Literatura

Hendl, J. (2004). *Přehled statistických metod zpracování dat*. Praha: Portál.

Lužný, D., & Navrátilová, J. (2001). Náboženství a sekularizace v České republice. *Sociální studia*, (6), 111–125.

Řehák, J. (1976). Základní deskriptivní míry pro rozložení ordinálních dat. *Sociologický časopis*, *12*(4), 416–431.

Swoboda, H. (1977). *Moderní statistika*. Praha: Svoboda.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.

Kapitola 4

Normální a standardizované normální rozdělení

4.1 Normální rozdělení

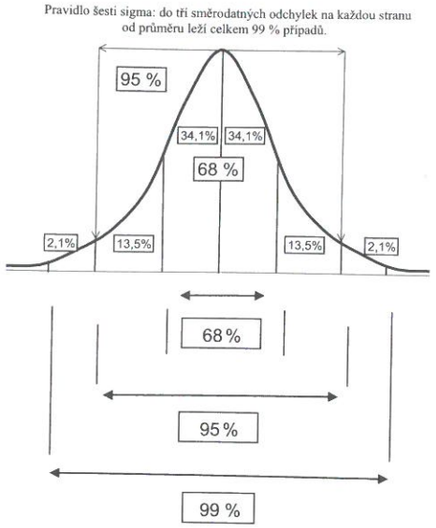
V předchozích lekcích jsme si ukázali, že předtím než začneme analyzovat data, je u proměnných měřených na intervalové úrovni (tedy proměnných spojitých, kardinálních) vždy dobré zjistit, jaký tvar má rozdělení jednotlivých znaků. Zajímá nás především, zdali má distribuce četností tvar rozdělení normálního. Ve statistice totiž provádíme řadu nejrůznějších testů – zde nemáme na mysli, že studenti jsou znovu a znovu zkoušeni z toho, co už ve statistice umí –, což znamená, že sledujeme, do jaké míry naše data odpovídají nějakému statistickému modelu (blíže k tomu v následující kapitole). Abychom mohli tyto testy provádět, musejí být splněny některé předpoklady – a právě předpoklad normálního rozdělení je často jedním z nich.

Normální rozdělení má podobu zvonovité křivky (však mu také angličtina říká *bell curve*, podobně němčina *Glockenkurve*) symetrické kolem střední osy (viz graf 4.1). Ve vědeckém jazyce se hovoří o **Gaussově křivce** (podle německého matematika a fyzika Karla Friedricha Gausse 1777–1855) nebo také o křivce normálního rozdělení.

Normální rozdělení je typické pro řadu biologických nebo psychických jevů, ale také pro některé vlastnosti sociální.[97] Podle francouzského matematika, statistika a astronoma belgického původu Adolpha Quételeta (1796–1874) normální rozdělení neznamená nic jiného než to, že příroda se snaží vytvořit ideální typ (reprezentovaný průměrem), avšak v různé míře (náhodně) chybuje.[98]

[97] Výraz „normální" je zde poněkud zavádějící – zvláště v sociálních vědách, kde mnoho proměnných je rozloženo jiným způsobem, takže mají podobu rozdělení ne-normálního. Slovo „normální" se v sousloví „normální rozdělení" vztahuje k staršímu významu „řídící se zákonem, předpisem nebo modelem".

[98] Reisenauer (1970) uvádí, že normální rozdělení je pozorováno při opakovaném měření téže veličiny za stejných podmínek. Jednotlivé naměřené hodnoty se v důsledku působení náhodných vlivů více či méně odchylují od skutečné hodnoty, jinými slovy jsou zatíženy tzv. náhodnými chybami.



Graf 4.1 Křivka normálního rozdělení a její základní charakteristiky (σ)

Gauss sám křivku normálního rozdělení vysvětloval takto: „Nesčetné dílčí vlivy vyvolávají větší nebo menší odchylky od *průměru*, který všude nacházíme, a tato náhodná kombinace náhodných vlivů podléhá nakonec *zákonům* hazardní hry, pravidlům binomického rozdělení s téměř nekonečným počtem pokusů. Tato úvaha nachází matematický výraz v centrální limitní větě – s její pomocí lze ukázat, že vždy se musí dojít alespoň přibližně k normálnímu rozdělení, jestliže je znak určen působením většího počtu navzájem nezávislých vlivů, ať již je každý z těchto jednotlivých faktorů rozdělen jakkoli." (In Swoboda, 1977, s. 80).

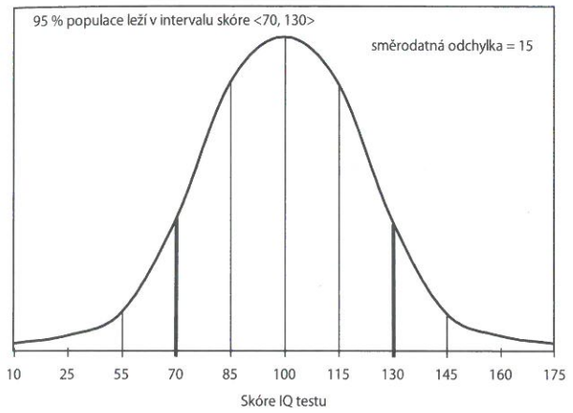
Rámeček 4.1

Pro statistiku je normální rozdělení navýsost důležité kvůli svým následujícím charakteristikám:

- Většina hodnot se soustřeďuje kolem průměru a jejich distribuce je symetrická: polovina hodnot je větších než průměr a polovina hodnot je menší než průměr. Průměr je tedy v normálním rozdělení také mediánem. Průměrná hodnota je také nejčastěji se vyskytující hodnotou, takže je současně i modem.
- Normální rozdělení má jeden vrchol (je tedy jednomodální). Má tvar zvonu, jeho levá strana je zrcadlovým obrazem pravé strany a obráceně; jeho aritmetický průměr, medián i modus mají stejnou hodnotu.
- Můžeme vždy vypočítat procento případů spadajících do určitého intervalu kolem průměru. Platí, že do jedné směrodatné odchylky na každou stranu spadne 68,26 % případů; do dvou směrodatných odchylek na každou stranu od průměru, tedy do čtyř sigma (σ = sigma je symbol pro teoretickou směrodatnou odchylku), spadne většina pozorovaných hodnot, přesně 95,34 %. Řečeno jinak: je 95% pravděpodobnost, že případ bude ležet v intervalu ± 2 σ kolem průměru –

v přesném vyjádření to je 1,96, ale zaokrouhlení na hodnotu 2 v našem případě zcela postačuje.[99] Do šesti sigma pak padne přesně 99,7 % pozorovaných hodnot (tedy v rozsahu +3 a −3 směrodatných odchylek – viz graf 4.1).

