LEKCE 3
NORMÁLNÍ A STANDARDIZOVANÉ NORMÁLNÍ ROZLOŽENÍ

V předchozích lekcích jsme si ukázali, že před tím, než začneme analyzovat data, je u proměnných měřených na intervalové úrovni vždy dobré se přesvědčit, jaký tvar má rozložení jednotlivých znaků. Zajímá nás především, zdali má distribuce četností tvar **rozložení normálního**.[1] Tato informace je ve statistické analýze dat navýsost důležitá.

Spousta biologických, psychických a některé sociální vlastnosti mají tu charakteristiku, že jsou rozloženy zvláštním symetrickým způsobem kolem střední hodnoty – totiž že jsou rozloženy normálně. Toto rozložení má podobu zvonovité křivky – nazývá se tak v angličtině (*Bell Curve*), ve francouzštině se hovoří o „křivce policejního klobouku". Ve vědeckém jazyce se hovoří o Gaussově křivce nebo také o křivce normálního rozložení (viz obr. 3.1).

Koncept normálního rozložení hraje ve statistice a především v její teorii extrémně důležitou roli. Je především základem teorie, které se využívá k odhadům (ke statistické inferenci) populačních parametrů z výběrových statistik (o tom blíže v lekci čtvrté). Normalita rozložení sledované proměnné (proměnných) je také předpokladem pro to, aby mohly být použity některé postupy statistické analýzy, především postupy tak zvaných parametrických testů.

Mnohé statistické procedury (statistické testy) jsou založeny na tom, že pracují s **parametrickými daty**. Aby data mohla být považována za parametrická, musí splňovat následující čtyři předpoklady (zpracováno podle: Field, Andy. 2000. *Discovering Statistics using SPSS for Windows*. Sage, London.)

1. **Normalita rozložení:** data musí být normálně rozložena – předpokládá se, že data pocházejí z populace, kde jsou normálně rozložena. Na tomto předpokladu je založeno veškeré testování statistických hypotéz. Jelikož však velmi často nemáme informace o celé populaci – vždyť právě z toho důvodu děláme výzkum, abychom se o populaci něco dozvěděli –, není úplně jednoduché tento předpoklad ověřit. Většina výzkumníků si zde zjednodušuje život a ověřuje tento předpoklad „okometricky". Udělají si histogram rozložení proměnné (jako jsme to udělali my s daty výsledků přijímací zkoušky – vit obr. 2.3 v lekci 2) a pokud je toto rozložení přibližně normální, předpokládají, že normálně bude rozloženo i v populaci, z níž jejich data byla vybrána. Jak uvidíme za chvíli, tuto okometrickou metodu bychom měli nahrazovat přesnějšími statistickými testy (testy normality rozložení).
2. **Homogenita rozptylu:** tento předpoklad znamená, že rozptyl v datech (v rozložení jedné proměnné) by se neměl systematicky měnit, že rozptyl náhodné složky je homoskedastický. Např. když bychom sledovali rozložení příjmu, jeho rozptyl by se neměl příliš odlišovat v různých věkových skupinách. Pokud by tomu tak bylo (např. příjem by byl více homogenní ve skupině seniorů než ve skupinách ostatních), tento předpoklad by byl porušen a my bychom měli v datech heteroskadiscitu. Homogenitu rozptylu v různých skupinách můžeme statisticky ověřit, my si později ukážeme Levenovu statistiku, která k tomuto účelu slouží.
3. **Intervalová data:** proměnná by měla být měřena přinejmenším na intervalové úrovni (tzv. kardinální proměnná)
4. **Nezávislost:** tímto předpokladem se myslí skutečnost, že data měřená na jednom subjektu nejsou závislá na jiném subjektu. Závislá data vznikají např. tak, že z výzkumného důvodu musíme měření na našich subjektech po nějakém čase opakovat, např. proto, abychom zjistili účinek nějaké intervence.

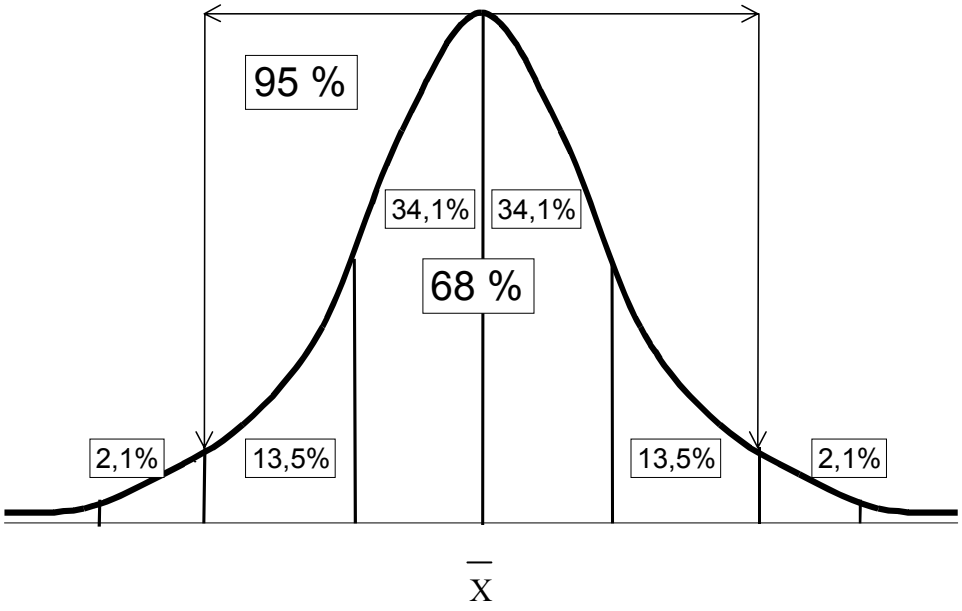
[1] David de Vaus upozorňuje, že výraz „normální" je poněkud zavádějící – zvláště v sociálních vědách, kde mnoho proměnných je rozloženo jiným způsobem (viz jeho pro studenty vynikající knížku Analyzing Social Science data, Sage, London vydanou v roce 2002).

Předpoklad č. 3 a 4 (intervalová data a nezávislost měření) se netestuje, neexistuje totiž na ně žádná metoda, o naplnění tohoto předpokladu musíme v konkrétních případech rozhodnout na základě našich znalostí. Předpoklad 2 se testuje odlišnými způsoby v závislosti na použité statistické proceduře. Na normalitu rozložení existují přesné postupy, které si nyní předvedeme.

