L11: ## Permutace

L12: P(n)=V(n,n) - permutace

L13: P\*(n1,n2,n3,....,nk) - permutace s opakováním, vstup bude vektor s jednotlivými počty unikátních entit

L14: vec\_n je vektro počtů hodnot př.: vec\_n = c(2,2,2,4,3)

L15: spočteme kolik máme hodnot celkem

L16: jejich faktoriál = hodnota v čitateli

L17: jednoduchý cyklus začíná příkazem for, pak v závorkách následuje název iterátoru a z

L18: jakého seznamu bude brán

L19: pocet je iterátor a postupně bude nabývat hodnot z vektoru vec\_n

L20: postupně dělíme faktoriálem každého počtu unikátních entit

L21: ## Kombinace

L22: C(n,k) - kombinace

L23: funkce for kombinace už existuje v Rku a jmenuje se choose

L24: C\*(n,k) - kombinace s opakováním

L25: použijeme známý vzorec

L26: # Úlohy na cvičení

L27: ## Příklad 1.

L28: V prodejně mají k dispozici tři typy zámků. Pro otevření prvního zámku je nutno zmáčknout čtyři z deseti tlačítek označených číslicemi 0 až 9. (Na pořadí nezáleží - tlačítka zůstávají zmáčknuta.) Druhý zámek se otevře pokud zmáčkneme šest tlačítek z deseti. Pro otevření třetího zámku je nutno nastavit správnou kombinaci na čtyřech kotoučích. Který z těchto zámků nejlépé chrání před zloději?

L29: ## Příklad 2.

L30: V prodejně nabízejí dva druhy zamykání kufříku. První kufřík se zamyká šifrou, která se skládá z šesti číslic. Druhý kufřík se zamyká dvěma zámky, které se otevírají současně. Šifra každého z nich se skládá ze tří číslic. Určete pro každý kufřík pravděpodobnost otevření zlodějem při prvním pokusu. Který typ zámku je bezpečnější?

L31: ## Příklad 3.

L32: V urně je 40 koulí - 2 červené a 38 bílých. Z urny náhodně vytáhneme 2 koule. S jakou pravděpodobností budou obě červené?

L33: ## Příklad 4.

L34: Student si měl ke zkoušce připravit odpovědi na 40 otázek. Na dvě otázky, které mu dal zkoušející, neuměl odpovědět a tak řekl „To mám smůlu! To jsou jediné dvě otázky, na které neumím odpovědět.“ S jakou pravděpodobností mluví pravdu?

L35: ## Příklad 5.

L36: Test z chemie žák složí, pokud v seznamu 40 chemických sloučenin podtrhne jediné dva aldehydy, které v seznamu jsou. Jaká je pravděpodobnost, že test složí žák, který provede výběr sloučenin náhodně?

L37: ## Příklad 6.

L38: Ze zahraničí se vracela skupina 40 turistů a mezi nimi byli 2 pašeráci. Na hranici celník 2 turisty vyzval k osobní prohlídce a ukázalo se, že oba dva jsou pašeráci. Zbylí turisté na to reagovali: „Celník měl opravdu štěstí!“, „Pašeráky někdo udal!“, . . .. Jak se postavit k těmto výrokům? Je oprávněné podezření, že pašeráky někdo udal?

L39: ## Příklad 7.

L40: Z urny se třemi koulemi, dvěma červenými a jednou bílou, budou současně vybrány dvě koule. Student a učitel uzavřou sázku. Pokud budou obě koule stejné barvy, vyhraje student. Pokud budou mít koule různou barvu, vyhraje učitel. Je hra férová? Jaké jsou pravděpodobnosti výhry učitele a studenta?

L41: funkce combn vyrobí kombinace o předepsané velikosti - první parametr je vektor hodnot, druhý velikost výběru

L42: ## Příklad 8.

L43: Hra popsaná v příkladu 7 nebyla férová. Jakou kouli (červenou nebo bílou) musíme do urny přidat, aby hra férová byla?

L44: ## Příklad 9.