Převedeme-li tento fakt do empirické roviny, znamená to například, že v IQ testech, kde se předpokládá, že průměr = 100 (je to vlastně standard určující, jakého skóre by měl „normální" jedinec daného věku v dané kultuře dosáhnout) a směrodatná odchylka (σ) je 15, spadne 68 % populace mezi hodnoty IQ = 85 a IQ = 115, tedy jednu σ na každou stranu od průměru 100, a 95 % populace se pohybuje mezi hodnotami IQ = 70 a IQ = 130 (viz graf 4.2). Jen asi 5 % populace se tomuto intervalu vymyká. Zkuste sami spočítat kolik procent jedinců má IQ nad 130.



Graf 4.2 Rozdělení skóre v IQ testu

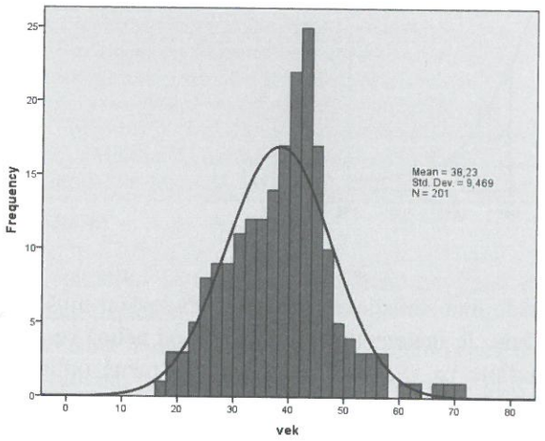
Normální rozdělení je, jako každé jiné statistické rozdělení, především myšlenkovým modelem a početní pomůckou. Je matematickým ideálem, od něhož se empirická rozdělení hodnot, které měříme ve výzkumech, tu více, tu méně odlišují. Je to ovšem tak důležitý statistický model, že všechna rozdělení při statistické analýze s ním srovnáváme a zajímáme se, do jaké míry se naše (empirické) rozdělení tomuto ideálu podobá. Zjistíme-li, že naše proměnné jsou rozloženy alespoň přibližně normálně, máme vyhráno, neboť ve statistické analýze můžeme použít mnohé postupy, které jsou na předpokladu normálního rozdělení založeny. První otázkou, kterou si tedy v analýze dat musíme položit, je, zdali je naše rozdělení normální. Pokud zjistíme, že tomu tak není, pak se musíme ptát, co s hodnotami proměnné udělat, aby se normálním stalo.

[99] Hodnotu 1,96 si obzvláště dobře zapamatujme, je totiž ve statistice svým způsobem hodnotou magickou.

4.1.1 Jak zjistit, zdali je rozdělení normální?

A) Nejjednodušším způsobem, jak zjistit, zdali je naše proměnná normálně rozložena, je nechat si v SPSS udělat **histogram** jejího rozdělení, do něhož vložíme křivku normálního rozdělení a „okometricky" zjistíme, jak se naše rozdělení odchyluje od rozdělení normálního. Potřebné příkazy již známe: *Analyze – Descriptive Statistics – Frequencies* (přitom odstraníme požadavek na tabulku frekvencí tak, že myší odklikneme políčko *Display frequency tables* vlevo dole) – *Charts – Histograms* (a klikneme na políčko *With normal curve*).

Jako příklad si ukažme rozdělení proměnné věk, která vznikla zaznamenáním věku u respondentů, kteří odebírají deník MF Dnes jako předplatitelé (pracujeme se souborem „predplatitele.sav", viz výstup 4.1). „Okometrická" analýza naznačuje, že rozdělení se do určité míry vychyluje od normálního, neboť je špičatější, než naznačuje normální křivka. Otázkou v takové situaci vždy je, zdali odchylka od normálu je natolik malá, abychom dané rozdělení mohli považovat alespoň za přibližně normální a mohli tak naplnit předpoklad mnoha statistických procedur, anebo je už natolik velká, že ji již ignorovat nelze.

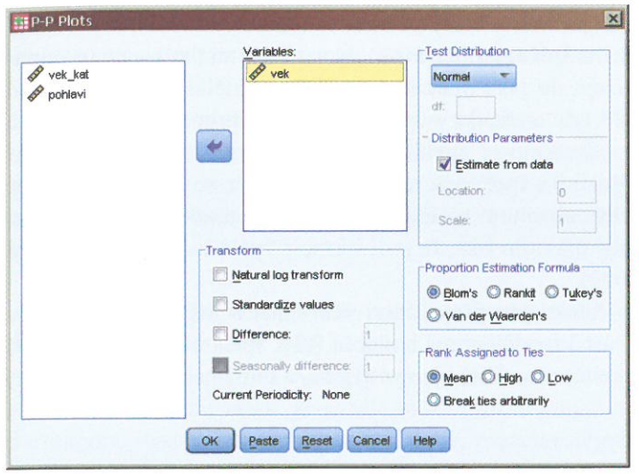


Výstup 4.1 Rozdělení proměnné věk předplatitelů MF Dnes (fiktivní data)
(Stupnice na ose X byla ve výstupu SPSS upravena na desetibodové rozpětí.)

B) Existuje ale ještě další grafický způsob, jak zjistit normalitu našeho rozdělení, a to prostřednictvím **P-P grafu** (*probability to probability plot, P-P plot*). Ten srovnává naše empirické hodnoty, které jsou určitým způsobem standardizovány, se standardizovanými očekávanými hodnotami – to jsou takové hodnoty, jakých by proměnná nabývala, pokud by rozdělení bylo normální.[100] P-P graf získáme tak, že v *Analyze*

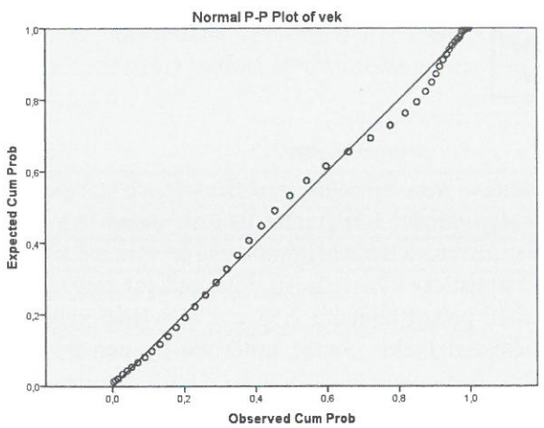
[100] Pracuje se zde určitým způsobem se z-skóry; co z-skóry jsou, si vysvětlíme později. O konceptu empirických a očekávaných hodnot se zmíníme také později, v dalších kapitolách.