Normální rozložení má – mimo nádherného a ladného symetrického tvaru – několik pěkných vlastností: předně, je přesně určena střední hodnotou a směrodatnou odchylkou (viz obr. 3.1). V normálním rozložení má průměr, medián i modus stejnou hodnotu. Většina hodnot se soustřeďuje kolem průměru. Navíc platí, že do čtyř sigma (σ = sigma je symbol pro směrodatnou odchylku), tedy do dvou směrodatných odchylek na každou stranu od průměru, spadne většina pozorovaných hodnot, přesně 95,34 %. Do šesti sigma pak padne přesně 99,7 % pozorovaných hodnot (tedy v rozsahu +3 a -3 směrodatných odchylek). Do jedné směrodatné odchylky na každou stranu spadne 68,26 % případů.

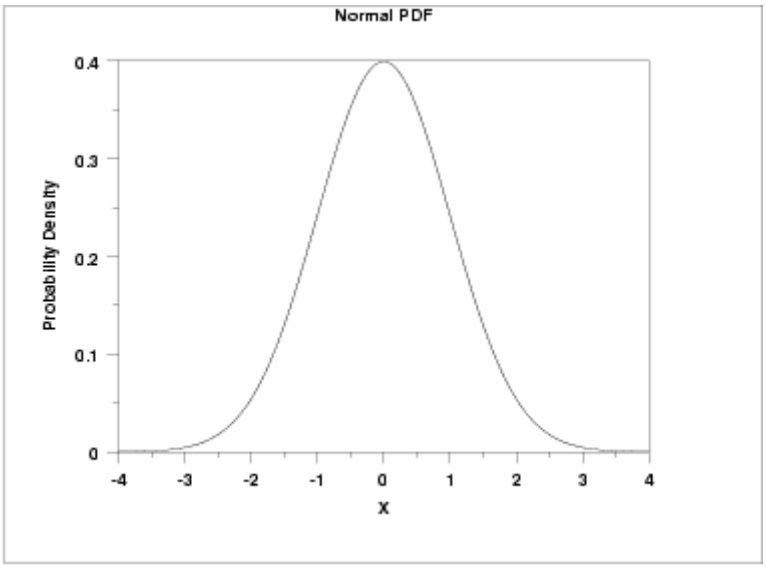
Obr. 3.1: Křivka normálního rozložení a její základní charakteristiky (σ)

Pravidlo šesti sigma: do tří směrodatných odchylek na každou stranu od průměru leží celkem 99,5 % případů.



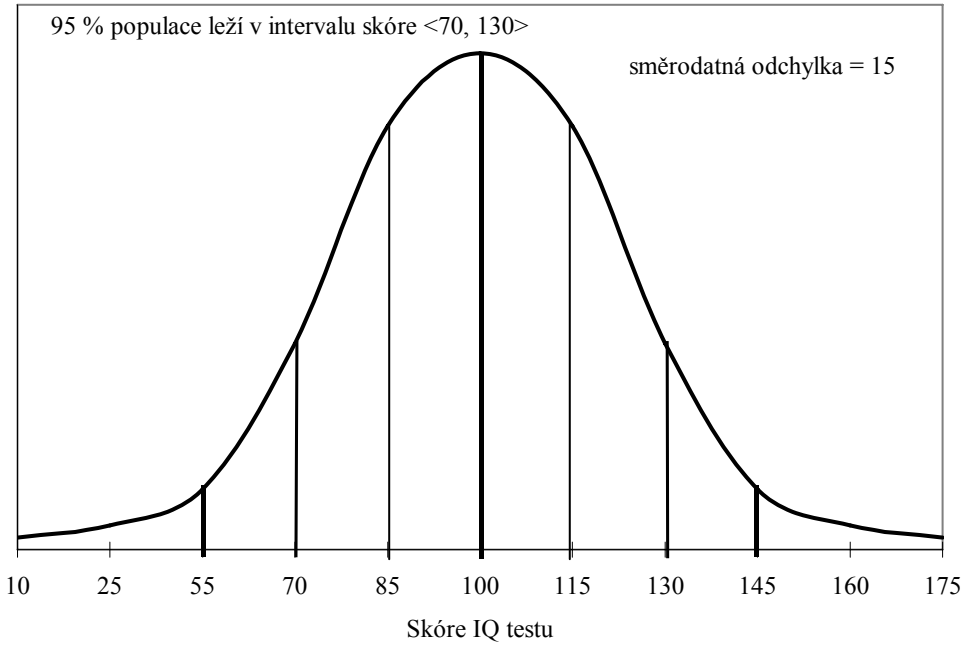
Speciálním případem normálního rozložení je standardizované (nebo také normované) normální rozložení. Jeho vlastností je, že průměr má hodnotu 0 a směrodatnou odchylku 1, jak ukazuje obr 3.1a.

Obr. 3.1a: Standardizované normální rozložení



Jelikož v sociologii pracujeme převážně s daty výběrového souboru, musíme se zajímat nejenom o to, zdali jsou normálně rozloženy charakteristiky výběrového souboru, ale také, zdali toto normální rozložení můžeme očekávat i v souboru základním.

Obr. 3.2: Rozložení skóre v IQ testu

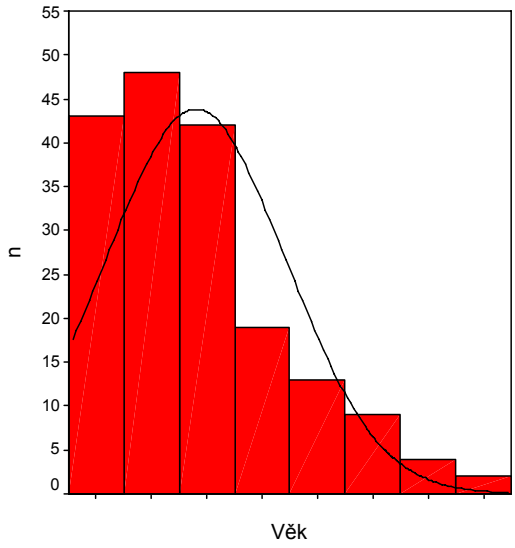


Jak zjistit, zdali je rozložení normální?

1. Nejjednodušším způsobem je nechat si udělat v SPSS histogram rozložení dané proměnné, do něhož vložíme křivku normálního rozložení.

