

Metodologie pedagogického výzkumu I

- kurz pro první ročník magisterského studia oboru pedagogiky, PedF UK
- rozsah kurzu: 1/1
- výuka probíhá blokově ve dnech:
 - pátek 6. listopadu, 14:25 až 20:25, R305
 - sobota 7. listopadu, 9:00 až 15:00, R305
 - neděle 8. listopadu, 9:00 až 15:00, R305
- **zakočení kurzu:** zápočet a zkouška
- **požadavky ke zkoušce:** ústní zkouška a seminární práce (výzkumná zpráva o realizaci vlastního výzkumu)
 - výsledek zkoušky: 60 % známky tvoří ústní zkouška, 40 % známky tvoří seminární práce;
 - ústní zkouška i seminární práce budou hodnoceny na škále 1 až 10
 - výsledná u zkoušky 1 bude odpovídat 8.5-10, známka 2 bude odpovídat 7-8.4, známka 3 bude odpovídat 5-6.9
 - **ústní zkouška** z metod pedagogického výzkumu a statistiky využívané v pedagogickém výzkumu
 - zkouška založena na odpřednášené látce během kurzu a diskutované literatuře během kurzu, studijní materiály jsou dostupné na internetové stránce www.zla-ryba.cz/hanicka/metodologie
 - v části metody pedagogického výzkumu si vylosujete dvě otázky, které budeme následně po vaší krátké přípravě diskutovat
 - v části statistika budete na počítači s využitím statistického softwaru Gretl vyhodnocovat požadovaným způsobem předložená

data (můžete si přinést vlastní notebook s jiným statistickým softwarem, který umíte ovládat, a vypracovat předloženou úlohu v něm)

- **seminární práce:** Výstižně popsat realizaci vlastního výzkumu!
 - Ve vlastním výzkumu je možné využít diskutovaných metod během kurzu, popř. jiných relevantních metod, které odpovídají povaze zkoumaného problému.
 - při psaní vaší výzkumné zprávy vyjděte z šesti kroků, které popisují jednotlivé kroky v provedení výzkumu (popis těchto šesti kroků bude diskutován na přednášce, jedná se konkrétně o obrázek popisující fáze výzkumu
metodologie_scanner_tables_graphs/ Figure-1-1-reasearch-steps.jpg
na následujících slajdech, tento obrázek stejně jako další studijní materiály si lze stáhnout na
www.zla-ryba.cz/hanicka/metodologie)
 - citace literatury může odpovídat požadavkům na citaci literatury pro časopis Pedagogická orientace :
http://www.ped.muni.cz/pedor/index.php?option=com_content&view=article&id=117&Itemid=96
 - **termíny k ústní zkoušce:**
8. prosince 2009 na katedře pedagogiky v R225, začátek v 13:00
14. prosince 2009 na katedře pedagogiky v R225, začátek v 13:00
 - **termín pro odevzdání seminární práce:**
do 20. ledna 2010, seminární práci poslat emailem, příjem práce vždy potvrdím
- **požadavky k zápočtu:** návrh vlastního výzkumu
 - při psaní návrhu vyjděte z šesti kroků, které popisují jednotlivé kroky v provedení výzkumu, viz obrázek
metodologie_scanner_tables_graphs/ Figure-1-1-reasearch-steps.jpg

- ve svém návrhu popište, jakým způsobem budete realizovat tyto kroky ve svém vlastním výzkumu (5 krok je např. analyzování a interpretování dat, ve svém návrhu můžete napsat, že hodláte využít regresní analýzy pro vysvětlení vztahů mezi vámi zkoumanými proměnnými)
 - termín odevzdání návrhů: 3. prosince 2009
 - návrhy můžu vrátit k přepracování
- **dotazy vždy zasílejte elektronicky na:**
h.vonkova@uvt.nl, h.vonkova@gmail.com, krizule@email.cz
 - **konzultace možné ve dnech:**
6.11.- až 11.11.
7.12. až 15.12.
nutno předem domluvit pomocí emailu
 - výklad **metod** pedagogického výzkumu je založen především na dvou knihách:
 - Gay, L.R., Mills, G.E., Airasian, P. *Educational Research. Competencies for Analysis and Application*. Upper Saddle River, NJ : Pearson Higher Education, 2008.
 - Chrástka, M. *Metody pedagogického výzkumu*. Praha : Grada, 2007.
 - výklad **statistiky** je založen především na knize:
 - Hinkle, D.E., Wiersma, W., Jurs, S.G. *Applied Statistics for the Behavioral Sciences*. Boston : Houghton Mifflin, 2003.
 - **Další studijní materiály** lze najít na internové adrese www.zla-ryba.cz/hanicka/metodologie
 - naskenované tabulky a obrázky z knih Gay (2008) a Chrástka (2009) v souboru `metodologie_scanner_tables_graphs.zip` (15 jpg souborů)

- příklady dotazníků v souboru metodologie_dotazniky_priklad.zip (4 příklady - dotazník o kázni, manipulaci, PISA dotazník a SHARE dotazník)
 - teoretické a praktické základy pojmového mapování v souboru metodologie_pojmove_mapy.pdf
 - datové soubory používané v příkladech diskutovaných během kurzu v souboru metodologie_data.zip (12 datových souborů, které jsou odděleně uloženy v csv souborech, všechny datové soubory jsou v excelovském souboru data.xls na jednotlivých listech)
- Statistika v pedagogickém výzkumu je v našem kurzu vysvětlována s minimálním použitím vzoreček a s důrazem na konkrétní využití v reálných příkladech.
 - Teorie statistiky je vysvětlena buď pomocí teoretických pouček či pomocí příkladů. Určitá část teorie je vysvětlena pouze a jen v příkladech.
 - K porozumění obsahu (především statistiky) je pro většinu studentů velmi vhodné chodit na přednášky a sledovat výklad.
 - Statistický software, který budeme využívat, se nazývá Gretl. Je to free software (nic nestojí) a lze si ho stáhnout z následující internetové adresy:
<http://gretl.sourceforge.net/win32/>
na první řádce této stránky naleznete soubor gretl-1.8.5.exe, stáhněte (uložte) si ho na svůj počítač. Následně ho otevřete - spustí se tím instalace. Velmi doporučuji si software stáhnout a provést v něm všechny příklady a cvičení, které budeme diskutovat během přednášky!

1 Metody pedagogického výzkumu

- Jednotlivé kroky v empirickém kvalitativním i kvantitativním výzkumu

metodologie_scanner_tables_graphs/ Figure-1-1-reasearch-steps.jpg

- Charakteristika dobře zvoleného výzkumného tématu
metodologie_scanner_tables_graphs/ Table-2-2-research-topic.jpg
- Rozdíly mezi kvalitativním a kvantitativním výzkumem
metodologie_scanner_tables_graphs/ Table-1-1-research-type.jpg
- Typy kvalitativního výzkumu
metodologie_scanner_tables_graphs/Table-1-2-research-qualitative.jpg
- Dotazník - jak formulovat položky 1
metodologie_scanner_tables_graphs/Figure-7-1-questionnaire-example-items.jpg
- Dotazník - jak formulovat položky 2
metodologie_scanner_tables_graphs/Table-7-1-questionnaire-items.jpg
- Typy škál pro měření postojů 1 - Likertova škála
metodologie_scanner_tables_graphs/Scales1-Likert.jpg
- Typy škál pro měření postojů 2 - bipolární škála, hodnotící škála
metodologie_scanner_tables_graphs/Scales2-diferencial-rating.jpg
- Typy měření
metodologie_scanner_tables_graphs/Scales2-diferencial-rating.jpg
- Příklady dotazníků
metodologie_dotazniky_priklad.zip (dotazníky PISA, SHARE, kázeň, manipulace)
- Metody sběru dat ve výzkumném šetření
metodologie_scanner_tables_graphs/Table-7-2-collection-methods.jpg
- Pozorování - příklad standardizovaného pozorování
metodologie_scanner_tables_graphs/Pozorovani1.jpg., Pozorovani2.jpg, Pozorovani3.jpg a Pozorovani4.jpg
- Pojmové mapování
metodologie_pojmove_mapy.pdf