– *Descriptive Statistics* klikneme na tlačítko *P-P Plots*. Objeví se toto dialogové okno, v němž do políčka *Variables* (viz obr. 4.1) vložíme tu proměnnou, kterou chceme zkoumat (v našem případě to je proměnná věk).



Obr. 4.1 Zadání pro zobrazení P-P grafu

Po kliknutí na *OK* dostaneme výstup 4.2. V něm srovnáváme, zdali naše empirická (pozorovaná) data odpovídají datům očekávaným – pokud tomu tak je, pak by měly body odpovídající našim datům (v grafu kroužky) ležet na diagonální přímce (ta je modelem normálního rozdělení), což značí, že naše rozdělení je normální. Pokud se od diagonály odchylují, odchylují se naše data i od normálního rozdělení.



Výstup 4.2 P-P graf

Jelikož výstup 4.2 ukazuje hadovitý charakter rozdělení proměnné věk (odchyluje se od přímky), máme další náznak, že naše rozdělení věku čtenářů MF Dnes není přísně normální, byť vychýlení empirických dat se příliš oproti očekávaným neliší.

C) Přesnějším způsobem, jak otestovat symetričnost našeho rozložení, je prozkoumat jeho **šikmost** (*skewness*) a **špičatost** (*kurtosis*). Jsou to dvě statistiky, které sumarizují tvar rozdělení a ukazují, do jaké míry se empirické rozdělení dat odlišuje od normálního rozdělení. SPSS tyto statistiky vypočítává např. v proceduře *Descriptive Statistics – Frequencies – Statistics* – zakliknout *Skewness* a *Kurtosis*. Normální rozdělení má hodnotu šikmosti 0 a špičatosti rovněž 0. Čím více se hodnoty šikmosti a špičatosti odchylují od 0 (v absolutní hodnotě), tím více se rozdělení vzdaluje od rozdělení normálního. Hrubé pravidlo říká, že je-li šikmost vyšší než 1, pak rozdělení není normální (de Vaus, 2002).

Hodnoty šikmosti a špičatosti pro proměnnou věk čtenářů MF Dnes zobrazuje výstup 4.3. Šikmost (*skewness*) rozdělení má hodnotu 0,36, špičatost (*kurtosis*) 0,83. Máme tedy určité indicie, že toto rozdělení je vychýlené, a tudíž není normální.

vek

N	Valid	201
	Missing	0
Mean		38,23
Median		39,00
Mode		42
Std. Deviation		9,469
Skewness		**,360**
Std. Error of Skewness		,172
Kurtosis		**,833**
Std. Error of Kurtosis		,341

Výstup 4.3 Údaje pro šikmost a špičatost rozdělení pro proměnnou *věk*

Pokud je vypočtený výsledek šikmosti nebo špičatosti podělený jejich standardní chybou (*Std. Error*) v absolutní hodnotě vyšší než **1,96** (vidíte, již jsme u oné slibované magické hranice), můžeme si být jisti, že rozdělení naší proměnné je výrazně šikmé nebo špičaté, neboť výsledek nabývá statistické významnosti.[101] Ve velkých souborech bychom jako kritérium asymetrie měli použít hodnotu 2,58 a v obzvláště velkých souborech, jak zdůrazňuje s vykřičníkem Field, „žádné kritérium by nemělo být

[101] Proč právě 1,96? Vzpomeňme si, že u normálního rozdělení spadá 95 % případů do dvou (přesněji do *1,96*) směrodatných odchylek na každou stranu. Těchto 95 % případů je považováno ve statistice za případy normální. To, co spadá nad tuto hranici (to je 5 % případů), je už podle statistické konvence odlišné od normality, a tudíž statisticky signifikantní.

aplikováno!“ (Field, 2009, s. 139), neboť u skutečně velkých souborů (kdy n = 200 a vyšší) vznikají statisticky významné (signifikantní) hodnoty i při malých odchylkách od normality. (Tuto větu si pamatujme, uslyšíme ji ještě mnohokrát při seznamování se s problematikou zobecňování výsledků z výběrového souboru na soubor základní.) U obzvláště velkých souborů proto Field doporučuje prozkoumat tvar rozdělení vizuálně (prostřednictvím histogramu) spolu s hodnotami šikmosti a špičatosti.

V našem rozdělení předplatitelů deníku MF Dnes je standardní chyba šikmosti 0,172 (viz řádek *Std. Error of Skewness* ve výstupu 4.3), z-skóre je tedy 0,36 / 0,172 = = 2,09. Pro špičatost platí 0,833 / 0,341 = 2,44. Vzhledem k tomu, že náš soubor má n = 201, je pro nás doporučenou kritickou hodnotou 2,58. Jelikož naše vypočtené hodnoty jsou pod touto kritickou hranicí, je toto rozdělení možno považovat za blízké normálnímu.

D) Dalším ze způsobů, jak testovat normalitu rozdělení, je použití **Kolmogorovova–Smirnovova testu (K-S test)**.[102] Tento formální statistický test ověřuje nulovou hypotézu, že naše data jsou výběrem z normálního rozdělení. Řečeno jinak, tento test statisticky hodnotí, zdali je rozdíl mezi pozorovaným rozdělením a teoretickým normálním rozdělením natolik malý, že jej můžeme připsat náhodě, to je výběrové chybě.[103] Pokud je ovšem tato diference větší, pak naše pozorované rozdělení není normální. Pro aplikaci K-S testu to znamená, že pokud vypočtená statistická významnost (signifikance, značená jako *Sig.*) bude větší než 0,05, pak test vychází jako statisticky nevýznamný (blíže k tomu v následující kapitole), což znamená, že rozdělení sledované proměnné se statisticky neodlišuje od normálního rozdělení, tudíž pochází z rozdělení normálního. Pokud ale vypočtená významnost testu bude menší než 0,05, pak se statisticky významně od normality odlišuje a naše rozdělení není normální.

