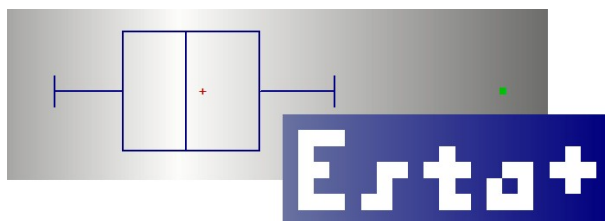


Metodologie ISK

Základy statistického zpracování dat

Ladislava Suchá, 28. dubna 2011

Programy na statistické zpracování dat



Aplikace na online dotazování, které zvládají některé základní i složitější statistické operace



SurveyMonkey.com
because knowledge is everything



easyresearch.biz



Fáze vyhodnocování dat

- Kódování
- Třídění prvního stupně
- (Úpravy znaků)
- Třídění druhého stupně

Kódování

- Jednotlivým variantám znaku jsou přiřazovány symboly (čísla) podle kódovacího čísla
- Kódování často probíhá přímo v terénu nebo ho provádí aplikace

Zápis do matice dat:

- Jednotlivé případy = řádky
- Jednotlivé proměnné = sloupce

Druhy proměnných

Nominální

- Známe hodnoty, ale můžeme o nich říci pouze to, že jsou různé
- Nelze provádět aritmetické operace
- Přiřazení znaku je symbolické

Pořadové

- Můžeme určit pořadí (vzdělání, spokojenost)
- Znaky = míra pořadovosti

Kardinální (intervalové, spojité)

- Můžeme říci, o kolik je jedna hodnota vyšší než druhá (měsíční příjem, počet dětí v domácnosti atd.)
- Přiřazení znaku = reálné číslo

 Nominal

 Ordinal

 Scale

Otázka v dotazníku a její zpracování

Považujete obor Informační studia a knihovnictví za perspektivní?

velmi perspektivní
spíše perspektivní
spíše neperspektivní
zcela neperspektivní

nevím, nemohu odpovědět
neodpověděl/a

1

2

3

4

-1

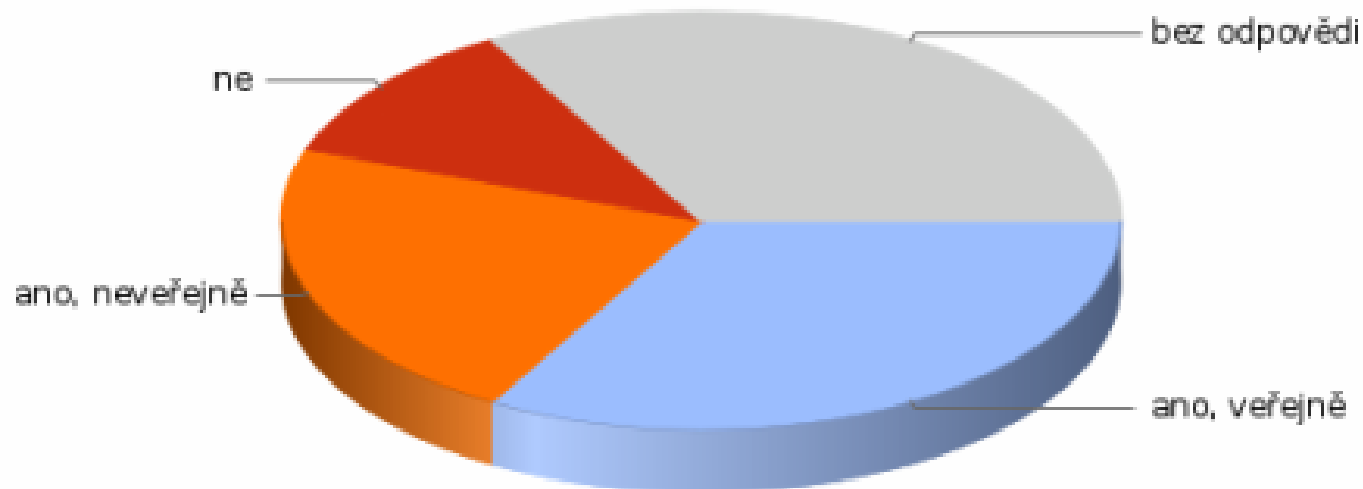
-2

Hodnoty
proměnné

Zapisujeme
jako „value
labels“

Chybějící
hodnoty
(missing
values)