L45: Chcete hrát Člověče nezlob se, ale ztratila se hrací kostka. Čím a jak lze nahradit hrací kostku, máte-li k dispozici hrací karty (balíček 32 karet) a 4 různobarevné kuličky?

L46: ## Příklad 10.

L47: Chcete hrát Člověče nezlob se, ale ztratila se hrací kostka. Jak lze nahradit hrací kostku, máte-li k dispozici 3 různobarevné kuličky?

L48: ## Příklad 11.

L49: V prodejně vozů Škoda mají v měsíci únoru prodejní akci. Ke standardnímu vybavení nabízejí 3 položky z nadstandardní výbavy zdarma. Nadstandardní výbava zahrnuje 7 položek: - tempomat, vyhřívání sedadel, zadní airbagy, xenonová světla, stropní okénko, bezpečnostní zámek převodovky, speciální odolný metalízový lak. Kolik možností má zákazník, jak zvolit 3 položky z nadstandardní výbavy?

L50: ## Příklad 12.

L51: Při zkoušce si do 5. řady sedlo 12 studentů. Zkoušející chce určit sám, jak tyto studenty v řadě rozesadit. - Kolik je možností jak studenty rozesadit? - Student Brahý žádá, aby mohl sedět na kraji a odejít dříve, aby stihl vlak. Kolik je možností jak studenty rozesadit, chce-li zkoušející vyhovět požadavku studenta Brahého? - Kolik je možností jak studenty rozesadit, nesmějí-li Pažout a Horáček sedět vedle sebe?

L52: a

L53: b

L54: c

L55: ## Příklad 13.

L56: Kolik anagramů lze vytvořit ze slova STATISTIKA?

L57: ## Příklad 14.

L58: V Tescu dostali nové zboží – 6 druhů chlapeckých trik. Od každého druhu mají alespoň 7 kusů. Maminka chce synovi koupit 4 trika. Kolik je možností, jak je vybrat - mají-li být všechna různá? - připouští-li, že mohou být všechna stejná?

L59: a

L60: b

L61: ## Příklad 15.

L62: Kolik hesel délky 5 můžeme vytvořit ze znaků abecedy - nejsou-li rozlišována velká a malá písmena? - jsou-li rozlišována velká a malá písmena?

L63: a

L64: b

%%%%%%%%%% cv11

L0: # Cvičení 11. Dvou-výběrové testy/Intervalové odhady ## Michal Béreš, Martina Litschmannová

L1: # Přehled testů/konstrukcí IO

L2: ## Párová vs. dvouvýběrová data

L3: - Párová data označují data, která jsou vstažena k dvěma měřením stejných entit -> datové sloupce jsou závislé. - Pokud jsou nezávislé, jedná se o dvouvýběrový test. - Pro párová data napočítáme rozdíl mezi sloupci (případně jinou funkci dle zadání) a použijeme jednovýběrové testy na tento rozdíl.

L4: ### Příklady párových dat:

L5: - měření žárovek při dvou různých teplotách (pokud každý kus je měřen dvakrát - při teplotě 1 a teplotě 2) - zde pozor, může se stát, že jsou testy např. destruktivní a nelze měřit dvakrát stejnou entitu (výrobek). Potom bychom uvažovali dva nezávislé výběry, každý pro jeden typ měření -> nezávislé datové sloupce -> dvouvýběrové testy - měření hodnot v krvi pacienta před a po podání léčiva - opět pozor na například testování léků ve dvou skupinách (placebo/skutečný lék) -> dvě nezávislé skupiny -> dvouvýběrové testy

L6: ## Obecně k dvouvýběrovým testům/IO

L7: - test je vždy svázán s příslušným IO -> stejné podmínky pro použití - pokud má test podmínky použití (např.: normalita dat, symetrie dat) pak musí tuto podmínku splnit \*\*oba soubory\*\*, pokud alespoň jeden nesplňuje, považujeme předpoklad za porušený - jeden z velmi důležitých předpokladů je nezávislost dat - např.: měření výrobků výrobce A a výrobků výrobce B - zde je rozumné předpokládat, že výrobky výrobce A jsou samostatné entity od výrobků výrobce B