Kolmogorovův–Smirnovův test získáme zadáním následujícího řetězce příkazů: *Analyze – Descriptive Statistics – Explore – Plots* – zaškrtnout *Normality plots with tests* (a odškrtnout *Stem-and-leaf*). Výstupem této procedury je několik tabulek a grafů, nás však zajímá tabulka testu normality (*Tests of Normality*), kterou ukazuje výstup 4.4.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
vek	,073	201	**,011**	,977	201	,002

a. Lilliefors Significance Correction

Výstup 4.4 Hodnoty Kolmogorovova–Smirnovova testu

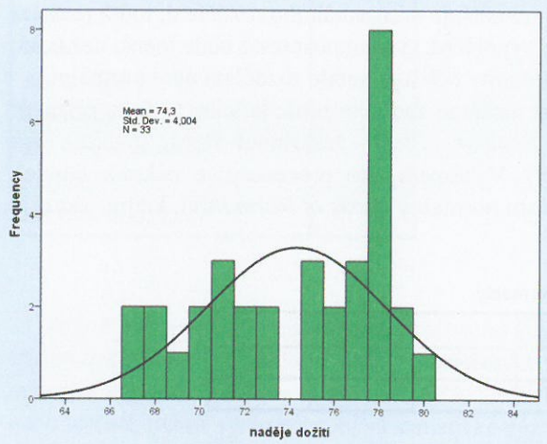
[102] Pro zvídavé dodáváme, že podobnost s vodkou stejně znějícího jména je čistě náhodná.

[103] V případě, že počet případů je menší než 50, SPSS tiskne automaticky v tabulce K-S testu také Shapirův–Wilkův test, který je v takové situaci vhodnější.

Test vychází statisticky významný, neboť hodnota významnosti (*Sig.*) je 0,011, což je méně než hladina 0,05. Z tohoto hlediska není tedy věk našeho souboru předplatitelů rozložen normálně. Musíme zde ale opět poznamenat spolu s Norušisovou (2010) a s Fieldem (2009), že při vyslovování závěrů o normalitě rozdělení našich empirických dat musíme vždy brát v úvahu velikost našeho výběrového souboru. Ve velkých souborech mohou i velmi malé odchylky od normality vyústit ve statisticky významný výsledek,[104] který ale nemusí být věcně důležitý (Norušis, 2010, s. 267). Proto je třeba vždy výsledky formálních statistických testů pečlivě zvažovat. Zlaté pravidlo v tomto kontextu praví: „Pokud je náš výběr velmi velký a rozdělení hodnot není extrémně vzdáleno od normálního, není třeba si dělat starosti. … Obecně řečeno, grafické prozkoumání předpokladů je více informativní než statistický test." (Norušis, 2010, s. 267) Pod grafickým zkoumáním se myslí především kontrola prostřednictvím histogramu, to je „okometricky". Ukažme si nyní celou proceduru kontroly normality rozdělení na skutečných datech.

Příklad 4.1

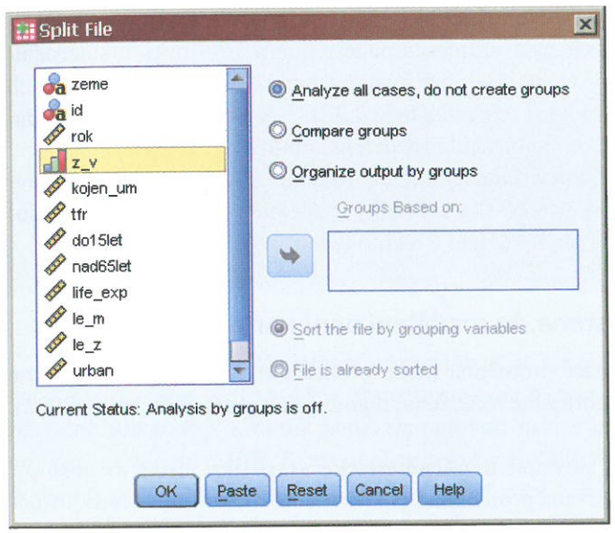
Použijeme demografické údaje o naději dožití z roku 1999 z demografického souboru „dmg-file.sav", to je souboru 33 evropských zemí, u nichž máme demografické charakteristiky – všimněme si, že jednotkou zde nejsou lidé, nýbrž země – a pracovat budeme s proměnnou *life_exp* (neboli *life expectancy*, což je anglický výraz pro demografický termín *naděje dožití*). Podle všeho by měl být tento údaj normálně rozložen. Výstup, který získáme (viz výstup 4.5), ovšem říká něco jiného. Na první pohled naznačuje, jako kdyby se v celkovém rozdělení skrývala rozdělení dvě, neboť vzdáleně připomíná rozdělení dvouvrcholové, bimodální.



Výstup 4.5 Histogram četností proměnné *naděje dožití*

[104] Často se hovoří o tom, že test normality je hodně citlivý na jakoukoliv odchylku od normality.

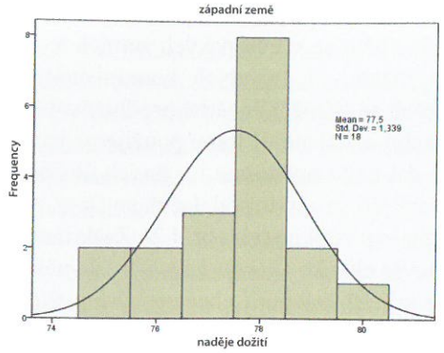
Nastane-li taková situace, vždy přemýšlejme, která vlastnost by se za ní mohla skrývat. Jelikož z demografie víme, že naděje dožití se v evropských zemích v posledních přibližně padesáti letech vyvíjela různě – v bývalých komunistických zemích de facto stagnovala, v zemích západních se (téměř) lineárně prodlužovala,[105] uděláme si analýzu pro země západní a východní odděleně. K tomu použijeme jednu důležitou proceduru SPSS, kterou v analýze dat často uplatníme i v jiných úlohách. Je jí procedura *Split file* (rozděl soubor). Jak na to? Vše je ukryto pod tlačítkem *Data*, pak kliknutím vybereme příkaz *Split file* a získáme dialogové okno (viz obr. 4.2). Zaškrtneme kroužek *Organize output by groups*, tím se otevře okénko *Groups based on*, do něhož pomocí šipky vložíme tu proměnnou, podle jejíchž kategorií chceme výstup třídit. V našem případě to je proměnná *z_v*, která jakožto dichotomická proměnná označuje země západní (kód 1) a východní (kód 2). Kliknutím na *OK* příkaz provedeme.



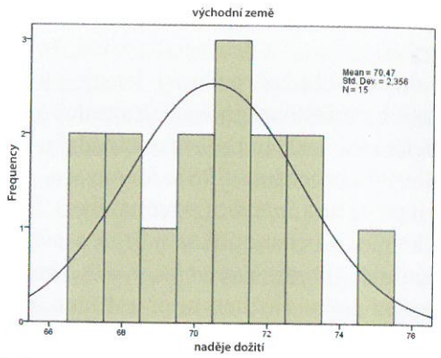
Obr. 4.2 Příkaz *Split file* (rozděl soubor)

Nyní již SPSS ví, že jakoukoliv početní proceduru bude zpracovávat zvlášť pro země západní a pro země východní. Získaný histogram pro naději dožití je zobrazen na výstupech 4.6a a 4.6b.