ANALYZE – DESCRIPTIVE STATISTICS – FREQUENCIES (odstraňte požadavek na tabulku frekvencí vlevo dole tak že odklinete zaškrtnuté políčko v *display frequency tables*) – *CHARTS – HISTOGRAMS* (*with normal curve*): vznikne obr. 3.3 (rozložení proměnné věk):

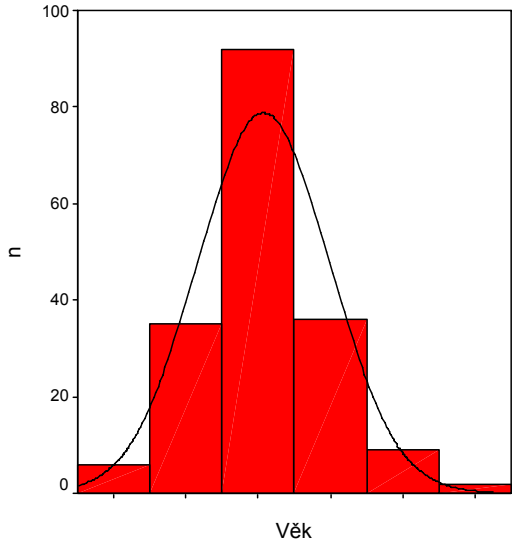
Obr. 3.3: Rozložení proměnné věk



Okometrická analýza obr. 3.3 naznačuje, že rozložení se vychyluje od normálního. Otázkou v takové situaci pak vždy je, zdali odchylka od normálu je natolik malá, abychom dané rozložení mohli považovat alespoň za přibližně normální a mohli tak naplnit předpoklad následných statistických procedur.

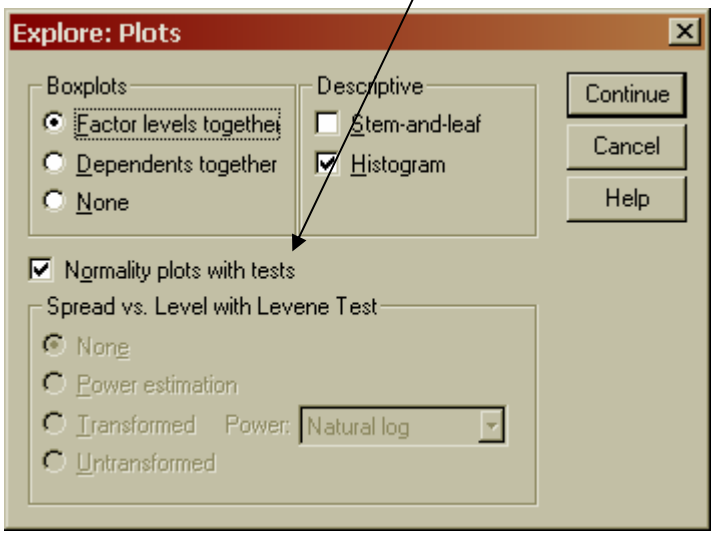
2. Přesnějším způsobem, jak otestovat symetričnost našeho rozložení, je prozkoumáním jeho **šikmosti** a **špičatosti**. Šikmost a špičatost jsou dvě statistiky, které sumarizují tvar rozložení a ukazují, do jaké míry se symetrická křivka odlišuje od svého ideálního symetrického tvaru. SPSS tyto statistiky vypočítává buď v proceduře *Descriptive Statistics – Frequencies* (nebo *Explore*). Normální rozložení má hodnotu šikmosti 0 a špičatosti rovněž 0. Šikmost, které má hodnotu vyšší než 1 (v absolutní hodnotě, neboť šikmost nabývá kladných hodnot, pokud je vrchol křivky posunut doleva a záporných hodnot, pokud je vpravo) indikuje, že rozložení je asymetrické, a tudíž se odchyluje on normálního rozložení. Špičatost rozložení, jehož ukázka je znázorněna na obr. 3.4, nabývá rovněž kladných a záporných hodnot. Záporná hodnota indikuje ploché rozložení, kladná hodnota signalizuje úzké rozložení s protaženým vrcholem.

Obr. 3.4: Ukázka kladně špičatého rozložení



Pro hodnoty šikmosti a špičatosti platí, že pokud se blíží 0, je proměnná normálně (symetricky) rozložena kolem průměru. Ale kdy si můžeme být jisti, že odchylka od nuly je již tak velká, že musíme považovat naše rozložení ze vychýlené? Jedním ze způsobů je hodnoty šikmosti a špičatosti standardizovat a vytvořit z nich tzv. **z-skóry**. Z-skór vzniká tak, že hodnoty znaku odečteme od průměru a výsledek podělíme směrodatnou odchylkou. Jelikož v případě šikmosti i špičatosti má hodnota v případě symetrického (normálního rozložení) velikost 0, nemusíme nic odečítat a můžeme v případě šikmosti její hodnotu přímo podělit její směrodatnou odchylkou (tu vypočítává SPSS). V případě špičatosti výsledek ještě odmocníme. Pokud je vypočtený výsledek vyšší než 2 (přesně řečeno 1,96, ale s tím si nemusíme dělat starosti a můžeme si klidně pamatovat hodnotu 2), můžeme si být jisti, že rozložení naší proměnné je výrazně zešikmeno. Máme-li malý soubor, kritériem pro symetrii je 2,5 (2,58). V případě „velmi velkých souborů", jak zdůrazňuje s vykřičníkem Andy Field, „žádné kritérium by nemělo být aplikováno!" (Field 2002:41).

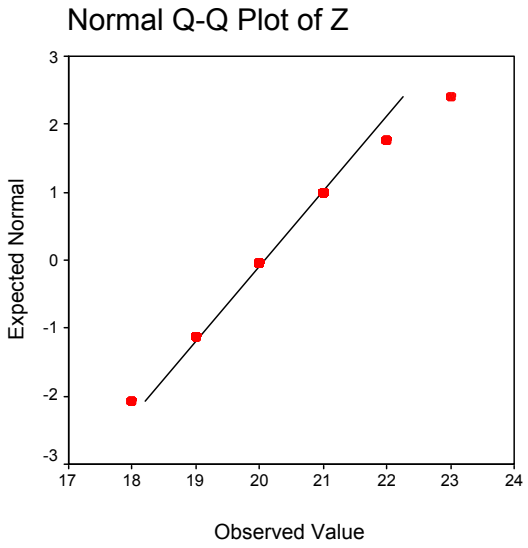
3. Dalším ze způsobů, jak testovat normalitu rozložení, je použití **Kolmogorova-Smirnova testu**. Tento test statisticky hodnotí, zdali je rozdíl mezi pozorovaným rozložením a teoretickým normálním rozložení natolik malý, že jej můžeme připsat náhodě, to je výběrové chybě.[2] Pokud je ovšem tato diference větší, pak naše pozorované rozložení není normální. Pro test normality rozložení tedy vycházíme z nulové hypotézy, že rozdíl mezi pozorovaným rozložením a rozložením teoretickým bude nulový, žádný (o nulové hypotéze se dozvíte v dalších lekcích). Pro aplikaci K-S testu to znamená, že pokud vypočtená signifikance (Sig.) bude menší než 0,05, není naše rozložení normální. Musíme zde ale poznamenat, že v případě velkých souborů i malá odchylka pozorovaného rozložení od rozložení teoretického bude vycházet statisticky významná, takže je potřeba opět nad rozložením uvažovat.
4. K tomu nám může dopomoci inspekce tzv. grafů normální pravděpodobnosti (*normal probability plots*), kterým se v SPSS říká **Normal Q-Q Plots** a **Detrended Normal Q-Q Plots.** Oba získáme v proceduře *Explore – Plots – Normality plots with test*