Ukázka – zahrnutí missing values (chyba)



Graf 1: V jakém rozsahu poskytovatelé uchovávají data o uživatelích svých služeb

Q1_prinos
Studium na
KISK hodnotím
jako:

- 1 velmi přínosné
- 2 spíše přínosné
- 3 spíše nepřínosné
- 4 zcela nepřínosné
- 1 nevím / nemohu
odpovědět
- 2 Neodpověděl/a

- Q8_1 Povinné (A) kurzy mají
logickou časovou
posloupnost.
- Q8_2 Obsahy jednotlivých
povinných (A) kurzů se
nepřekrývají.
- Q8_3 Jsem spokojen/a s
tematickou šíří nabídky
povinně volitelných (B) kurzů.
- Q8_4 Jsem spokojen/a s
počtem nabízených povinně
volitelných (B) kurzů.

Případy
(cases)

*anketa2010 [DataSet1] - PASW Statistics Data Editor

File Edit View Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

4 : q6_prinos Děláním to co mě baví a neruší mě 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000

Visible: 32 of 32 Variables

	q1_prinos	q2_perspektiv	q3_doporučení	q4_znovuznání	zapory	q8_1_posl	q8_2_prinos	q8_3_tema	q8_4_pocetkurzu	q8_5_prace	q9_1_navrhkurzu	q9_2_navrhrkurzu	q9_3_navrhrkurzu	q10
1	1								3	-1	viz bod 12			10
2	1							1	3	2				3
3	2		4	2				2	2	-1				10
4	1	2	1	2				2	2	1	Práce ve ...			3
5	1	2	1	2										2
6	2	1	1	1										-1
7	1	1	2	1										20
8	1	1	1	1	1									30
9	2	2	1	1	Volnost, relativně zajímavé n...	Knihovnictví...					formatika	Datab...		25
10	2	2	4	2	určitý teoretický nadhled	příliš mnoho					kladní ...			-1
11	2	2	4	2	Inovativní přístup, nové trendy	Inovativní přis					rz zam...	nějaký...	kurz z...	15
12	2	1	1	2	Rozmanitost a pestrost před...	Občas má čl					bioped...	Psych...	Anglič...	
13	1	2	2	1	Důraz na samostatné myšle...	Vadí mi, jak								-1
14	1	2	2	1	Přínosem je mi to, že můžu ...									
15	2	1	1	1	volnost, mnoho času na oso...	Studium mi p					ormačn...	SEO, ...	více pr...	20
16	1	2	2	2	Poznal a dostal jsem se (as...	Žádné zápor								10
17	2	2	2	1	rozšíření znalostí	Nizká kredita								28
18	2	2	2	1	Celostní osobnostní rozvoj, ...	Je tu moc kn					arketin...	Co nej...		8
19	2	2	2	1	Studuji dálko									10
20	2	2	2	1										-1
21	2	2	2	1										-1
22	2	2	2	1	- vícedenní kurzy_x000D_- ...	- chtělo by to								7
23	2	2	1	1	Moderní, sveží forma. Osob...	Nejsem si jis					y jazy...	Více p...		9