2 Statistika v pedagogickém výzkumu

2.1 Úvod, základní pojmy

- **Populace** zahrnuje všechny členy definované skupiny.
- **Výběr** je podmnožina členů populace.
- **Deskriptivní statistika** je kolekce metod pro klasifikování a sumarizování numerických dat.
- **Inferenční statistika** je kolekce metod, která umožňuje činit závěry o charakteristikách populace na základě příslušných charakteristik příslušného výběru.
- Proces **kódování** zahrnuje přepisování numerických hodnot kategoriálním proměnným. (Zopakuj rozdíly mezi kategoriální, ordinální, intervalovou a poměrovou proměnnou.)
- Data jsou v datovém souboru většinou organizována tak, že každý řádek odpovídá jednomu individu a sloupec obsahuje data for měřenou proměnnou.

2.2 Deskriptivní statistika

2.2.1 Tabulka absolutních, relativních a kumulativních četností

Příklad

Učitel biologie zadal ve své třídě test z biologie, v němž žáci dopadli následujícím způsobem (uvedeny známky z testu):

1,2,3,2,2,5,4,2,2,3,2,1,4,5,4,3,1,1,2,2.

Sestavte tabulku absolutních, relativních a kumulativních četností pro zpřehlednění výsledků žáků z testu.

Řešení

známka	četnosti		
	absolutní	relativní (v %)	kumulativní (v %)
1	4	20	20
2	8	40	60
3	3	15	75
4	3	15	90
5	2	10	100
celkem	20	100	

Cvičení

Sestavte tabulku četností pro následující hodnoty:

0,1,1,2,2,0,1,1,2,0,1,1,2,0,2,2,2,0,2,1,2,1,1,1,1,1,1,1,1

2.2.2 Míry polohy

Míry polohy indikují centrální tendenci naměřených hodnot proměnné.

Průměr

- **Průměr**(mean) vypočítáme ho tak, že všechny hodnoty sečteme a tento součet podělíme počtem hodnot.
- Průměr je nejčastější používanou mírou polohy dat.
- Průměr je velmi ovlivněn extrémními hodnotami, tj. buď extrémně malými či extrémně velkými hodnotami. (Průměr není robustní statistikou.)
- příklad: průměr z hodnot 1, 2, 1, 1, 2, 1, 1 je roven 1.29; průměr z hodnot 1, 2, 1, 1, 2, 1, 1000 je roven 144 → jedna hodnota v datech zcela změnila průměr

- Průměr nemá význam počítat u nominálních a ordinálních proměnných. Využíváme ho u intervalových a poměrových proměnných.

Medián

- **Medián** je bod, pod kterým leží 50 procent hodnot (z toho vyplývá, že nad ním leží také 50 procent hodnot). Medián lze také nazvat 50ti procentním percentilem.
- příklad: urči medián pro skóry 1000, 18, 3, 6, 12, 19, 21
řešení: data nejprve uspořádáme podle velikosti od nejmenší po největší hodnotu 3,6,12,18,19,21,1000 ; prostřední hodnota je 18 (před ní jsou 3 hodnoty, za ní jsou 3 hodnoty), medián je tudíž roven 18
- příklad: urči medián pro skóry 1000, 18, 3, 6, 1, 12, 19, 21
řešení: data nejprve uspořádáme podle velikosti 1, 3, 6, 12, 18, 19, 21, 1000, vzhledem k tomu, že máme lichý počet hodnot, tak medián vypočítáme jako průměr dvou prostředních hodnot 12 a 18. Medián je tedy roven $(12+18)/2=15$
- Medián je oproti průměru robustní statistikou, tj. není citlivý na extrémní hodnoty. Viz první příklad pro medián.
- cvičení: Porovnej průměrný a mediánový plat v České republice. Je průměrný plat nižší, stejný, či vyšší než mediánový plat?
- Medián nemá význam počítat u nominálních a ordinálních proměnných. Využíváme ho u intervalových a poměrových proměnných.

Modus

- **Modus** je nejčastější hodnota v datech.

- příklad: urči modus pro následující data 1,2,1,3,2,7,1000,2,2,6,2
řešení: nejčastěji se vyskytuje hodnota 2, modus je tedy roven 2.
- Modus je robustní statistikou, viz předchozí příklad (extrémní hodnota nemá na modus vliv).
- Modus můžeme určit pro všechny typy proměnných, tj. nominální, ordinální, intervalové i poměrové proměnné.

Minimum a maximum

- Minimum je nejmenší hodnota, maximum je největší hodnota.
- příklad: urči minimum a maximum pro následující data 2,-4,3,-50,20,13,-14,23,-41
řešení: minimum je -50, maximum je 23.
- Minimum i maximum nemá význam počítat u nominálních a ordinálních proměnných. Využíváme je u intervalových a poměrových proměnných.

2.2.3 Míry variability

Míry variability indikují, jak naměřené hodnoty kolísají, tj. jakou mají variabilitu.

Rozptyl, standardní odchylka

- **Rozptyl** je definován jako průměr čtvercových odchylek jednotlivých hodnot od průměrné hodnoty.
- Postup výpočtu rozptylu: Máme-li dané hodnoty, musíme nejdříve spočítat průměr z těchto hodnot. Následně spočítáme rozdíl

naměřených hodnot od vypočítané průměrné hodnoty. Dále každý rozdíl vynásobíme sám sebou (je-li rozdíl roven 3, pak spočítáme $3*3=9$). Z těchto hodnot spočítáme průměr.

- příklad: mějme naměřené hodnoty 1,3,5. Spočítejte rozptyl.
řešení: průměr z naměřených hodnot je roven $(1+3+5)/3=3$
rozdíly hodnot od průměru jsou 1-3,3-3,5-3, tj. -2,0,2
každý rozdíl vynásobíme sám sebou $-2*(-2)$, $0*0$, $2*2$, tj. 4,0,4
průměr z předchozích hodnot 4,0,4 je roven $(4+0+4)/3 = 2.67$
rozptyl je roven 2.67
- Rozptyl je citlivý na extrémní hodnoty.
- cvičení: spočítej rozptyl z hodnot 1,1,1,10
- cvičení: spočítej rozptyl z hodnot 1,1,1,1

Směrodatná odchylka

- Směrodatná odchylka je rovna odmocnině z rozptylu.
- Postup výpočtu: Nejprve spočítáme rozptyl, následně z rozptylu spočítáme druhou odmocninu.
- příklad: mějme naměřené hodnoty 1,3,5. Spočítejte směrodatnou odchylku.
řešení: rozptyl je roven 2.67 (viz předchozí příklad)
druhá odmocnina z 2.67 je rovna $\sqrt{2.67} = 1.63$
směrodatná odchylka je rovna 1.63
- Směrodatná odchylka je oproti rozptylu vyjádřena v původních jednotkách měření, tj. na té samé škále, na které měříme hodnoty proměnné.
- Směrodatná odchylka je citlivá na extrémní hodnoty.