Ukázku prvního z obou jmenovaných grafů přináší obrázek. 3.5. V něm je každá pozorovaná (naše empiricky zaznamenaná) hodnota (horizontální osa) vynesena proti hodnotě očekávané z normálního rozložení (vertikální osa). Pokud body grafu (tečky) vytvářejí přímku nebo jsou alespoň kolem přímky přibližně rozloženy, naznačuje to normální rozložení. Což je případ obr. 3.5.

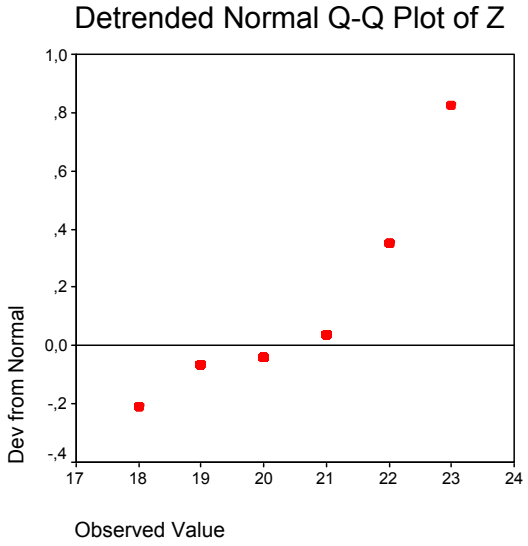
[2] V případě, že počet případů je menší než 50, SPSS tiskne automaticky v tabulce K-S testu také Shapir-Wilkův test, který je v takové situaci vhodnější.

Obr. 3.5: Graf Q-Q proměnné Z



Příklad druhého grafu (*Detrended Normal Q-Q Plot*) je na obr. 3.6. Zde by body, pokud má být rozložení považováno za normální, neměly vytvářet žádné shluky a většina z nich by měla být blízko přímky. Což v našem případě nenastalo.

Obr. 3. 6: Příklad *Detrended Normal Q-Q* grafu.



Co dělat, když zjistíme, že rozložení není normální?

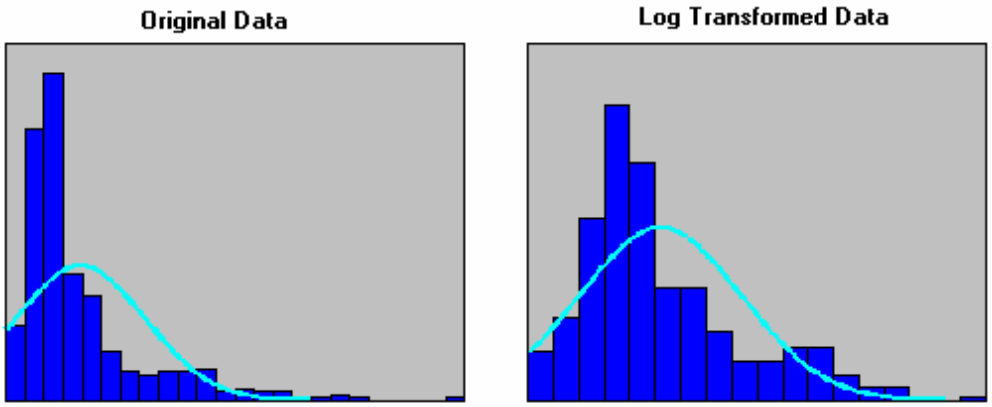
V situaci, kdy zjistíme, že naše proměnná (nebo proměnné), kterou (které) chceme statisticky analyzovat, máme, jak napovídá de Vauss (2002), tři možnosti.

1. **Použít některý z postupů neparametrické statistiky**. To jsou postupy, které nevyžadují, aby analyzovaná proměnná byla normálně rozložena a my si je v naší příručce postupně ukážeme v příslušných kapitolách.

2. **Transformovat statisticky** distribuci (rozložení) naší proměnné, např. tak, že proměnnou logaritmujeme, nebo ji odmocníme, umocníme na druhou případně na třetí apod. Příklad logaritmické transformace uvádí SPSS:

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



Hodnoty o prodeji výrobku byly logaritmovány (viz druhý, tučně orámovaný řádek v tabulce) a původní zešikmené rozložení (obrázek vlevo nazvaný *Original data*), které mělo hodnotu šikmosti výrazně vyšší než l (2,112), což indikuje odchylku od normality, se změnilo na rozložení s menším zešikmením (šikmost se přiblížila 1), což také naznačuje druhý obrázek (*Log transformed data*).

3. **Nebudeme si odchylky od normality všímat** a klidně použijeme statistiku parametrickou. Pro to, že si to můžeme dovolit, existují dva dobré důvody:

a) statistikové postupně ukázali, že porušení požadavku na normalitu nemá tak závažné následky 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.

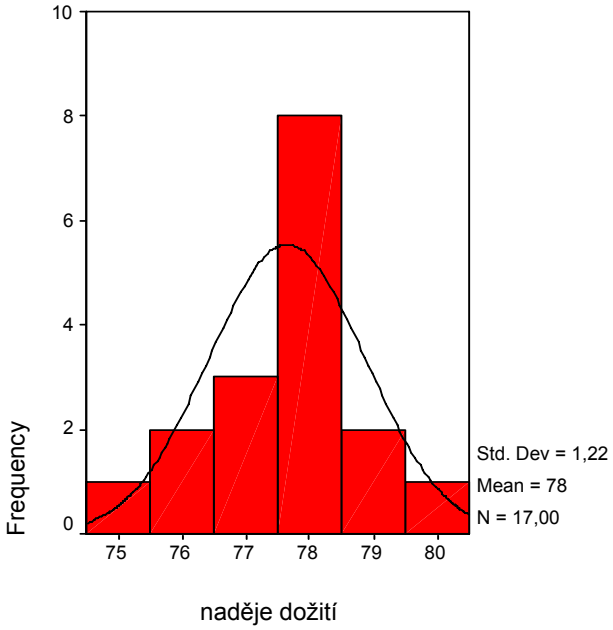
b) ve statistice platí centrální limitní věta (teorém), která stanovuje velmi důležitý princip: se vzrůstající velikostí výběrového souboru se jeho rozložení blíží rozložení normálnímu (podrobněji si rozebereme níže). Což v praxi znamená, že pokud rozložení naší analyzované proměnné není normální, ale pokud náš výběrový soubor je dostatečně velký (rozuměj větší než 100), je možné využívat i statistických postupů, které předpokládají normální rozložení. Hezkou grafickou simulaci tohoto principu je možné nalézt na http://www.statsoft.com/textbook/graphics/an_sampl.gif/ [3]