Data View Variable View

Definování proměnných

Druh
proměnné

*anketa2010_19024010.sav [DataSet1] - PASW Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Wi...	De...	Label	Values	Missing	Col...	Align
1	q1_prinos	Numeric	1	0	Vnímáte studium na KISK jako přínosné?	{1, velmi pří...	None	6	Right
2	q2_perspektiva	Numeric	1	0	Považujete obor Informační studia a knihovnictví za perspe...	{1, velmi per...	None	6	Right
3	q3_doporuceni	Numeric	1	0	Doporučil/a byste studium na KISK svým přátelům?	{1, rozhodn...	None	5	Right
4	q4_znovustud	Numeric	1	0	Pokud byste se měl/a rozhodovat znovu o svém studiu s tí...	{1, ano}...	None	5	Right
5	q5_seberealizace	String	3	0	Máte pocit, že při studiu můžete uplatnit to, co umíte nejlé...	None	None	5	Left
6	q6_prinos	String	713	0	V čem spatřujete největší přínos svého studia na KISK FF ...	None	None	19	Left
7	q7_zapory	String	1148	0	Co Vám naopak studium na KISK vzalo, co se vám na stu...	None	None	18	Left
8	q8_1_posl	Numeric	1	0	Povinné (A) kurzy mají logickou časovou posloupnost.	{-2, neodpov...	-1, -2	4	Right
9	q8_2_prekr	Numeric	1	0	Obsahy jednotlivých povinných (A) kurzů se nepřekrývají.	{-2, neodpov...	-1, -2	6	Right
10	q8_3_tema	Numeric	1	0	Jsem spokojen/a s tematickou šíří nabídky povinně volitelných...	{-2, neodpov...	-1, -2	8	Right
11	q8_4_pocetkurzu	Numeric	1	0	Jsem spokojen/a s počtem nabízených povinně volitelných...	{-2, neodpov...	-1, -2	6	Right
12	q8_5_praxe	Numeric	1	0	Absolvování povinné praxe bylo přínosem.	None	0, 0	5	Right
13	q9_1_navrhkurzu	String	162	0	Navrhovaný kurz 1	None	None	7	Left
14	q9_2_navrhkurzu	String	264	0	Navrhovaný kurz 2	None	None	5	Left
15	q9_3_navrhkurzu	String	196	0	Navrhovaný kurz 3	None	None	5	Left

Zápis baterie
otázek

Třídění prvního stupně

- Sleduje se četnost výskytu jednotlivého znaku
 - Kolik je v souboru mužů a žen
 - Kolik je v souboru lidí, kteří chodí do knihovny atd...
- Sledujeme základní **statistické míry znaků**

Třídění prvního stupně

Absolutní četnosti

- Absolutní číslo – kolik případů má danou vlastnost
- Součet absolutních četností u všech hodnot (včetně missing values) = celkový počet respondentů

(V souboru je 71 žen.)

Relativní četnosti

- Jaký podíl (v procentech z výběrového souboru) představují případy s jednotlivou vlastností

(V souboru je 34 % osob se středoškolským vzděláním.)

Kumulativní relativní četnosti

(V souboru je 52 % osob s alespoň středoškolským vzděláním.)

Rozložení hodnot proměnných

Statistics

Považujete za dostačující komunikaci KISKu se studenty?

N	Valid	442
	Missing	3

„Missing“ do grafů
nezařazujeme

(počítáme jen s
těmi, kteří
odpověděli)