- cvičení: spočítej směrodatnou odchylku z hodnot 1,1,1,10
- cvičení: spočítej směrodatnou odchylku z hodnot 1,1,1,1

Variační rozpětí

- **Variační rozpětí** je rovno rozdílu maxima a minima, k němuž přičteme 1.
- příklad: spočítej variační rozpětí z hodnot -2,3,-10,6,9
řešení: variační rozpětí je rovno $9 - (-10) + 1 = 20$
- cvičení: spočítej variační rozpětí z hodnot -4,9,0,63,5,-50,-31,2

Gretl a datové soubory

- Pro splnění všech následujících příkladů je nutné využít nějaký statistický software. V našich přednáškách využijeme Gretl.
- natáhnutí dat do Gretlu: *File* → *Open data Import* → Zvolte formát, ve kterém máte data uložena (např. .xls pro Excel, .csv pro comma separated soubor)
- Gretl se Vás může při natahování dat zeptat "The imported data have been interpreted as undated (cross-sectional). Do you want to give the data a time-series or panel interpretation?" Ve všech datových souborech, se kterými budeme během hodin pracovat, nejsou data uspořádána ani jako časová řada ani jako panel. Je tedy nutno zvolit odpověď "No".
- všechny datové soubory, které budeme používat, lze najít v excelovském souboru metodologie_data.xls na jednotlivých listech; jednotlivé datové soubory lze najít jako .csv soubory (viz zla-ryba.cz/hanicka/metodologie)

Příklad (data 01_descriptive_normal_IQ.csv)

V datovém souboru jsou hodnoty IQ pro pět set individuů.

1. Sestavte tabulku četností (absolutních, relativních a kumulativních), kde velikost jednoho třídícího intervalu je rovna 5 a minimální hodnota je rovna 50. Určete modus.
2. Sestavte tabulku četností (absolutních, relativních a kumulativních), kde je počet intervalů roven 11.
3. Reprezentujte data graficky pomocí histogramu, v němž velikost jednoho třídícího intervalu je rovna 5 a minimální hodnota je rovna 50.
4. Reprezentujte data graficky pomocí histogramu, v němž je počet intervalů roven 11.
5. Znázorněte data graficky pomocí boxplot. Určete minimum, první kvartil (hodnota, po níž leží 25 % všech hodnot), medián, třetí kvartil (hodnota, pod níž leží 75 % všech hodnot) a maximum.
6. Spočítejte průměr, medián, minimum, maximum, standardní odchylku a roztyl.
7. Zvonovitý tvar histogramu indikuje normální rozložení zkoumané veličiny. Na základě histogramu pro IQ posuďte, zda má tato veličina tendenci být normálně rozložená.

Řešení

1. Gretl: *Variable* → *Frequency distribution* → *Minimum value, left bin* zvol 50 a *Bin width* zvol 5

Frequency distribution for IQ, obs 1-500
 number of bins = 20, mean = 99.3317, sd = 14.679

interval	midpt	frequency	rel.	cum.
< 55.000	52.500	0	0.00%	0.00%
55.000 - 60.000	57.500	3	0.60%	0.60%
60.000 - 65.000	62.500	2	0.40%	1.00%
65.000 - 70.000	67.500	5	1.00%	2.00%
70.000 - 75.000	72.500	6	1.20%	3.20%
75.000 - 80.000	77.500	29	5.80%	9.00% **
80.000 - 85.000	82.500	38	7.60%	16.60% **
85.000 - 90.000	87.500	52	10.40%	27.00% ***
90.000 - 95.000	92.500	62	12.40%	39.40% ****
95.000 - 100.00	97.500	69	13.80%	53.20% ****
100.00 - 105.00	102.50	65	13.00%	66.20% ****
105.00 - 110.00	107.50	52	10.40%	76.60% ***
110.00 - 115.00	112.50	43	8.60%	85.20% ***
115.00 - 120.00	117.50	36	7.20%	92.40% **
120.00 - 125.00	122.50	16	3.20%	95.60% *
125.00 - 130.00	127.50	14	2.80%	98.40% *
130.00 - 135.00	132.50	2	0.40%	98.80%
135.00 - 140.00	137.50	4	0.80%	99.60%
140.00 - 145.00	142.50	1	0.20%	99.80%
>= 145.00	147.50	1	0.20%	100.00%

Modus je roven 97.5 (střední bod=midpoint intervalu, který má největší četnost).

2. Gretl: *Variable* → *Frequency distribution* → *Number of bins* zvol 11

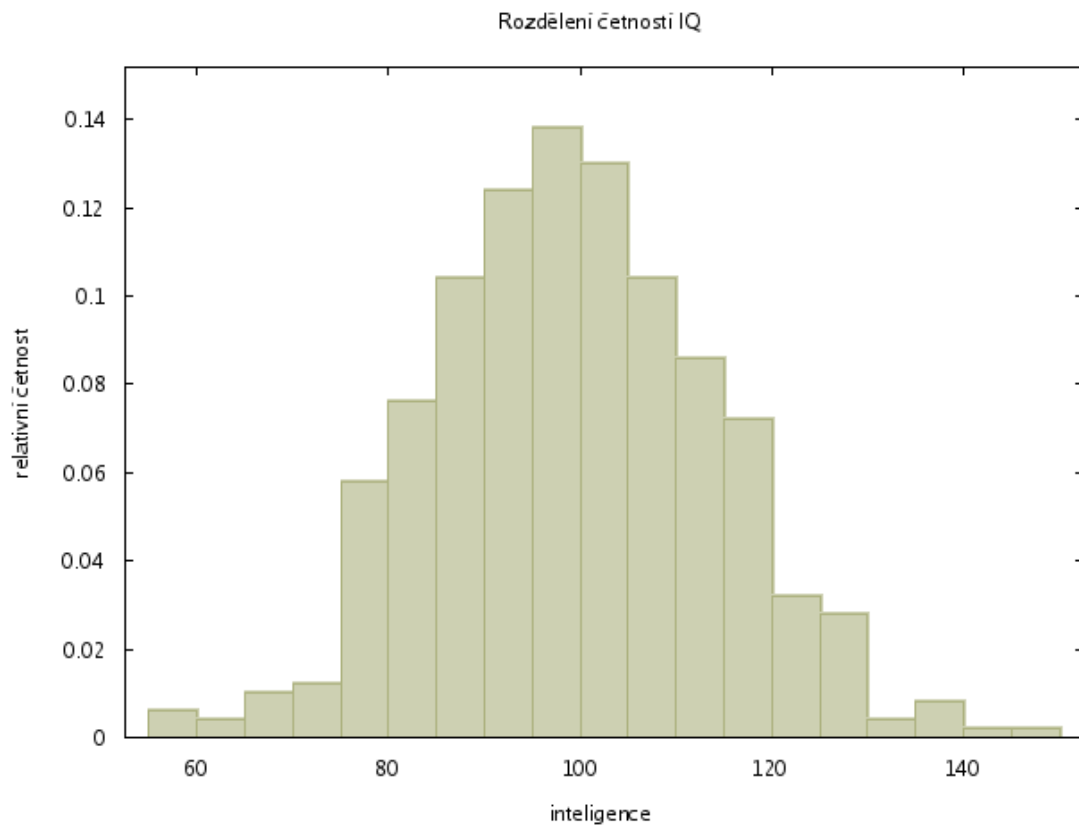
Frequency distribution for IQ, obs 1-500
 number of bins = 11, mean = 99.3317, sd = 14.679

interval	midpt	frequency	rel.	cum.
< 63.015	58.440	4	0.80%	0.80%
63.015 - 72.165	67.590	9	1.80%	2.60%
72.165 - 81.315	76.740	46	9.20%	11.80% ***
81.315 - 90.465	85.890	80	16.00%	27.80% *****