Příklad P3.1: V našem demografickém souboru dmg_file.sav máme údaj o naději dožití. Podle všeho by tento údaj měl být normálně rozložen. Ale jelikož z demografie víme, že naděje dožití se v bývalých komunistických zemích vyvíjela jinak než v zemích západních, udělejme si analýzu pouze pro země západní.

Řešení:
Nejdříve si vybereme podsoubor západních zemí. V SPSS je na to procedura *Data – Select Cases*, kterou se naučíme v lekci 5. V podsouboru západních zemí si pak necháme udělat histogram rozložení s proloženou normální křivkou. Výsledek je na obr. 3.7.

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

Obr. 3.7: Rozložení naděje dožití pro obě pohlaví v západních evropských zemích v roce 1999.



Již pouhá okometrická analýza obr. 3.7 naznačuje, že rozložení by mohlo být z hlediska zešikmenosti v pořádku, ale že bude pravděpodobně vychýlené z hlediska špičatosti.

K testování normality použijme nám známých prostředků.
1. Testujeme, zdali se rozložení podstatně odlišuje od normality z hlediska **šikmosti** a **špičatosti**. K tomu potřebujeme výpočet šikmosti a špičatosti a jeho směrodatnou chybu. Tyto údaje nám poskytne procedura *Explore*, kterou již známe (*Analyze – Descriptive Statistics – Explore*). Jsou tabelovány v tabulce 3.1.

Tab. 3.1: Vypočtené charakteristiky proměnné naděje dožití

Descriptives

			Statistic	Std. Error
LIFE_EXP nadeje dožití	Mean		77,65	,296
	95% Confidence Interval for Mean	Lower Bound	77,02	
		Upper Bound	78,28	
	5% Trimmed Mean		77,66	
	Median		78,00	
	Variance		1,493	
	Std. Deviation		1,222	
	Minimum		75	
	Maximum		80	
	Range		5	
	Interquartile Range		1,00	
	Skewness		-,387	,550
	Kurtosis		,505	1,063

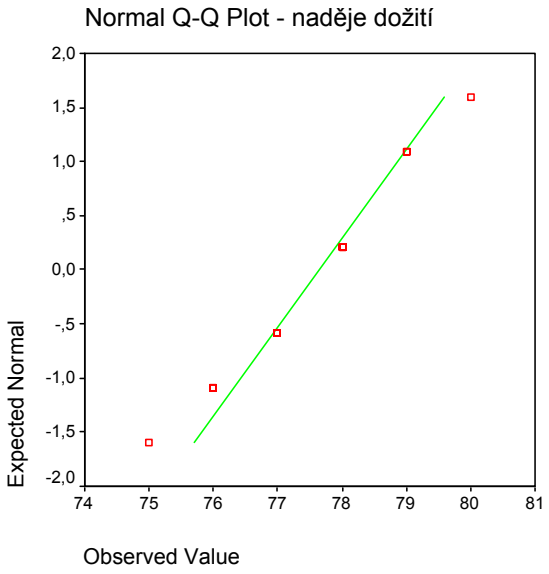
Jak šikmost, tak i špičatost mají hodnoty poměrně nízké, odchylka od normality by tedy nemusela být velká. Podělme hodnoty šikmosti (*Skewness*) a špičatosti (*Kurtosis*) jejich směrodatnými chybami (*Std. Error*) a vypočítejme tak z-skóry. V našem případě, jak vidíme z tabulky 3.1, je šikmost -0,387 a její směrodatná chyba 0,550. Z-skóre je tedy: -0,387/0,550 = -0,70. Z-skóre špičatosti je $\sqrt{0{,}505/1{,}063}$ = 0,69. I tyto hodnoty naznačují, že odchylka od normality není z hlediska šikmosti a špičatosti příliš velká.

Testujme rozložení dále, nyní graficky. Zadání pro Q-Q graf je:

Analyze – Descriptive Statistics – Explore — Plots — Normality Plots with Tests

a jeho výsledek je na obr. 3.8. Normalita rozložení je podle tohoto výsledku na vážkách, body grafu nejsou v některých případech příliš blízko přímce.

Obr. 3.8: Q-Q graf naděje dožití v západních zemích



Součástí výstupu výše uvedené procedury je i následující tabulka Kolmogorova-Smirnova testu normality, výpočet jsme získali současně se zadáním grafického testu normality. **Kolmogorov-Smirnovův test** testuje nulovou hypotézu, že data pocházejí z normálního rozložení. Pokud náš soubor obsahuje méně než 50 případů (jednotek), v tabulce se vytiskne i Shapiro-Wilkův test, který je v daném případě adekvátnější. Výsledek testu je uveden v tab. 3.2.

Tab. 3.2: Test normality rozložení

Tests of Normality

	Kolmogorov-Smirnov [a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
LIFE_EXP naděje dožití	,261	17	,003	,921	17	,152

a. Lilliefors Significance Correction

Jelikož v našem souboru západních zemí bylo pouze 17 případů, použijeme S-W testu. Vypočtená signifikance (ve sloupci *Sig.*) je vyšší než 0,05, takže můžeme naše rozložení považovat za normální.

* * *

Centrální limitní teorém

Jedním z důležitých principů teorie pravděpodobnosti, která tvoří podstatnou část statistické teorie, je centrální limitní teorém. Ten říká, že když provedeme mnoho výběrů o určité velikosti[4] založených na pravděpodobnostním principu (pokud jste již zapomněli, co to znamená, a chcete si stručně a rychle obnovit vaše vědomosti, pročtěte si část 2. kapitoly, str. 37–59 v knize Jana Hendla *Přehled statistických metod zpracování dat*. Portál, Praha, 2004), pak rozložení (distribuce) výběrových průměrů se přiblíží normálnímu rozložení a celkový průměr těchto průměrů se bude podobat průměru v populaci. A to nezávisle na tom, jak jsou hodnoty proměnné rozloženy v populaci.