L8: ## Dvouvýběrové testy/IO - rozdíl měr polohy

L9: vyrobíme si testovací data - tak je lze použít všude

L10: ### Dvouvýběrový Studentův t-test

L11: - Testuje/odhaduje rozdíl středních hodnot: $H\_0: \mu\_{1} - \mu\_{2} = a$ - požadavky: - Normalita dat - Homoskedasticita (shoda rozptlyů) - nezávislost výběrů - funkce musí mít parametr var.equal = TRUE

L12: H0: mu1 - mu2 = 2

L13: HA: mu1 - mu2 != 2

L14: H0: mu1 - mu2 = 2

L15: HA: mu1 - mu2 > 2

L16: H0: mu1 - mu2 = 2

L17: HA: mu1 - mu2 < 2

L18: ### Aspinové-Welshův test

L19: - Testuje/odhaduje rozdíl středních hodnot: $H\_0: \mu\_{1} - \mu\_{2} = a$ - požadavky: - Normalita dat - nezávislost výběrů - funkce musí mít parametr var.equal = FALSE

L20: H0: mu1 - mu2 = 2

L21: HA: mu1 - mu2 != 2

L22: H0: mu1 - mu2 = 2

L23: HA: mu1 - mu2 > 2

L24: H0: mu1 - mu2 = 2

L25: HA: mu1 - mu2 < 2

L26: ### Mannův-Whitneyův test

L27: - Testuje/odhaduje rozdíl mediánů: $H\_0: X\_{0.5,1} - X\_{0.5,2} = a$ - požadavky: - nezávislost výběrů - (stejný typ rozdělení) - vyžaduje conf.int = TRUE, pro spočtení IO

L28: H0: X0.5,1 - X0.5,2 = 2

L29: HA: X0.5,1 - X0.5,2 != 2

L30: H0: X0.5,1 - X0.5,2 = 2

L31: HA: X0.5,1 - X0.5,2 > 2

L32: H0: X0.5,1 - X0.5,2 = 2

L33: HA: X0.5,1 - X0.5,2 < 2

L34: ## Dvouvýběrové testy/IO - podíl rozptylů

L35: ### F-test

L36: - Testuje/odhaduje podíl rozptylů: $H\_0: \sigma^2\_{1} / \sigma^2\_{2} = a$ - požadavky: - normalita dat - nezávislost výběrů

L37: H0: sigma1^2/sigma2^2 = 1

L38: H0: sigma1^2/sigma2^2 != 1

L39: H0: sigma1^2/sigma2^2 = 1

L40: H0: sigma1^2/sigma2^2 > 1

L41: H0: sigma1^2/sigma2^2 = 1

L42: H0: sigma1^2/sigma2^2 < 1

L43: ### Leveneův test

L44: - Testuje rovnost rozptylů: $H\_0: \sigma^2\_{1} = \sigma^2\_{2}$ ! - požadavky: - nezávislost výběrů - vyžaduje data ve standardním datovém formátu - funkce leveneTest v balíčku car

L45: vyrobíme data ve standardním datovém formátu

L46: install.packages("car")

L47: H0: sigma1^2 = sigma2^2

L48: HA: sigma1^2 != sigma2^2

L49: ## Dvouvýběrové testy/IO - rozdíl pravděpodobností

L50: ### Test homogenity dvou binomických rozdělení

L51: - Testuje shodu/odhaduje rozdíl pravděpodobností: $H\_0: \pi\_{1} - \pi\_{2} = 0$ - požadavky: - dostatečná velikost výběrů: $n\_i>\frac{9}{p\_i(1-p\_i)}$ - nezávislost výběrů

L52: vyrobíme si vhodná data

L53: H0: pi1 - pi2 = 0

L54: HA: pi1 - pi2 != 0

L55: H0: pi1 - pi2 = 0

L56: HA: pi1 - pi2 > 0

L57: H0: pi1 - pi2 = 0

L58: HA: pi1 - pi2 < 0

L59: # Příklady

L60: ## Příkald 1.