90.465 - 99.615	95.040	123	24.60%	52.40%	*****
99.615 - 108.77	104.19	108	21.60%	74.00%	*****
108.77 - 117.92	113.34	78	15.60%	89.60%	*****
117.92 - 127.07	122.49	37	7.40%	97.00%	**
127.07 - 136.22	131.64	11	2.20%	99.20%	
136.22 - 145.37	140.79	3	0.60%	99.80%	
>= 145.37	149.94	1	0.20%	100.00%	

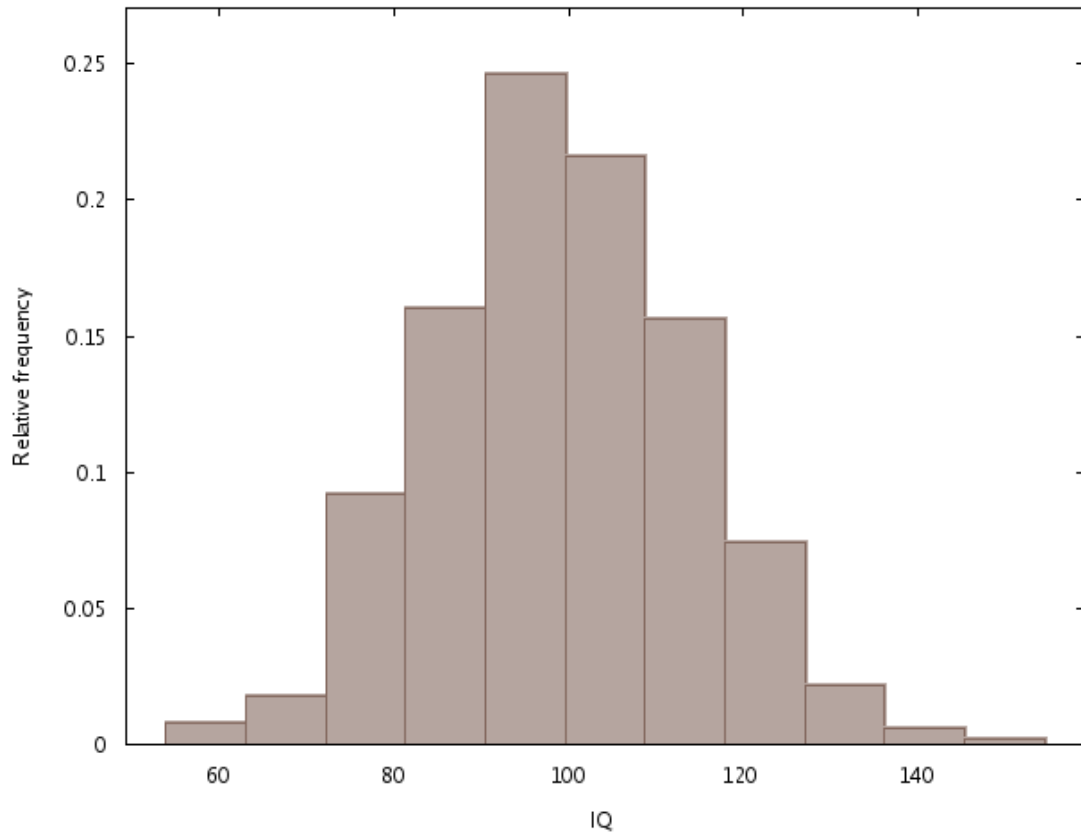
3. Gretl: *Variable* → *Frequency plot* → *Minimum value*, *left bin* zvol 50 a *Bin width* zvol 5

Figure 1: Histogram IQ 1



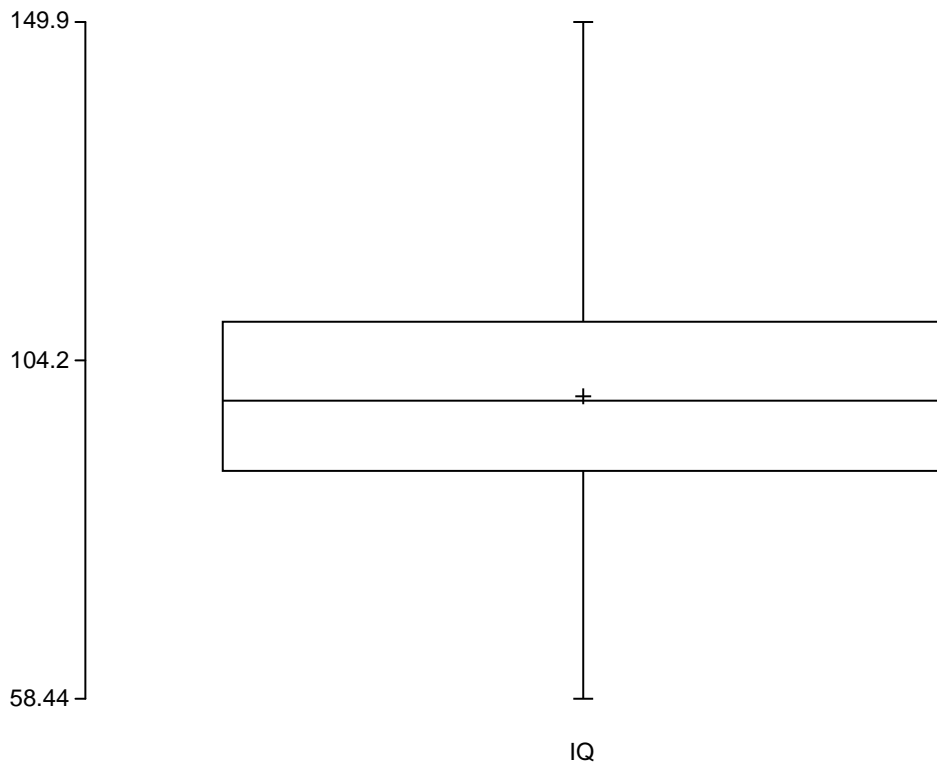
4. Gretl: *Variable* → *Frequency plot* → *Number of bins* zvol 11

Figure 2: Histogram IQ 2



5. Gretl: *View* → *Graph specified vars* → *Boxplot*

Figure 3: Boxplot



Klikni myší na obrázek boxplotu, zvol *Numerical summary*

Numerical summary

	mean	min	Q1	median	Q3	max	
IQ	99.332	58.44	89.248	98.74	109.41	149.94	(n=500)

6. Gretl: *Variable* → *Summary statistic*

Summary Statistics, using the observations 1 - 500
for the variable 'IQ' (500 valid observations)

Mean	99.332
Median	98.740
Minimum	58.440
Maximum	149.94
Standard deviation	14.679
C.V.	0.14778
Skewness	0.11914
Ex. kurtosis	-0.010735

7. Histogram IQ má zvonovitý tvar, což indikuje normální rozdělení.

Cvičení (data 02_descriptive_test_oblibenost_atd.csv)

Výzkumník má záměr zkoumat vztah mezi skórem v testu z matematiky a dalších proměnných jako je hodnocení respondentů o jejich oblíbenosti matematiky (škála: 1=velmi oblíbená až 5=zcela neoblíbená), hodnocení respondentů toho, jak jim přijde matematika obtížná (škála: 1=velmi obtížná až 5=velmi snadná), bydliště (1=město, 0=vesnice) a pohlaví (1=žena, 0=muž). Výzkumník provedl náhodný výběr 33 studentů, od kterých sebral všechny údaje. Zpřehledněte data pomocí deskriptivní statistiky. Konkrétně se můžete zaměřit na následující:

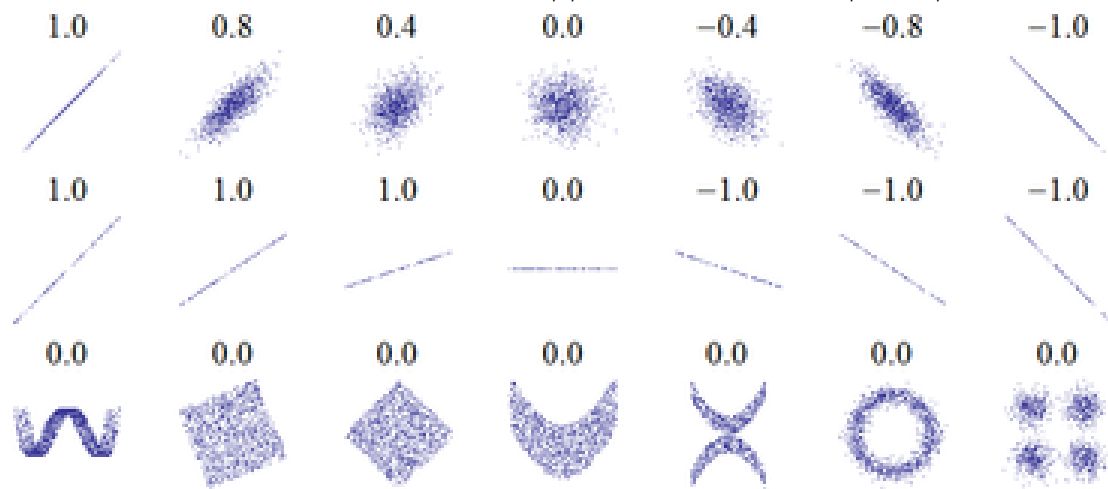
- Sestavte tabulku četností (absolutních, relativních a kumulativních) pro všechny proměnné.
- Spočítejte průměr, medián, minimum, maximum, standardní odchylku a roztyl.
- Reprezentujte data pomocí vhodně zvoleného grafu (histogram, sloupcový graf atd.)

”Deskriptivní statistika je deskriptivní.” Používej jen takové míry polohy a variability, které slouží k zpřehlednění dat a účelu tvé studie.

2.2.4 Korelační koeficient

- Korelační koeficient udává míru lineárního vztahu mezi dvěma proměnnými.
- Jeho hodnoty se pohybují mezi -1 a 1.
- Podle znaménka korelace ("+" či "-") můžeme usoudit, zda je vztah mezi proměnnými kladný či záporný. Negativní hodnota korelačního koeficientu naznačuje, že vztah mezi dvěma proměnnými je záporný, tj. zvětšíme-li hodnotu jedné proměnné, zmenší se hodnota druhé proměnné. Pozitivní hodnota korelačního koeficientu naznačuje, že vztah mezi dvěma proměnnými je kladný, tj. zvětšíme-li hodnotu jedné proměnné, zvětší se hodnota i druhé proměnné.
- Vzdálenost korelačního koeficientu od nuly indikuje těsnost lineárního vztahu mezi dvěma proměnnými:
 - do 0.2 - lineární vztah je zanedbatelný
 - od 0.2 do 0.4 - lineární vztah je nepříliš těsný
 - od 0.4 do 0.7 - lineární vztah je středně těsný
 - od 0.7 do 0.9 - lineární vztah je velmi těsný vztah
 - od 0.9 - lineární vztah je extrémně těsný
- Je-li hodnota korelačního koeficientu nízká až nulová, neznamená to, že mezi proměnnými nemůže být žádný vztah. Znamená to pouze, že mezi veličinami je lineární vztah zanedbatelný.
- Vysoká hodnota korelačního koeficientu nemusí znamenat, že je mezi proměnnými kauzální vztah. Znamená pouze predikční vztah.

Figure 4: Korelace - zdroj <http://cs.wikipedia.org/wiki/Korelace>



Příklad (data 03_korelace_vek_plat.csv)

Výzkumník chtěl zjistit míru lineárního vztahu mezi věkem a platem. Náhodně vybral 19 respondentů, kterých se dotázal na jejich věk a hodinový plat. Následující tabulka shrnuje získané údaje:

respondent	vek	plat
1	30	116
2	45	140
3	32	119
4	56	152
5	60	157
6	23	105
7	25	110
8	48	142
9	57	158
10	63	166
11	49	145
12	52	149
13	61	161
14	44	135
15	36	126
16	53	147
17	35	125
18	63	164
19	49	145

Vypčítejte korelační koeficient. Jaký směr má vztah mezi věkem a platem (kladný, záporný)? Jak těsný je vztah mezi věkem a pohlavím (zanedbatelný, nepříliš těsný vztah, středně těsný vztah, velmi těsný vztah a extrémně těsný vztah)?

Řešení

Gretl: *View* → *Correlation matrix*

corr(vek, plat) = 0.99647103

Under the null hypothesis of no correlation:

t(17) = 48.9478, with two-tailed p-value 0.0000

Korelační koeficient mezi věkem a platem je v našem příkladu roven 0.996. Směr vztahu je kladný. Vztah je extrémně těsný.

2.3 Inferenční statistika

2.3.1 Jednovýběrový t-test

Jednovýběrový t-test se používá pro testování toho, zda-li je střední hodnota (průměr) v nějaké populaci rovna předem stanovené hodnotě.

Příklad (data 04_ttest_pocetzaku.csv)

Výzkumník chtěl zjistit, zda-li je průměrný počet žáků v jedné třídě odlišný od 20. Zaměřil se na populaci žáků v osmých ročnících na základních školách. Aby mohl provést tento test, provedl náhodný výběr ze všech tříd osmých ročníků základních škol. U těchto tříd zjistil počet žáků ve třídě:

třída	počet
1	30
2	12
3	25
4	20
5	18
6	19
7	14
8	13
9	15
10	20
11	14
12	17
13	31
14	35
15	8
16	17
17	16
18	19
19	20
20	7
21	32
22	20
23	14
24	25
25	26
26	24
27	22
28	23
29	21

Na hladině významnosti 10 procent testujte, zda-li je průměrný počet žáků ve třídě odlišný od 20.

Řešení

Nulová hypotéza $H_0 : \mu = 20$, alternativní hypotéza $H_1 : \mu \neq 20$

Gretl: *Tools* → *Test statistic calculator* → *mean*

Null hypothesis: population mean = 20

Sample size: n = 29

Sample mean = 19.8966, std. deviation = 6.82613

Test statistic: $t(28) = (19.8966 - 20)/1.26758 = -0.0816108$

Two-tailed p-value = 0.9355

(one-tailed = 0.4678)

Na hladině významnosti 10 procent nemůžeme zamítnout nulovou hypotézu, protože p-hodnota 0.9355 je větší než 0.1 (10 procent), tj. nemůžeme říci, že průměrný počet žáků v jedné třídě je odlišný od 20. (Žáky myslíme žáky osmých ročníků základních škol.)

Cvičení (data 05_ttest_obtiznost.csv)

Výzkumník chtěl zjistit, jak hodnotí studenti prvních ročníků gymnázií obtížnost předmětu biologie. Provedl náhodný výběr těchto studentů. Následně jim položil otázku, jak hodnotí obtížnost předmětu biologie na rating škále od 1 (velmi snadný předmět) do 10 (velmi obtížný předmět). Hodnocení studentů je shrnuto v následující tabulce:

zak	obtiznost
1	5
2	9
3	6
4	1
5	2
6	1
7	3
8	2
9	4
10	2
11	2
12	1
13	1
14	3

Na hladině významnosti 5 procent testujte, zda-li se hodnocení obtížnosti biologie liší od 5 (ani snadný, ani obtížný předmět).