Má to svou logiku. Budeme-li z populace (např.z ekonomicky aktivní populace ČR) dělat nové a nové pravděpodobnostní výběry, a budeme-li na nich měřit např. hrubý měsíční příjem jednotlivce, abychom zjistili, jaké je rozložení této charakteristiky v ČR, pak rozložení průměrných hrubých měsíčních příjmů začne postupně tvořit normální rozložení. Průměr z těchto průměrů se pak bude velmi podobat skutečnému průměru v celé populaci (který samozřejmě neznáme). To znamená, že hodnoty těchto výběrových průměrů se budou stále více přimykat k populačnímu průměru a původně dlouhé konce rozložení se začnou zkracovat. Směrodatná odchylka tohoto rozložení průměrů se nazývá směrodatnou chybou průměru (*standard error of the mean*). Má tu vlastnost, že se zvyšující se velikostí výběru se snižuje. Což znamená, že značná velikost pravděpodobnostního výběru zaručuje dobrou možnost zobecňovat výsledky zjištěné ve výběrovém souboru na soubor základní (populaci).

Vztah mezi velikostí výběrového souboru a výběrovou chybou není lineární, neboť od určité velikosti výběrového souboru se velikost výběrové chyby již dále nesnižuje. Z tohoto důvodu platí, že soubory větší než, řekněme, 1500 jednotek již nemohou přinést vyšší přesnost zobecňovaných výsledků. Praktické hledisko zde však ale stále platí: pokud víme, že budeme muset z výzkumných důvodů v analýzách pracovat s podsoubory (např. v případě příjmů bychom rádi srovnávali podsoubor vysokoškolsky vzdělaných muž a žen, kteří pracují ve státních službách se souborem stejně vzdělaných osob v privátním sektoru), musíme velikost výběrové souboru rozšířit. To je jeden z důvodů, proč v sociologických výzkumech klidně naleznete výběry o velikost např. čtyř tisíc osob.

Příklad P3.2: Naznačme si platnost centrální limitní věty (*central limit theorem*) která říká: ať je rozdělení základního souboru jakékoliv, rozdělení střední hodnoty výběrového souboru bude vždy normální, jestliže rozsah výběrového souboru dosáhne alespoň jisté minimální velikosti, tedy alespoň 30 (viz Helmut Swoboda, *Moderní statistika*, str. 153).

Jak známo, z populace je možné teoreticky udělat nekonečné množství výběrových souborů. Představme si nyní, že soubor 1 908 respondentů, kteří odpovídali na naše otázky ve výzkumu o hodnotách v ČR v roce 1999, je naším základním souborem. Že jsme tedy provedli vyčerpávající zjišťování (de facto census) v nějakém malém státečku, který má 1908 obyvatel. Z tohoto základního souboru můžeme prostřednictvím SPSS udělat celou řadu náhodných výběrových souborů o velikosti, řekněme 20 % z celého souboru. Procedura k tomu je následující:

Data — Select Cases — Random Sample of Cases — Sample — Approximately … % of all cases

My těch náhodných výběrů uděláme pouze 10 a budeme sledovat, jak se mění hodnota průměru proměnné q33 *Jak důležitý je Bůh ve Vašem životě?* (soubor EVS99_cvicny). Výsledky uvádí tabulka 3.3.

[4] Určení velikosti tohoto výběrů v učebnicích kolísá. Někteří autoři tvrdí, že může mít jenom 30 jednotek, jiní se domnívají, že by to měla být přinejmenším stovka. Požadovaná velikost výběru závisí ovšem nejenom na statistické teorii, ale také na praktických hlediscích. Chceme-li provádět třídění vyšších stupňů a většina našich proměnných je nominálních nebo ordinálních, což je častý případ např. v sociologii, musíme získávat mnohem větší výběrové soubory. Proto N > 1000 není v sociologických výzkumech ničím neobvyklým.

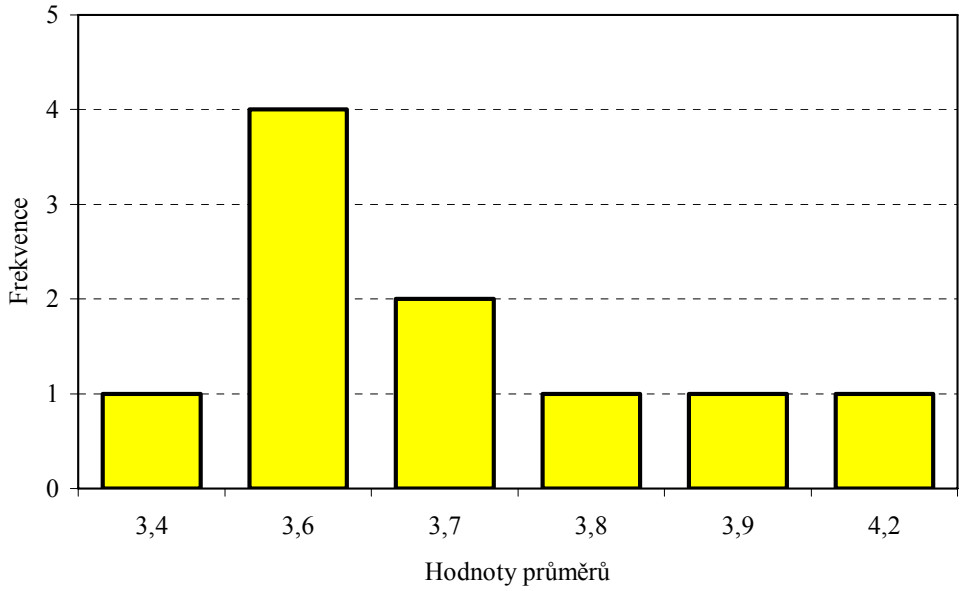
Připomeňme si, jak již víme z lekce druhé, že hodnota „skutečného“ průměru (to je průměru populace našeho imaginárního ministátu) byla 3,63 a jeho směrodatná odchylka 3,06.

Tab. 3.3: Různé náhodné výběry a měnící se hodnoty průměrů proměnné q33

Výběr	Průměr	Směrod. odchylka	N
1.	4,15	3,28	371
2.	3,58	3,05	375
3.	3,75	3,02	361
4.	3,85	3,12	406
5.	3,41	2,92	368
6.	3,74	3,11	371
7.	3,64	3,08	373
8.	3,58	2,97	388
9.	3,71	3,01	362
10.	3,56	3,05	367

Když z hodnot průměrů a jejich frekvence uděláme příslušný graf (do grafu jsme zanesli hodnoty průměrů zaokrouhlené na jedno desetinné místo), vidíme, že rozložení průměrů nabývá tvaru, které začíná připomínat normální rozložení (viz obr. 3.9).