L61: Data v souboru cholesterol2.xls udávají hladinu cholesterolu v krvi mužů dvou různých věkových skupin (20-30 letých a 40-50 letých). Ověřte na hladině významnosti 0,05 hypotézu, zda se hladina cholesterolu v krvi starších mužů neliší od hladiny cholesterolu v krvi mladších mužů.

L62: Načtení dat

L63: Převod do standardního datového formátu

L64: Explorační analýza

L65: Odstranění odlehlých pozorování:

L66: pozor v datech máme NA a musíme s tím dále počítat!!!

L67: (např. u zjištění délky)

L68: zaokrouhlování -> 3 platné cifry -> dle sd na tisíciny

L69: \*\*Test o shodě středních hodnot / mediánů\*\*

L70: Ověření normality

L71: normalita na hl. významnosti 0.05 OK

L72: Ověření shody rozptylů

L73: Exploračně

L74: explorační posouzení: poměr nejvetšího ku nejmenšímu je > než 2

L75: -> nepředpokládám shodu rozptylu

L76: Exaktně pomocí F-testu

L77: H0: sigma.starsi = sigma.mladsi

L78: Ha: sigma.starsi <> sigma.mladsi

L79: vyberu si žádaná data

L80: Na hl. významnosti 0.05 zamítáme předpoklad o shodě rozptylů

L81: Pozorovanou neshodu mezi rozptyly lze na hladině významnosti 0,05

L82: označit za statisticky významnou.

L83: Ověření shody středních hodnot (Aspinové-Welchův test)

L84: H0: mu.starsi - mu.mladsi = 0

L85: Ha: mu.starsi - mu.mladsi != 0

L86: na hl. významnosti 0.05 zamítáme H0-> existuje stat. významný rozdíl.

L87: H0: mu.starsi = mu.mladsi (mu.starsi - mu.mladsi = 0)

L88: Ha: mu.starsi > mu.mladsi (mu.starsi - mu.mladsi > 0)

L89: Na hladině významnosti 0,05 zamítáme předpoklad o shodě středních

L90: hodnot cholesterolu ve skupinách mladších a starších mužů ve prospěch

L91: alternativy, že starší muži mají vyšší střední hladinu cholesterolu

L92: než muži mladší

L93: Dle výsledků výběrového šetření očekáváme, že střední obsah

L94: cholesterolu v krvi straších mužů bude cca o 0,524 mmol/l vyšší než

L95: střední obsah chol. u mladších mužů. Dle 95% levostranný interv.

L96: odhadu daného rozdílu očekáváme střední obsah cholesterolu u

L97: starších mužů minimálně o 0,457 mmol/l větší než stř. hodnota

L98: cholesterolu u mladších mužů.

L99: ## Příklad 2.

L100: Údaje v souboru deprese.xls představují délku remise ve dnech z prostého náhodného výběru ze dvou různých skupin pacientů (pacienti s endogenní depresi a pacienti s neurotickou depresí). Ověřte, zda je pozorovaný rozdíl mezi průměrnou délkou remise u těchto dvou skupin pacientů statisticky významný.

L101: Načtení dat z xlsx souboru (pomoci balíčku readxl)

L102: Převod do standardního datového formátu

L103: Explorační analýza

L104: Data neobsahují odlehlá pozorování.

L105: zaokrouhlování -> 3 platné cifry -> dle sd na jednotky

L106: \*\*Test o shodě středních hodnot / mediánů\*\*

L107: Ověření normality

L108: Předpoklad normality ověříme Shapirovovým - Wilkovovým testem.

L109: Na hl. významnosti 0,05 zamítáme předpoklad normality

L110: alespoň orientačne zkontrolujeme podobnost rozdělení

L111: vybereme si data pro jednodušší spracování

L112: Ověření shody mediánů (Mannův - Whitneyho test)

L113: Dle histogramů předpokládáme, že data mají stejný typ rozdělení.