[105] V současné době se ale některé východoevropské země začaly těm západním svou hodnotou naděje dožití přibližovat – platí to především o Slovinsku a České republice, částečně i Chorvatsku.



Výstup 4.6a



Výstup 4.6b

Obě rozdělení se zřetelně odlišují. Liší se i svými průměry a směrodatnými odchylkami: v západních zemích byla průměrná naděje dožití 77,5 roků a směrodatná odchylka 1,34 (viz Mean a Std. Dev. na výstupu 4.6a), ve východních zemích byla pouze 70,5 roků (a směrodatná odchylka byla 2,36).[106] A histogram pro východní země již „okometricky" říká, že o normalitě rozdělení nemůže být ani řeč.

Z tohoto cvičení vyplývá jeden důležitý závěr. Pokud je ve statistické úloze hlavním cílem srovnání skupin, pak z hlediska normality rozdělení není důležité kontrolovat rozdělení celého souboru, ale rozdělení v těchto jednotlivých skupinách.

4.1.2 Co dělat, když zjistíme, že rozdělení není normální?

V situaci, kdy zjistíme, že naše proměnná (nebo proměnné), kterou (které) chceme statisticky analyzovat, není normálně rozložena, máme, jak napovídá de Vaus (2002), tři možnosti.

1. Použijeme některý z postupů **neparametrické statistiky**. Jsou to postupy, které nevyžadují, aby analyzovaná proměnná byla normálně rozložena. V naší příručce si je postupně představíme v příslušných kapitolách (stručné vysvětlení toho, co je to neparametrická statistika, najde čtenář na konci této kapitoly).

2. Transformujeme statisticky distribuci naší proměnné. Pozor, prosím, znovu opakujeme: **transformace proměnné** neznamená, že upravujeme jednotlivé hodnoty proměnné například proto, aby se „potvrdila" naše hypotéza – to by byl nepřípustný podvod, za který se vyobcovává z vědeckého společenství! Transformujeme-li data, pak každou hodnotu upravujeme stejnou matematickou funkcí. Například tak, že proměnnou logaritmujeme, odmocňujeme, umocňujeme, převedeme na převrácenou hodnotu (1/x) apod. Detailněji si způsob transformace (v jejím širším pojetí, ne pouze v kontextu

[106] Vyšší hodnota směrodatné odchylky ve východních zemích indikuje, že data jsou mnohem více rozptýlena kolem průměru, takže naděje dožití v nich byla v r. 1999 různorodější než v zemích západních.

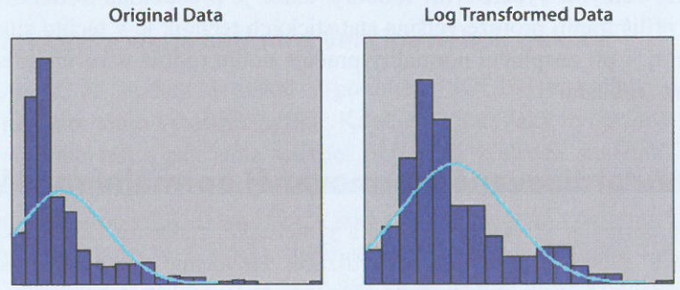
normálního rozdělení) ukážeme v následující kapitole. Na tomto místě si pouze dovolíme malou ukázku. Jde o příklad logaritmické transformace, jak ji uvádí SPSS.[107]

Ilustrace

Hodnoty o prodeji výrobku (řádek *1996 Sales*) byly logaritmovány (viz v tabulce druhý, tučně orámovaný řádek nazvaný *Log of Sales*) a původní výrazně špičaté rozdělení (obrázek vlevo nazvaný *Original data*) se změnilo na rozdělení méně špičaté.

Frequencies Statistics

	Mean	Median	Std. Deviation	Skewness	Kurtosis
1996 Sales	$371,893	$307,500	$171,311	2,112	5,247
Log of Sales	5,5367	5,4878	,1603	1,110	,791



Zda transformovat, či netransformovat data, je ale složitá otázka. Jak upozorňuje např. Field (2009, s. 155–156), transformujeme-li data, pak také poněkud měníme původní konstrukty, s nimiž jsme formulovali naši výzkumnou otázku a jež jsme určitým způsobem měřili. To má pochopitelně následky pro naše interpretace výsledků. Podobně, pokud použijeme neadekvátní transformaci, může to mít pro naše výsledky mnohem horší efekt než práce s původními, netransformovanými proměnnými.

Dodejme ještě v kontextu transformací, že data transformujeme často nejenom kvůli snaze získat normální rozdělení. Poměrně častým způsobem transformace je vytváření tzv. **centrovaných dat**. Vzniknou tak, že od každé hodnoty dané proměnné odečteme průměr této proměnné (mimochodem, tuto operaci jsme již prováděli jako mezivýpočet při hledání rozptylu v kapitole 3). No a samozřejmě velmi užívanou transformací je výpočet z-skórů (též standardizovaných skórů), který převádí původní hodnoty na takové, jejichž průměr je roven 0 a směrodatná odchylka rovna 1.

3. Třetí možností, jak říká de Vaus (2002) a také další moderní statistikové, je nedělat si s tvarem rozdělení starosti a bez obav použít postupy **parametrické statistiky** (ty jsou na předpokladu normálního rozdělení založeny). Statistikové postupně

[107] Mimochodem, vyhledejte si zajímavou nápovědu ke statistickým operacím SPPS na liště *Help–Tutorial*.

ukázali, že porušení požadavku na normalitu nemá tak závažný vliv na výsledky analýzy, jak se původně myslelo. Ačkoliv z teoretického hlediska je porušení předpokladu normality neospravedlnitelné, v praxi se ukazuje, že výsledkům to příliš neškodí.

Ve statistice totiž platí centrální limitní věta/teorém, která stanovuje velmi důležitý princip: **se vzrůstající velikostí (náhodně vybraného) výběrového souboru se výběrová distribuce blíží normálnímu rozdělení** (podrobněji si vše rozebereme v kapitole 5).

Což v praxi znamená, že i když rozdělení naší analyzované proměnné není normální, je možné využívat i statistických postupů, které normální rozdělení předpokládají, je-li náš výběrový soubor dostatečně velký (rozuměj větší než 100).