Obr. 3.9: Rozložení hodnot průměrů proměnné q33 z deseti náhodně vybraných vzorků



Když navíc vypočteme z průměrů jednotlivých výběrů celkový průměr, dostaneme hodnotu 3,70, která není příliš vzdálena od průměru 3,63.

* * *

Z skóry (standardizovaná směrodatná odchylka)

V některých úlohách potřebujeme porovnat, jak jsou vzdáleny jednotlivé hodnoty od průměru. Předpokládejme, že v testu ze statistické analýzy dat někdo získal 77 bodů a jiný 66 bodů. Když víme, že průměrný výsledek v testu byl 70 bodů, můžeme vypočítat, jaká je pozice těchto dvou výsledků vzhledem k celkovému rozložení hodnot výsledků testu. Nástrojem k tomu jsou tzv. Z-skóry. Potřebujeme k tomu znát kromě průměru navíc směrodatnou odchylku, neboť vzorec pro výpočet této charakteristiky je:

Z-skór = (hodnota znaku – průměr) / směrodatná odchylka.

Víme-li, že směrodatná odchylka od průměrného bodového skóre v testu z analýzy dat byla 5, pak výsledek studenta, jenž získal 66 bodů jej umisťuje do vzdálenosti -0,8 směrodatné odchylky od průměru, neboť (66–70)/5 = -0,8. Výsledek 77 bodů znamená +1,4 směrodatné odchylky od průměru, neboť (77–70)/5 = 1,4.

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. V SPSS je možno Z-skóry nejen vypočítat, ale uložit jako novou proměnnou a dále s ní pracovat.

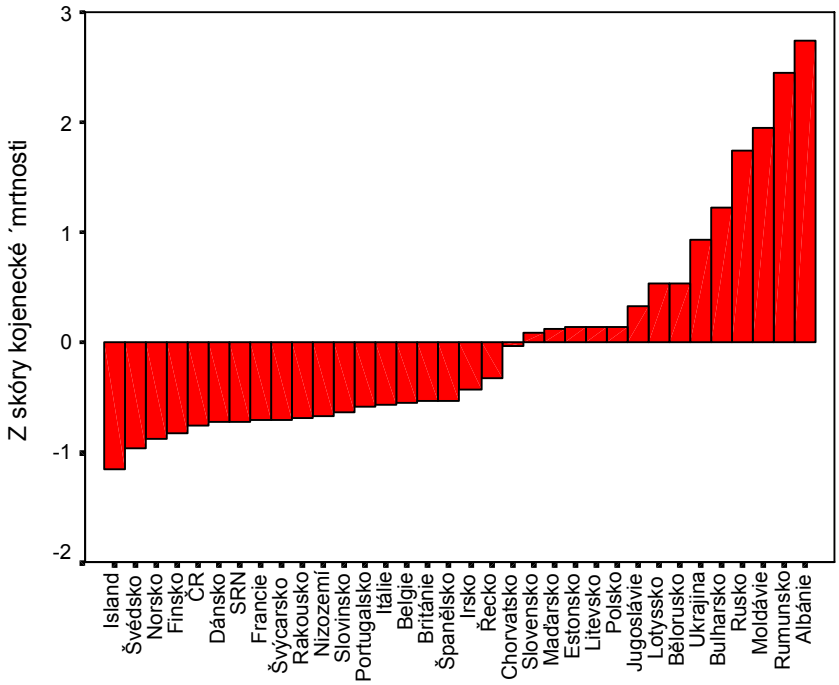
Z-skóry jsou de facto standardizované hodnoty, neboť převádějí původní proměnné, měřené v různých jednotkách, do jednotné metriky: vzdálenosti od průměru. Touto standardizací tak získáváme možnost srovnávat zdánlivě nesrovnatelné, např. i ony okřídlené hrušky a jablka. Zjistíme např., že respondent má 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 rozložení 3.1 a uvědomíte si, že do plochy nad dvě směrodatné odchylky spadá jen 2,14 % případů s nejvyššími hodnotami nad průměrem, je to příjem velmi vysoký). Ve vzdělání je však pod průměrem. 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.

Příklad P3.3: A nyní příklad z reálných dat. Z demografické statistiky máme údaje o kojenecké úmrtnosti (viz soubor *dmg-data.sav*, proměnná *kojen_um*). Když si prostřednictvím procedury

Analyze–Descriptive statistics–Descriptives–Savestandardized values as variables

necháme uložit z-skóry této proměnné, uloží se nám jako nová proměnná na konec matice s názvem, který opakuje název původní proměnné s tím, že před něj předřadí písmeno z. Z proměnné *kojen_u* se tak stane proměnná ***z****kojen_u*. Nechejme si nyní celý soubor utřídit pomocí procedury *Data – Sort cases – Sort by (zkojen_u)*, 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 3.10.

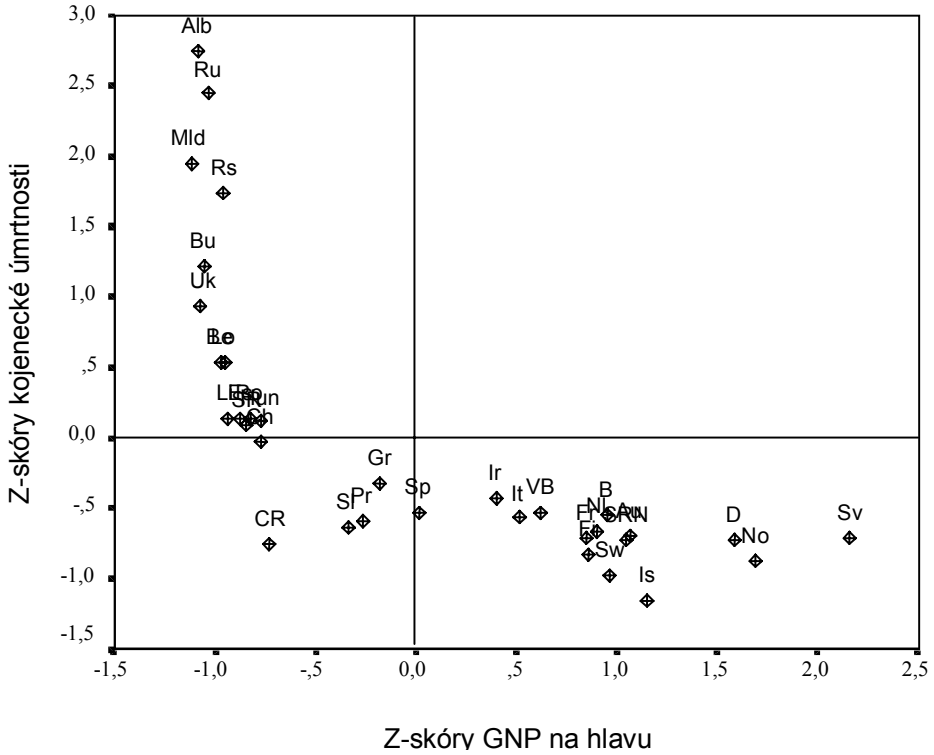
Obr. 3.10: Pořadí evropských zemí podle z-skórů kojenecké úmrtnosti v roce 1999.