L114: H0: med.neuro = med.endo (med.neuro - med.endo = 0)

L115: Ha: med.neuro != med.endo (med.neuro - med.endo != 0)

L116: na hl. významnosti 0.05 zamítáme H0-> existuje stat. významný rozdíl

L117: H0: med.neuro = med.endo (med.neuro - med.endo = 0)

L118: Ha: med.neuro > med.endo (med.neuro - med.endo > 0)

L119: Na hladině významnosti 0,05 zamítáme hyp. o shodě mediánů dob do

L120: remise onemocnění pro obě skupiny pacientů ve prospěch alternativy

L121: Medián doby remise je u pacientů s neurotickou depresí statisticky

L122: významně větší než u pacientů s endogenní depresí.

L123: Doba remise pacientů s neurotickou depresí je cca o 191 dnů delší

L124: než u pacientů s endogenní depresí. Dle 95% levostranného

L125: intervalového odhadu očekáváme, že pacienti s neuro. depresí mají

L126: minimálně o 168 dní delší dobu remise než pacienti s endo. depresí.

L127: ## Příklad 3.

L128: Sledujeme osmolalitu moči na lůžkové stanici v 08:00 hodin a v 11:00 hodin u 16 mužů. Na základě výsledků uvedených v souboru osmolalita.xls ověřte, zda se osmolalita statisticky významně zvýšila.

L129: Načtení dat

L130: Výpočet nárůstu osmolality

L131: Explorační analýza

L132: Data obsahují odlehlá pozorování.

L133: Odstranění odlehlých hodnot

L134: Explorační analýza pro data bez odlehlých pozorování

L135: zaokrouhlování -> 2 platné cifry -> dle sd na jednotky

L136: Ověření normality

L137: Předpoklad normality ověříme Shapirovým - Wilkovým testem.

L138: Na hl. významnosti 0.05 nelze předpoklad normality zamítnout

L139: (Shapirův-Wilkův test, W = 0,949, p-hodnota=0,545).

L140: Párový t-test

L141: H0: mu.narust = 0 mm

L142: Ha: mu.narust > 0 mm

L143: Dle výběrového šetření lze očekávat, že osmolalita moči se

L144: mezi 8 a 11 hodinou zvýší o cca 24 mmol/kg. Dle 95% intervalového odhadu

L145: lze očekávat, že dojde k navýšení osmolality minimálně o 10 mmol/kg).

L146: Na hladině významnosti 0,05 lze tento nárůst označit za statisticky

L147: významný (párový t-test, t = 3,1, df = 13, p-hodnota = 0,005).

L148: ## Příklad 4.

L149: Byly testovány polovodičové součástky dvou výrobců - MM a PP. MM prohlašuje, že její výrobky mají nižší procento vadných kusů. Pro ověření tohoto tvrzení bylo z produkce MM náhodně vybráno 200 součástek, z nichž 14 bylo vadných. Podobný experiment byl proveden u firmy PP s výsledkem 10 vadných ze 100 náhodně vybraných součástek.

L150: ### a)

L151: Otestujte tvrzení firmy MM čistým testem významnosti.

L152: Ověření předpokladů

L153: Dále pro obě firmy předpokládáme, že n/N < 0.05, tj. že daná populace (součástek) má rozsah alespoň 20 \* n, tj. 20 \* 200 (4 000), resp. 20 \* 150 (3 000) součástek, což je asi vcelku reálný předpoklad.

L154: Pearsonův X2 test

L155: H0: pi.PP = pi.MM

L156: Ha: pi.PP > pi.MM

L157: Vzhledem k p-hodnotě > hl. významnosti 0.05 nezamítáme H0 - tedy předpokl.

L158: shodné chybovosti. Nelze tedy říci, že firma MM má kvalitnější produkci.

L159: Pearsonův X2 test

L160: H0: pi.PP = pi.MM

L161: Ha: pi.PP != pi.MM

L162: ### b)

L163: Otestujte tvrzení firmy MM prostřednictvím intervalového odhadu na hladině významnosti 0,05.

L164: Na základě 95% Clopperova - Pearsonova pravostranného intervalového odhadu

L165: (-0,036; 1,000) lze pozorovaný rozdíl v kvalitě výroby označit za

L166: statisticky nevýznamný. Ke stejným závěrům můžeme dojít i na základě

L167: Pearsonova pravostranného testu