Pro sociální vědy – a sociologii obzvláště – je velkým štěstím, že většinou pracují s dostatečně velkými výběrovými soubory, takže je předpoklad normality rozdělení nemusí příliš trápit, protože většina statistických technik je v těchto situacích tzv. robustních, tj. i při nesplnění normality pracuje dobře (odolá narušení předpokladu, odtud výraz „robustní“).

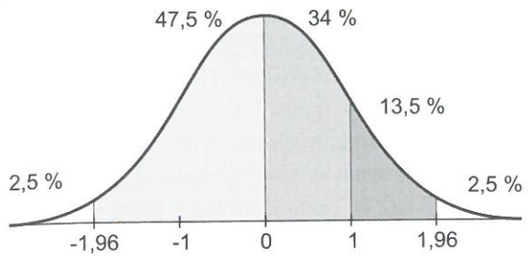
4.2 Standardizované (normované) normální rozdělení

I když jsou normální křivky pravidelné a symetrické, získávají velký praktický význam teprve dalším procesem standardizace. Hovoříme o tzv. standardizované nebo někdy také normované náhodné veličině (z-skóre), jejíž matematickou úpravou se získá tzv. **standardizované normální rozdělení**. Právě toto rozdělení je základem pro tzv. inferenční statistiku, nebo lépe řečeno je základem teorie, které se využívá k odhadům **populačních parametrů z výběrových statistik**[108] (o tom blíže v následujících kapitolách).

Standardizované normální rozdělení je souměrné podle osy, jíž je hodnota průměru (viz obr. 4.6). Průměr je zde roven nule a směrodatná odchylka je rovna jedné. V ploše vymezené mezi −1 a 1 směrodatnou odchylkou leží 68 % pozorování (proč asi?). Do dvou směrodatných odchylek (−2 a +2) leží 95 % případů a do tří směrodatných odchylek (−3 a +3) 99,7 % pozorování. Jednotlivé případy lze vyjádřit nejen „pozorovanou hodnotou“ spojité proměnné, ale i jako počet (násobek) směrodatných

[108] V tomto kontextu si nyní připomeňme některé důležité statistické koncepty. **Parametr** je neznámá vlastnost základního souboru, kterou však můžeme (s určitou pravděpodobností – viz následující kapitolu) odhadovat z **výběrové statistiky**. Takovými parametry může být průměr či rozptyl základního souboru. Průměr základního souboru označujeme symbolem μ (mí). Pro směrodatnou odchylku základního souboru používáme symbolu σ (sigma) a pro rozptyl průměru základního souboru, logicky, symbol σ^2. Výběrové statistiky, vypočtené na základě údajů, které jsme naměřili ve výběrovém souboru, označujeme jinými symboly – to proto, abychom je odlišili od parametrů. Průměr výběrového souboru označujeme symbolem x (někdy se také značí jako x s pruhem), pro směrodatnou odchylku od průměru výběrového souboru používáme označení s a pro rozptyl s^2.

odchylek polohy této hodnoty případu od aritmetického průměru. Vše osvětlí následující pasáž o z-skórech.



Obr. 4.6 Standardizované normální rozdělení

4.2.1 Standardizovaná náhodná veličina neboli z-skóre

Když řekneme, že student X získal v předmětu SOC107 statistická analýza dat 66 bodů, nic nám tento výsledek neříká. Když ale víme, jaký byl průměrný bodový výsledek v tomto testu, pak jsme schopni říci, zdali výsledek studenta X byl lepší nebo horší než průměr. A pokud budeme navíc znát směrodatnou odchylku od tohoto průměru, pak jsme schopni také říci, v jaké standardizované vzdálenosti od průměru se tento výsledek nachází. Takže, zjistili jsme, že v testu z předmětu SOC107 získal student X 66 bodů a student Y 81 bodů. Když víme, že průměrný výsledek v testu byl 70 bodů, můžeme vypočítat, jaká je pozice těchto dvou výsledků vzhledem k celkovému rozdělení hodnot výsledků testu. Nástrojem k tomu jsou **z-skóry**. Abychom je vypočítali, potřebujeme znát kromě průměru také směrodatnou odchylku, neboť vzorec pro výpočet této charakteristiky říká, že od hodnoty dané proměnné odečteme průměr a podělíme směrodatnou odchylkou:

$$\text{z-skór } x' = \frac{x - \bar{x}}{s_x}.$$

Víme-li, že směrodatná odchylka od průměrného bodového skóre v testu z analýzy dat byla 5, pak výsledek studenta X, jenž získal 66 bodů, říká:

a) Student X byl o 4 body (66 − 70 = −4) horší, než byl celkový průměr.

b) Jeho výsledek jej umisťuje do vzdálenosti −0,8 směrodatné odchylky od průměru, neboť (66 − 70) / 5 = −0,8 (= z-skór).

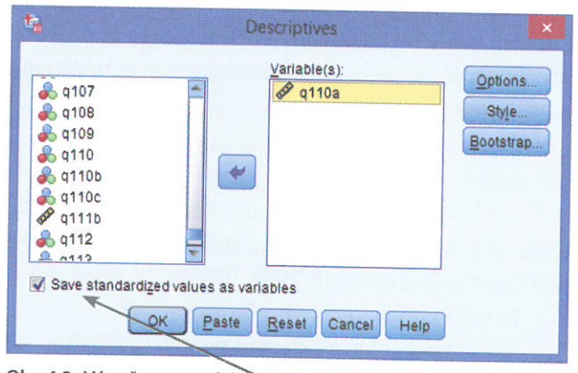
Výsledek studenta Y 81 bodů znamená, že byl o 11 bodů (81 − 70 = 11) lepší než průměr (gratulujeme!) a že byl vzdálen +2,2 směrodatné odchylky od průměru, neboť (81 − 70) / 5 = 2,2 (= z-skór). Tento student dosáhl vskutku skvělého výsledku, neboť již víme, že v normálním rozdělení leží do dvou směrodatných odchylek 95 % (normálních) případů. Jelikož student Y byl +2,2 směrodatné odchylky od průměru, byl ve skupině 2,5 % nejlepších. (Víte proč?)

Hodnota *z* (z-skór) tedy říká, kolik standardních odchylek je určitý případ pod nebo nad průměrem. Je-li vypočtený z-skór roven 0, je případ přesně na průměru, je-li roven +1, je případ jednu směrodatnou odchylku nad průměrem. SPSS z-skóry standardně nejen vypočítává, ale umí je současně i uložit jako novou proměnnou do datové matice, s níž lze dále pracovat.

Z-skóry jsou de facto standardizované hodnoty, neboť převádějí původní proměnné, měřené v různých jednotkách, do jednotné metriky (způsobu měření vzdáleností): do vzdálenosti od průměru vyjádřené počtem směrodatných odchylek.