2.3.2 Dvouvýběrový t-test

Dvouvýběrový t-test se používá (mimo jiné) pro porovnání středních hodnot (průměrů) ve dvou základních populacích (nezávislých populacích). Toto porovnání provádíme na základě náhodného výběru z jedné a následně náhodného výběru z druhé populace.

Příklad (data 06_ttest_spokojenost_pohlavi.csv)

Výzkumník chtěl zjistit, zda-li se liší spokojenost se vzdělávacím systémem v dané zemi mezi ženami a muži. Provedl náhodný výběr jedenácti žen a osmi mužů a zeptal se jich zda-li jsou spokojeni se vzdělávacím systémem. Své hodnocení měli respondenti uvést na rating škále od jedné do pěti, na níž jedna reprezentuje "velmi nespokojen" a pět "velmi spokojen". Data, která výzkumník získal jsou následující:

ženy	muži
4	5
5	1
2	2
1	2
5	3
4	2
2	1
3	3
2	
1	
2	

Na hladině významnosti 5 procent testujte, zda-li je spokojenost mužů a žen se vzdělávacím systémem odlišná.

Řešení

- Testováním odlišnosti průměrné spokojenosti mužů a žen musíme nejprve provést jiný test, abychom určili, zda je variance (rozptýlenost) spokojenosti mužů a žen odlišná či nikoli. Závěr testu pro porovnání dvou variancí použijeme jako předpoklad pro testování průměrné spokojenosti mužů a žen. Test pro porovnání dvou rozptylů nazýváme F-test pro porovnání dvou rozptylů.
- Provedení F-testu pro porovnání rozptylu jedné populace σ_1^2 a rozptylu druhé populace σ_2^2 na hladině významnosti 5 procent
 Nulová hypotéza $H_0: \sigma_1 = \sigma_2$, alternativní hypotéza $H_1: \sigma_1 \neq \sigma_2$
 Gretl: *Tools* → *Test statistic calculator* → *2 variances*

Null hypothesis: The population variances are equal

Sample 1:

n = 11, variance = 2.16364

Sample 2:

n = 8, variance = 1.69643
Test statistic: $F(10, 7) = 1.27541$
Two-tailed p-value = 0.7684
(one-tailed = 0.3842)

P-hodnota je větší než 0.05. Na hladině významnosti 5 procent tudíž nemůžeme zamítnout nulovou hypotézu o shodnosti rozptylů. T-test pro porovnání průměrů dvou populací provedeme s předpokladem, že rozptyly (standardní odchylky) v těchto dvou populacích jsou shodné.

- Provedení t-testu pro porovnání dvou průměrů na hladině významnosti 5 procent
Nulová hypotéza $H_0: \mu_1 = \mu_2$, alternativní hypotéza $H_1: \mu_1 \neq \mu_2$
Gretl: *Tools* → *Test statistic calculator* → *2 means* (Předpoklad: Zaškrtni okénko u "Assume common population standard deviation")

Null hypothesis: Difference of means = 0

Sample 1:

n = 11, mean = 2.81818, s.d. = 1.47093
standard error of mean = 0.443502
95% confidence interval for mean: 1.83 to 3.80637

Sample 2:

n = 8, mean = 2.375, s.d. = 1.30247
standard error of mean = 0.460493
95% confidence interval for mean: 1.28611 to 3.46389

Test statistic: $t(17) = (2.81818 - 2.375)/0.65239 = 0.679321$
Two-tailed p-value = 0.5061
(one-tailed = 0.253)

P-hodnota je větší než 0.05. Na hladině významnosti 5 procent tudíž nemůžeme zamítnout nulovou hypotézu o shodnosti průměrů, tj. nemůžeme říci, že průměrná spokojenost se vzdělávacím systémem je mužů a žen odlišná.

Cvičení (data 07_ttest_esej_mapa.csv)

Výzkumník chtěl porovnat účinek dvou vyučovacích metod (psaní esejí a využití concept mapping) na to, jak studenti na konci kurzu rozumí vyučované látce. Aby mohl účinek těchto dvou metod porovnat, provedl experiment. Rozdělil náhodně studenty do dvou skupin. Jedna skupina měla během kurzu využívat ke strukturaci učiva eseje (během kurzu museli studenti napsat dvě eseje) a druhá skupina měla využívat metodu pojmového mapování (během kurzu museli studenti sestavit dvě pojmové mapy). Studenti tak během kurzu získávali nové vědomosti, zamýšleli se nad novými otázkami a ke strukturaci a shrnutí svých znalostí používali buď eseje či mapy. Na konci kurzu šli ke zkoušce, kde měli prokázat porozumění nově naučené látce. (Jako měřítko porozumění látce byla zvolena známka u zkoušky.) Výsledky studentů u zkoušky (známka 1 až 5) shrnuje následující tabulka:

esej	mapa
1	2
1	3
2	1
3	1
3	2
2	2
1	3
3	1
4	1
4	2
3	1
2	2
4	1
3	1
	1
	2

Přepokládejte, že studenti v obou skupinách jsou náhodným výběrem z populace studentů. Na hladině významnosti 10 procent testujte, zda-li je účinek těchto dvou vyučovacích metod v populaci studentů odlišný.

2.3.3 T-test pro korelační koeficient

Příklad (data 08_koreltest_vzdelani_prijem.csv)

Často zkoumaným vztahem v sociálních vědách je vztah mezi příjmem a vzděláním. Abychom tento vztah mohli zkoumat, byl proveden náhodný výběr patnácti osob z ekonomicky aktivních lidí (populace), kteří byli dotázáni na jejich vzdělání (měřeno počtem let vzdělání) a jejich příjem (měřeno v tisících). Následující tabulka shrnuje získaná data:

individum	vzdělání	příjem
1	9	12
2	14	30
3	10	10
4	13	20
5	14	28
6	10	13
7	12	15
8	15	33
9	17	25
10	13	20
11	14	30
12	13	16
13	13	25
14	17	45
15	20	40

1. vypočítej korelační koeficient mezi vzděláním a příjmem

- testuj na hladině významnosti 5 %, zda-li je korelační koeficient signifikantně odlišný od nuly
nulová hypotéza $H_0 : \rho = 0$, alternativní hypotéza $H_1 : \rho \neq 0$

Řešení

Gretl: *View* → *Correlation*

```
corr(vzdelani, prijem) = 0.86691624
```

Under the null hypothesis of no correlation:

```
t(13) = 6.27081, with two-tailed p-value 0.0000
```

- korelační koeficient mezi vzděláním a příjmem je roven 0.87
- korelační koeficient je signifikantně odlišný od nuly na hladině významnosti 5%, protože p-hodnota 0.0000 je menší než 0.05.

Cvičení

- Z populace žáků osmých ročníků byli náhodně vybráni tři žáci, u nichž byla zjištěna známka z českého jazyka na vysvědčení na konci osmého ročníku a známka z testu, kterou dostali z posledního písemného testu z českého jazyka.

	známka	
žák	vysvědčení	test
1	1	2
2	2	3
3	3	7

Vypočítej korelační koeficient a testuj, zda-li je na hladině významnosti 5 % signifikantně odlišný od nuly.

2. (data 09_koreltest_vysvedceni_test.csv) Z populace žáků osmých ročníků bylo náhodně vybráno patnáct žáků, u nichž byla zjištěna známka z českého jazyka na vysvědčení na konci osmého ročníku a známka z testu, kterou dostali z posledního písemného testu z českého jazyka.