Z dat víme, že průměr kojenecké úmrtnosti byl v Evropě v roce 1999 8,34 zemřelých dětí do jednoho roku na 1000 živě narozených a směrodatná odchylka byla 4,97. Z obrázku je pak patrné, jak mnoho se jednotlivé evropské země v tomto ukazateli odlišují. Na průměrné hodnotě je Chorvatsko, hodnoty nižší než průměr, mají všechny západoevropské země, k nimž se z bývalých komunistických zemí řadilo v roce 1999 pouze ČR (a zdůrazněme, že naše kojenecká úmrtnost je jedna z nejnižších na světě i v současnosti) a Slovinsko.

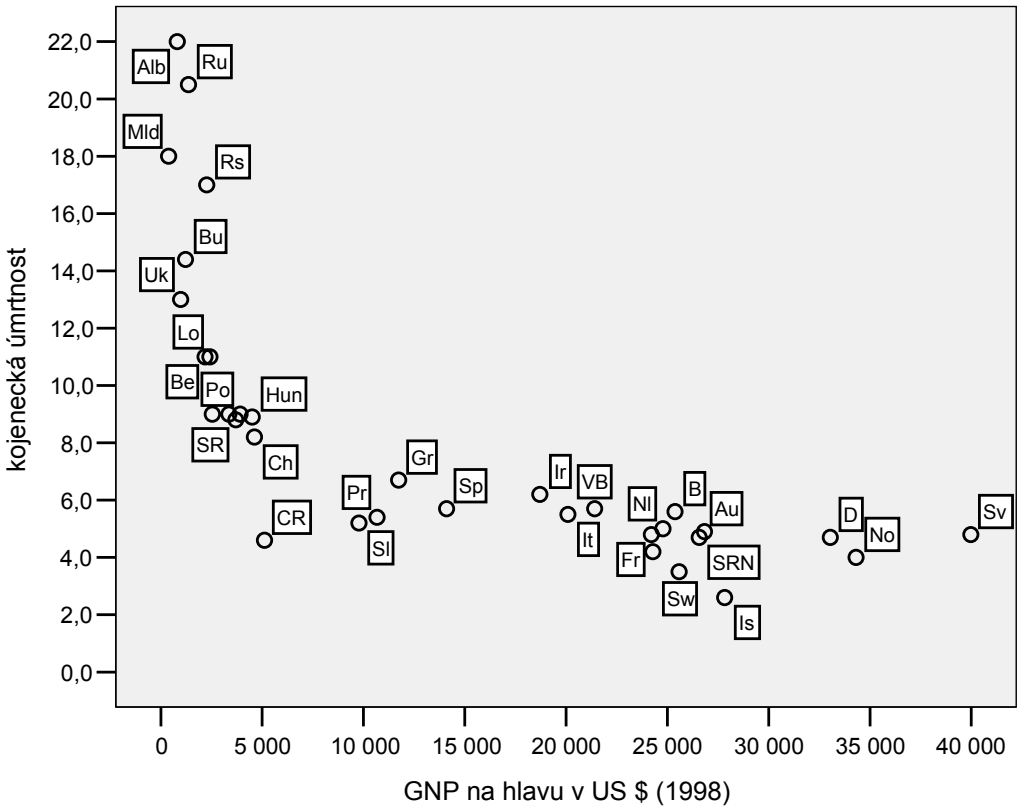
Nechejme si vypočítat z-skóry ještě pro další proměnnou tohoto souboru, a to pro proměnnou hrubý národní produkt na hlavu (*gnp_head*), který je uváděn v US dolarech na hlavu (tato data zachycují situaci v roce 1998). Když tuto novou proměnnou (*zgnp_hea*) umístíme do *scatter* grafu spolu se z-skóry kojenecké úmrtnosti, získáme velmi zajímavý obrázek (viz obr. 3.11). Říká nám, že země s nadprůměrným hrubým národním produktem – v Evropě byl průměr GNP na hlavu 13 899 US dolarů a směrodatná odchylka byla 12 086 – mají také obvykle podprůměrnou kojeneckou úmrtnost a naopak země s podprůměrným GNP (chudé země) mají obvykle vysokou kojeneckou úmrtnost.

Obr. 3.11: Evropské země podle z-skórů hrubého národního příjmu na hlavu (GNP) a kojenecké úmrtnosti v roce 1999



Podobný obrázek (viz obr. 3.12)ovšem získáme i tehdy, když nebudeme pracovat se z-skóry, ale s reálnými, to je netransformovanými proměnnými, to je s *kojen_um* a *gnp_head*. Zásadní rozdíl v obou obrázcích spočívá v tom, že v obr 3.11 se z-skórovými proměnnými máme dobrou představu, jak jsou které země vzdáleny od průměru. Z hlediska kojenecké úmrtnosti a hrubého národního produktu na hlavu bylo v roce 1999 průměrnou zemí Řecko (Gr). Albánie (Alb) byla zemí odlehlou, jež měla jak v kojenecké úmrtnosti, tak v GNP nejhorší výsledky, výrazně vzdáleny od průměru. Pozice České republiky je zajímavá. V GNP byla asi 0,7 směrodatné odchylky pod průměrem, což nás nijak netěší, v kojenecké úmrtnosti byla rovněž 0,7 směrodatné odchylky pod průměrem, z čehož však již máme radost.

Obr. 3.12: Evropské země podle hrubého národního příjmu na hlavu (GNP) a kojenecké úmrtnosti v roce 1999



Oba obrázky ovšem naznačují, že Evropa byla v roce 1999 v kontextu GNP a kojenecké úmrtnosti značně heterogenním celkem. Jak by asi vypadala analýza provedená zvlášť v podsouboru tzv. západoevropských a tzv. východoevropských zemí?