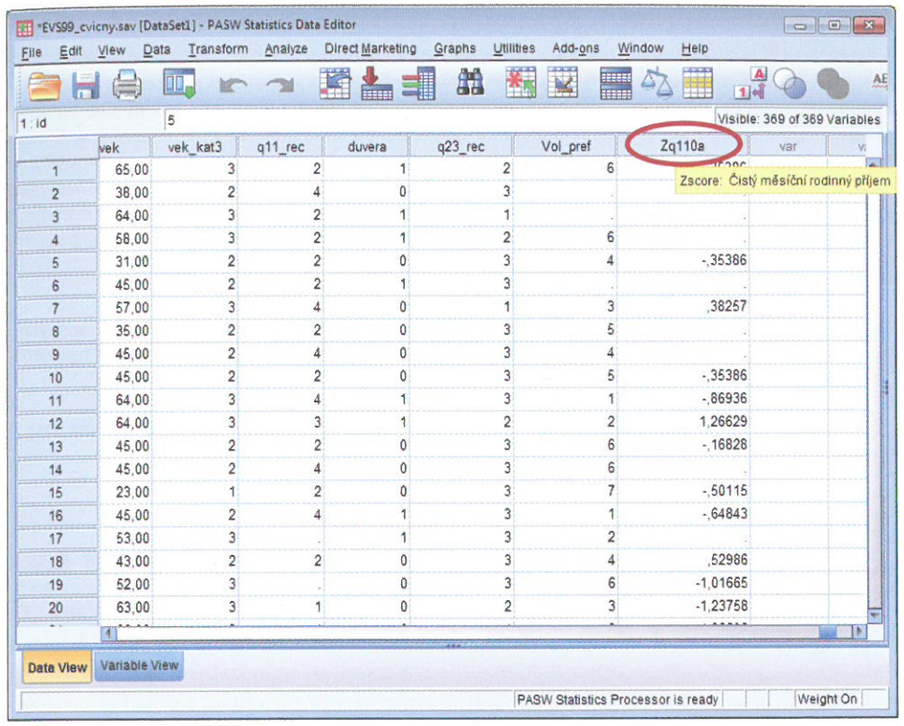
Touto standardizací tak získáváme možnost srovnávat zdánlivě nesrovnatelné, třeba i ony okřídlené hrušky a jablka. Zjistíme například, že respondent má v příjmu hodnotu z-skóru +2,1 a ve vzdělání má –1. Znamená to tedy, že tato osoba je v příjmové kategorii více než dvě směrodatné odchylky nad průměrem (a když se podíváte na obrázek normálního rozdělení a uvědomíte si, že do plochy nad dvě směrodatné odchylky spadá jen 2,14 % případů s nejvyššími hodnotami nad průměrem, pak je to příjem velmi vysoký). Úrovní vzdělání je však dotyčný pod průměrem (jen 34 % osob má nižší či stejné vzdělání). Je to tedy člověk, který ač má nízké vzdělání, patří mezi osoby s nejvyššími příjmy (kdopak to asi je?). Bez standardizace prostřednictvím z-skóru by takovéto srovnání nebylo tak jednoduché, neboť každá proměnná má jiné jednotky měření, odlišné průměry a odlišné směrodatné odchylky. Proto z-skóry umožňují srovnávat hrušky s jablky, což, pokud bychom se řídili pravidlem selského rozumu, by v zásadě nemělo být možné.

Výpočet z-skóre zadáme v proceduře *Analyze – Descriptive Statistics – Descriptives* (viz obr. 4.3, soubor „EVS99-cvicny.sav“).



Obr. 4.3 Výpočet pro z-skóre

Zadáme proměnnou, jejíž z-skóre pro každý z případů chceme vypočítat (zde *q110a*), a zaškrtneme políčko *Save standardized values as variables*. Výsledkem je nová proměnná (SPSS ji při uložení dá automaticky nové jméno předřazením *Z* před jméno původní proměnné: zde tedy *Zq110a*), která bude v matici dat zařazena jako nový (poslední) sloupec a každému případu do něho bude zapsána vypočítaná

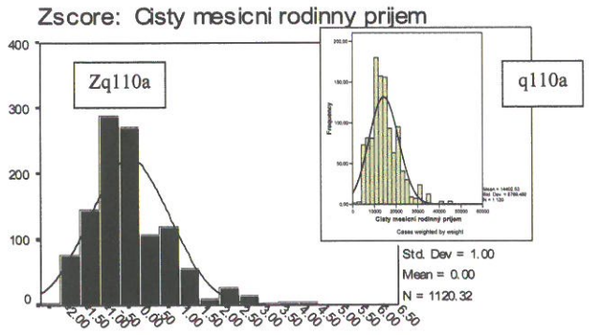


	vek	vek_kat3	q11_rec	duvera	q23_rec	Vol_pref	Zq110a
1	65,00	3	2	1	2	6	.
2	38,00	2	4	0	3	.	.
3	64,00	3	2	1	1	.	.
4	58,00	3	2	1	2	6	.
5	31,00	2	2	0	3	4	-,35386
6	45,00	2	2	1	3	.	.
7	57,00	3	4	0	1	3	,38257
8	35,00	2	2	0	3	5	.
9	45,00	2	4	0	3	4	.
10	45,00	2	2	0	3	5	-,35386
11	64,00	3	4	1	3	1	-,86936
12	64,00	3	3	1	2	2	1,26629
13	45,00	2	2	0	3	6	-,16828
14	45,00	2	4	0	3	6	.
15	23,00	1	2	0	3	7	-,50115
16	45,00	2	4	1	3	1	-,64843
17	53,00	3	.	1	3	2	.
18	43,00	2	2	0	3	4	,52986
19	52,00	3	.	0	3	6	-1,01665
20	63,00	3	1	0	2	3	-1,23758

Obr. 4.4 Datová matice s nově uloženými hodnotami z-skórů proměnné *q110a*

hodnota (standardizovaná vzdálenost od průměru: kladná či záporná) – viz obr 4.4.