žák	známka	
	vysvědčení	test
1	1	1
2	2	3
3	2	1
4	1	1
5	3	3
6	4	4
7	2	3
8	3	3
9	4	4
10	1	2
11	1	1
12	1	1
13	3	3
14	3	5
15	4	4

Vypočítej korelační koeficient a testuj, zda-li je na hladině významnosti 5 % signifikantně odlišný od nuly.

3. Porovnej korelační koeficienty v předchozích dvou cvičeních. Porovnej závěry testů (na hladině významnosti 5 %) o odlišnosti korelačního koeficientu od nuly. Porovnej tyto dva závěry!

2.3.4 Chí-kvadrát test

Příklad (data 10_chitest_nazor_pohlavi.csv)

Vyučující chtěl zjistit, zda-li souvisí názor studentů o obtížnosti kurzu s pohlavím studenta. Náhodně vybral 166 studentů, u kterých zaznamenal názor na obtížnost kurzu (obtížné, snadné) a jejich pohlaví (viz datový soubor nazor_pohlavi). Na hladině významnosti 10 % testuj, zda-li názor ohledně obtížnosti kurzu souvisí s pohlavím studenta.

Řešení

Nulová hypotéza H_0 : názor a pohlaví navzájem nesouvisí, alternativní hypotéza H_1 : názor a pohlaví spolu souvisí

Gretl: *View* → *Cross Tabulation*

Cross-tabulation of nazor (rows) against pohlavi (columns)

	[0]	[1]	TOT.
[0]	42	33	75
[1]	27	64	91
TOTAL	69	97	166

Pearson chi-square test = 11.7349 (1 df, p-value = 0.000613377)

Na hladině významnosti 10 % (=0.1) zamítáme nulovou hypotézu, protože p-hodnota je menší než 0.1 . Na hladině významnosti 10 % (=0.1) lze říci, že názor ohledně obtížnosti kurzu a pohlaví spolu navzájem souvisí.

2.3.5 Lineární regrese

- slouží k predikci či odhadu jedné proměnné Y na základě znalosti další proměnné X (proměnných)
- slovo "lineární" označuje, že předpokládáme lineární vztah mezi proměnnou Y a X , tj. proměnné mohou být reprezentovány grafem scatterplot, v němž se body mají tendenci nacházet kolem přímky

- tato přímka je nazývána přímkou lineární regrese
- tato přímka reprezentuje, jak souvisí změna proměnné X se změnou proměnné Y

Příklad (data 11_regrese_seminar_zkouska.csv)

Vysokoškolský učitel chtěl zjistit, zda-li souvisí počet seminářů, které student během semestru navštívil, s výsledným počtem bodů v zkouškovém testu. U náhodného výběru 20 studentů si zaznamenal počet navštívených seminářů během semestru (rozmezí 0-13) a počet bodů v zkouškovém testu (rozmezí 0-100 procent):

student	pocet seminaru	vysledek zk
1	0	3
2	13	50
3	5	40
4	13	90
5	13	70
6	12	100
7	11	97
8	4	20
9	2	10
10	10	56
11	9	80
12	13	90
13	12	78
14	14	83
15	1	2
16	4	24
17	10	80
18	3	34
19	0	7
20	1	2

1. Uveďte popisné statistiky (průměr, medián, minimum, maximum a standardní odchylka) pro obě zkoumané proměnné (počet seminářů, výsledek u zkoušky)
2. Reprezentujte data pomocí grafu scatterplot, zakreslete výběrovou regresní přímku (odhad regresní přímky)
3. Na hladině významnosti 5 procent testujte, zda-li je koeficient u počtu navštívených seminářů signifikantně odlišný od nuly, tj. zda-li počet navštívených seminářů pomáhá signifikantně vysvětlit výsledek ve zkouškovém testu
4. Interpretujte koeficient u počtu navštívených seminářů.
5. Jaký výsledek (počet bodů) ve zkouškovém testu může dle našeho regresního modelu očekávat student, který navštívil 7 seminářů? Jaký výsledek může očekávat student, který navštívil 9 seminářů?
6. Porovnej predikci výsledku v testu pro studenta, který navštívil 9 seminářů se sebranými údaji vysokoškolského profesora. (Je predikce výsledku shodná s daty, které učitel naměřil? Proč tomu tak je?)
7. Je mezi počtem navštívených seminářů a výsledku v zkouškovém testu kauzální vztah?

Řešení

1. Gretl: *View* → *Summary statistics*

Summary Statistics, using the observations 1 - 20
for the variable 'pocet_seminaru' (20 valid observations)

Mean	7.5000
Median	9.5000
Minimum	0.0000
Maximum	14.000

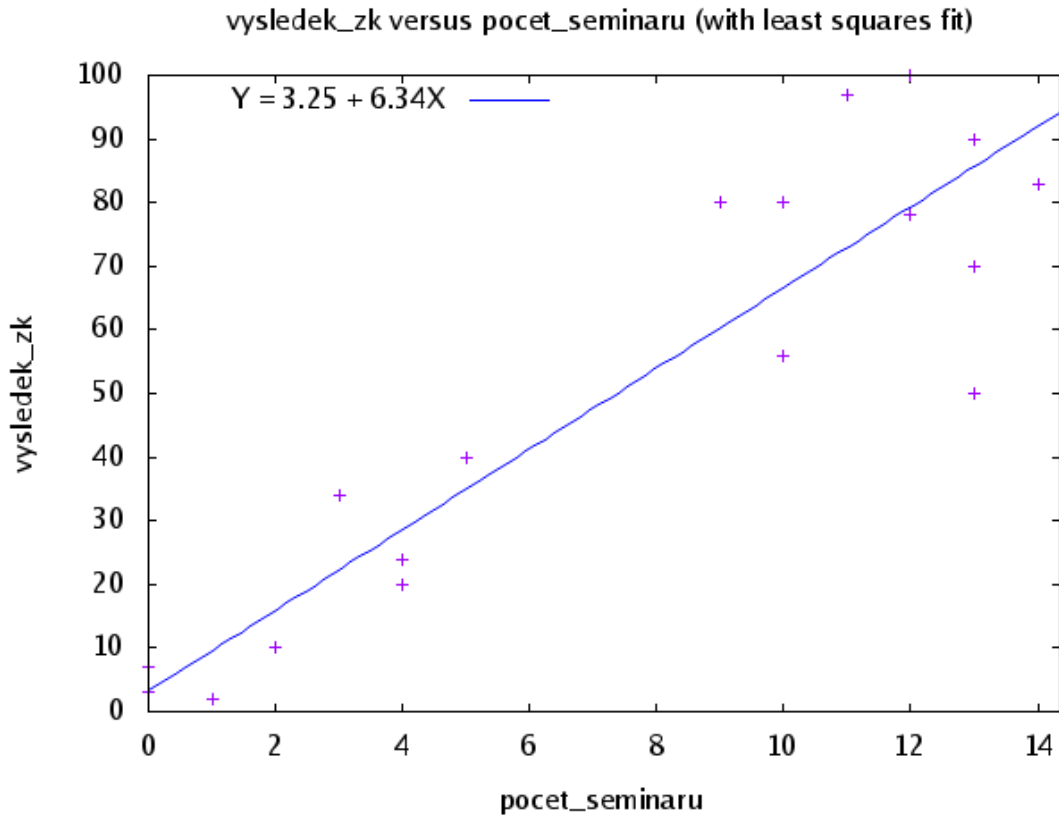
Standard deviation	5.1759
C.V.	0.69011
Skewness	-0.21497
Ex. kurtosis	-1.6057

Summary Statistics, using the observations 1 - 20
for the variable 'vysledek_zk' (20 valid observations)

Mean	50.800
Median	53.000
Minimum	2.0000
Maximum	100.00
Standard deviation	35.691
C.V.	0.70258
Skewness	-0.11800
Ex. kurtosis	-1.5362

2. Gretl: *View* → *Graph specified vars*

Figure 5: Scatterplot



3. Gretl: *Model* → *Ordinary least squares*

Model 1: OLS estimates using the 20 observations 1-20
 Dependent variable: vysledek_zk

	coefficient	std. error	t-ratio	p-value
const	3.25088	5.77839	0.5626	0.5807
pocet_seminaru	6.33988	0.639287	9.917	1.01E-08 ***

Mean of dependent variable = 50.8
 Standard deviation of dep. var. = 35.6911
 Sum of squared residuals = 3744.4
 Standard error of the regression = 14.423
 Unadjusted R-squared = 0.84529

Adjusted R-squared = 0.83670
Degrees of freedom = 18
Log-likelihood = -80.7016
Akaike information criterion (AIC) = 165.403
Schwarz Bayesian criterion (BIC) = 167.395
Hannan-Quinn criterion (HQC) = 165.792

- Výběrová regresní přímka je: $V = 3.25 + 6.34S$, kde S je počet seminářů a V je výsledek u zkoušky
 - Koeficient u počtu seminářů je tedy roven 6.34. Tento koeficient je signifikantně odlišný od nuly na hladině významnosti 5 procent, protože p-hodnota $1.01E-08$ je menší než 0.05 (5 procent). (Porovnej tento závěr se závěrem testu o tom, zda je korelační koeficient mezi počtem seminářů a výsledkem u zkoušky signifikantně odlišný od nuly na hladině významnosti 5 %.)
4. Pokud se počet navštívených seminářů zvýší o jeden, lze očekávat, že procentuální výsledek ve zkuškovém testu v průměru o 6.34 procentního bodu.
 5. Predikce výsledku testu pro studenta, který navštívil 7 seminářů je roven $3.25 + 6.37 \cdot 7 = 47.84$ procent. Predikce výsledku testu pro studenta, který navštívil 9 seminářů je roven $3.25 + 6.37 \cdot 9 = 60.58$ procent.
 6. Vysokoškolský učitel má ve svém výběru jednoho studenta, který navštívil 9 seminářů. Jeho výsledek ve zkuškovém testu je 80 procent. Dle našeho modelu lze pro studenta, který navštívil 9 seminářů predikovat výsledek 60.58 procent. Rozdíl mezi těmito závěry lze vysvětlit např. chybou měření výsledku studenta. Je možné, že při opravě testu či zaznamenávání výsledku tohoto studenta udělal učitel chybu. Dalším důvodem by mohlo být, že použitý model lineární regrese není správným modelem pro tuto situaci. Je možné, že jiný model vysvětluje výsledek testu na základě počtu seminářů přesněji.

7. Daný vztah mezi počtem seminářů a výsledkem v testu je predikčním vztahem. Na základě počtu seminářů predikujeme výsledek v testu. O kauzálním vztahu nelze jednoznačně nic říci. Nemůžeme tedy říci, že zvýšení počtu seminářů o jeden je příčinou zvýšení výsledku v testu o 6.34 procentního bodu. (Příčinou dobrého výsledku u zkoušky může být např. velká píle studenta. Proměnná pilnost studenta však v našem regresním modelu není zahrnuta. Tato proměnná je však korelována s počtem navštíveným seminářů, který v našem modelu je zahrnut. Reálně tak může být vliv počtu seminářů na výsledek u zkoušky nesiginifikantní (nevýznamný; není signifikantně odlišný od nuly). Ale vzhledem ke korelaci s nepozorovanou proměnnou píle studenta vyjde v modelu koeficient u počtu navštívených seminářů nadhodnocený a signifikantně odlišný od nuly.)

Cvičení (data 12_regrese_test_IQ_konzultace.csv)

1. Učitel chtěl zjistit vztah mezi počtem hodin, které s ním student konzultoval, a výsledkem v testu z matematiky. Provedl náhodný výběr dvaceti studentů, u kterých si zaznamenal procentuální výsledek v testu a počet hodin, které student využil pro konzultování příkladů, kterým v průběhu semestru méně nerozuměl. Proměnnou, kterou učitel nepozoroval je výše IQ. Všechna data (tj. ta, které učitel měl i neměl k dispozici) shrnuje následující tabulka:

student	test	IQ	konzultace
1	71.32	89	2.1
2	78.58	96	1.4
3	74.50	91	1.5
4	93.64	116	4.2
5	75.34	92	2.2
6	83.06	102	1.8
7	72.34	89	0.7
8	79.06	98	1.3
9	78.30	97	0.0
10	77.66	97	2.3
11	84.88	101	2.9
12	65.20	80	0.5
13	82.54	101	2.2
14	94.28	116	3.4
15	79.78	98	2.4
16	76.00	93	3.0
17	80.82	98	2.6
18	87.18	108	3.4
19	92.04	112	3.2
20	77.92	95	1.1

- (a) Uveďte popisné statistiky (průměr, medián, minimum, maximum a standardní odchylka) pro proměnné, které učitel měl i neměl k dispozici (výsledek v testu, počet konzultačních hodin a IQ).
- (b) Reprezentujte data pro výsledek v testu a počet konzultačních hodin pomocí grafu scatterplot, na vodorovnou osu naneste počet konzultačních hodin a na svislou osu výsledek v testu. Zakreslete výběrovou regresní přímku (odhad regresní přímky).
- (c) Na hladině významnosti 5 procent testujte, zda-li je koeficient u počtu konzultačních hodin signifikantně odlišný od nuly, tj. zda-li počet konzultačních hodin pomáhá sig-

nifikantně vysvětlit počet bodů ve testu

- (d) Interpretujte koeficient u počtu konzultačních hodin.
- (e) Jaký výsledek (počet bodů) ve zkouškovém testu může dle našeho regresního modelu očekávat student, který konzultoval s učitelem 50 minut?
- (f) Nyní se zaměříme na proměnnou, kterou učitel nepozoroval, tj. IQ. Znázorněte graficky vztah mezi IQ a výsledkem v testu z matematiky.
- Odhadněte model lineární regrese pro IQ jako vysvětlující proměnnou a výsledek v testu jako vysvětlovanou proměnnou.
 - Je koeficient u výsledku v testu signifikantní na hladině významnosti 5 procent?
- (g) Model lineární regrese lze použít i v případě, kdy máme více než jednu vysvětlující proměnnou. V našem případě budeme chtít vysvětlit výsledek v testu pomocí počtu konzultačních hodin i IQ.
- Odhadněte model lineární regrese, kde jako vysvětlující proměnné (independent variables) použijete počet konzultačních hodin a IQ, tj. odhadni parametry a, b, c v rovnici $\text{výsledek} = a + b \cdot \text{IQ} + c \cdot \text{konzultace}$.
 - Jsou jsou odhadnuté koeficienty u IQ a počtu hodin konzultací signifikantně odlišné od nuly.
 - Jaká je interpretace těchto koeficientů?
 - Porovnej signifikanci a interpretaci koeficientu u konzultačních hodin v dvou regresních modelech: modelu, který má jednu vysvětlující proměnnou (počet konzultačních hodin), a modelu, který má dvě vysvětlující proměnné (počet konzultačních hodin i IQ).
 - Je vztah mezi počtem konzultačních hodin a výsledkem v testu kauzální?