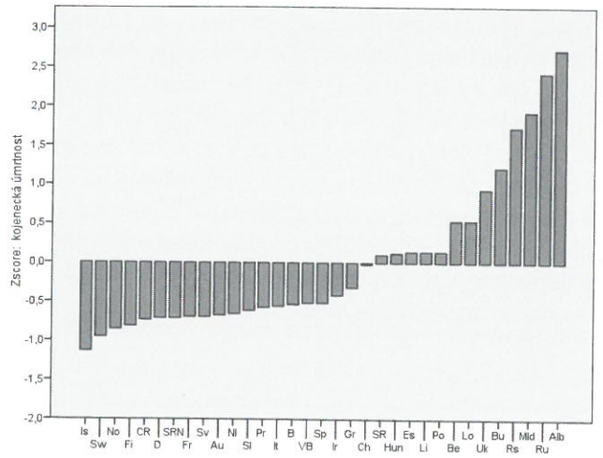
Výstup 4.7 nabízí výsledek rozdělení původní proměnné *q110a* a její standardizované podoby – z-skóry.



Výstup 4.7

Příklad 4.2

Z demografické statistiky máme údaje o kojenecké úmrtnosti v evropských zemích v roce 1999 (viz soubor „dmg-file.sav", proměnná *kojen_um*). Prostřednictvím procedury *Analyze – Descriptive Statistics – Descriptives – Save standardized values as variables* si necháme uložit z-skóry této proměnné. Jak jsme již ukázali před chvílí, uloží se nám jako nová proměnná na konec matice s názvem, který opakuje název původní proměnné s tím, že je před něj předřazeno písmeno *Z*. Proměnná *kojen_um* se tak změní na proměnnou *Zkojen_um*. Nechejme si nyní celý soubor utřídit pomocí procedury *Data – Sort cases – Sort by (Zkojen_um)*, čímž se pořadí matice změní tak, že v prvním řádku se objeví země s nejnižší hodnotou z-skóre kojenecké úmrtnosti (Island) a na posledním (34.) místě Albánie. Když si pak tuto novou proměnnou necháme zpracovat do grafu, získáme obrázek ve výstupu 4.8.[109]

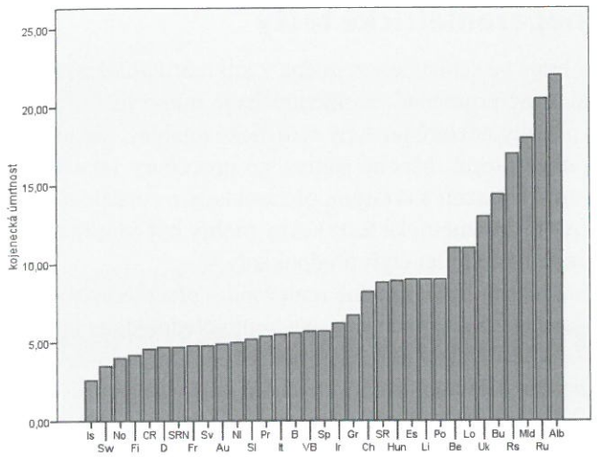


Výstup 4.8 **Pořadí evropských zemí podle z-skórů kojenecké úmrtnosti v roce 1999**

Z obrázku je zřetelně vidět, jak mnoho se jednotlivé evropské země v tomto ukazateli odlišují. Na průměrné hodnotě je Chorvatsko (Ch), hodnoty nižší než průměr (záporné z-skóry), a tudíž nižší kojeneckou úmrtnost, mají všechny západoevropské země, k nimž se z bývalých komunistických zemí řadilo v roce 1999 pouze ČR (naše kojenecká úmrtnost je jedna z nejnižších na světě i v současnosti) a Slovinsko. Bývalé tzv. socialistické země jsou v kladných hodnotách z-skórů, mají tedy vyšší kojeneckou úmrtnost, než je evropský průměr, a mnohé ji mají dokonce velmi vysokou – od průměru jsou vzdáleny o více než dvě směrodatné odchylky.

Je zřejmé, že srovnávat kojeneckou úmrtnost v jednotlivých zemích bychom mohli i prostřednictvím netransformovaných (nestandardizovaných) hodnot. V roce 1999, jak říkají naše údaje, byl průměr kojenecké úmrtnosti v Evropě 8,34 dětí na 1000 živě narozených a směrodatná odchylka byla 4,97. Abychom si uvědomili, v čem spočívá rozdíl mezi nestandardizovanými a standardizovanými hodnotami, vytvořili jsme graf i z hodnot nestandardizovaných (viz výstup 4.9). I tento graf je ilustrativní, avšak nám se zdá, že mnohem větší vypovídací a interpretační schopnost skýtá graf z-skórů.

[109] Graf jsme vytvořili pomocí procedury *Graphs – Chart builder*. Postup při vytváření grafu zde neuvádíme, jsme přesvědčení, že ti, kdo umí grafy v Excelu, si s ním intuitivně poradí i v SPSS.



Výstup 4.9 **Kojenecká úmrtnost v evropských zemích v r. 1999**

4.2.2 K čemu může z-skóre být?

Pro standardizaci hodnot spojitých proměnných (pro z-skóry) existuje řada důvodů, z nichž jsme již některé zmínili. Zejména důležitá je role, kterou standardizované rozdělení hraje v inferenční statistice při odhadu parametrů, to je odhadu hodnot platných v základním souboru z vypočítaných statistik vycházejících z hodnot zjištěných v konkrétním výběrovém souboru. Zmiňme dále, že z-skóre dovolují:

- **porovnávat hodnoty dvou různých distribucí**, například skóre jedince ve dvou různých testech. Může jít o stejný test před a po určité výuce neboli o pre-test a post-test nebo může jít o dva zcela odlišné testy – například jeden, který uchazeč o studium sociologie psal na Masarykově univerzitě, a druhý, který psal na Univerzitě Palackého (dovoluje porovnat i skóre dvou jedinců, z nichž každý absolvoval jinou verzi testu). Dosažená skóre nelze srovnávat přímo, protože testy mohou být různě obtížné nebo různě opravované a bodované;
- **standardizovat data**, a tím je činit souměřitelná. Představme si, že do shlukové analýzy[110] okresů ČR vstupují proměnné o různém měřítku: míra nezaměstnanosti v % (má možnost nabývat hodnot od 0 do 100 %), průměrná výše příjmu okresu (má možnost nabývat hodnot například od 0 Kč do 1 000 000 Kč), podíl osob s vysokoškolským vzděláním atd. Váha proměnných o různém řádu by ve výpočtu byla nesouměřitelná a ve výsledku by proměnné s větším řádem měly nezaslouženě větší váhu. Proto naměřené hodnoty našich proměnných zaměníme za z-skóre.

[110] Shluková analýza (*cluster analysis*) je vícerozměrná statistická technika, která na základě několika proměnných udává, jak jsou si jednotky analýzy navzájem podobné či nepodobné. Umožňuje mimo jiné vytvářet typologie. Více se o ní dozvíte v samostatné kapitole 14.