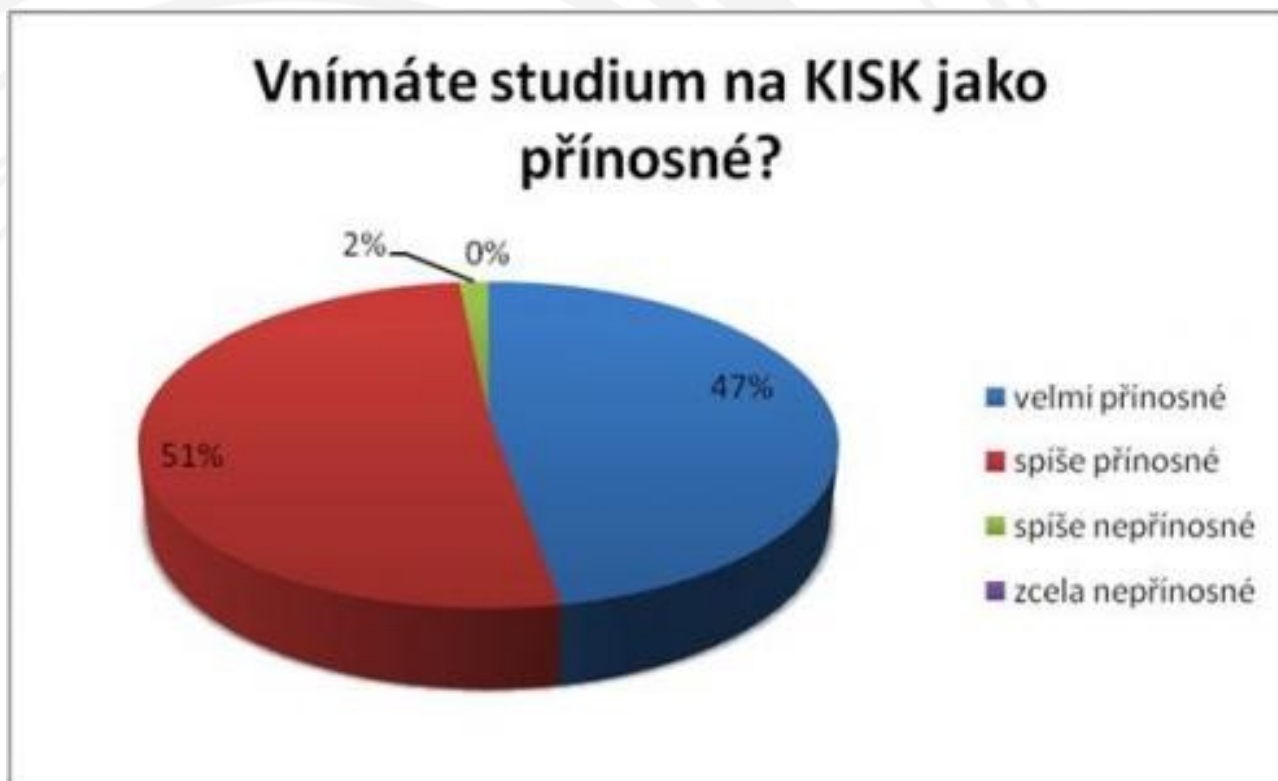
Relativní
četnosti bez
„missing
values“

Považujete za dostačující komunikaci KISKu se studenty?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ano	390	87,6	88,2	88,2
	s výhradami, jak kteří vyučující	22	4,9	5,0	93,2
	ne	30	6,7	6,8	100,0
	Total	442	99,3	100,0	
Missing	neodpověděli	3	,7		
Total		445	100,0		

Zobrazování výsledků

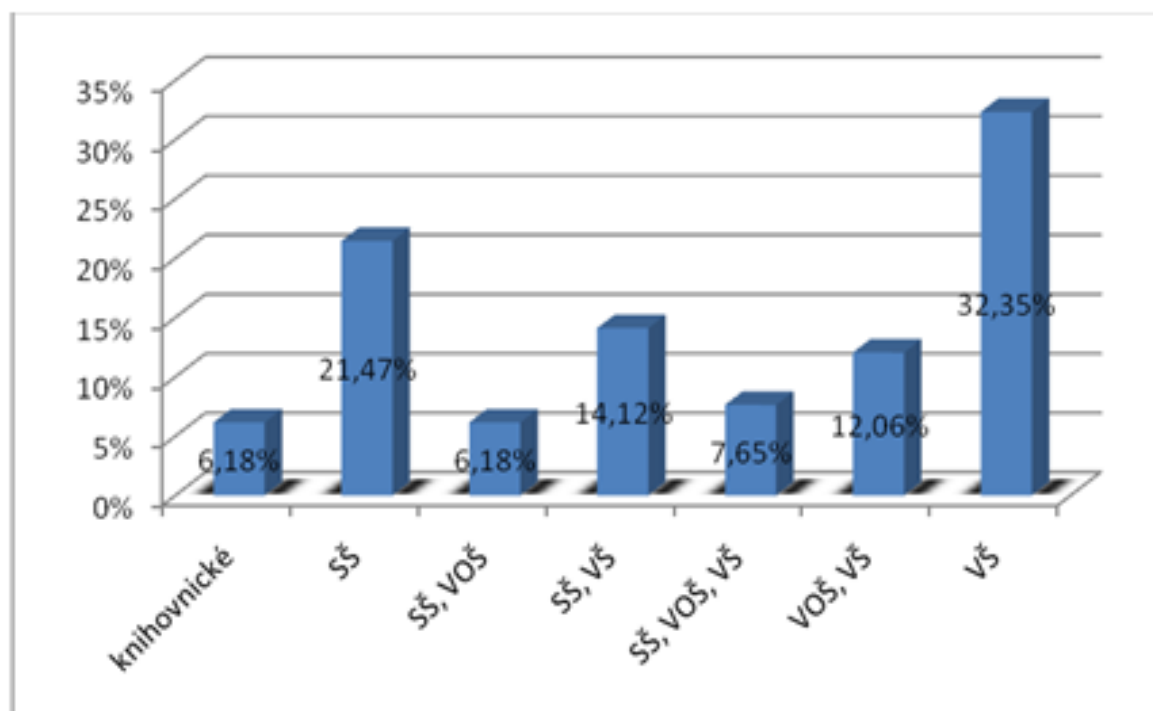
Koláčové, sloupcové grafy



Zobrazování výsledků

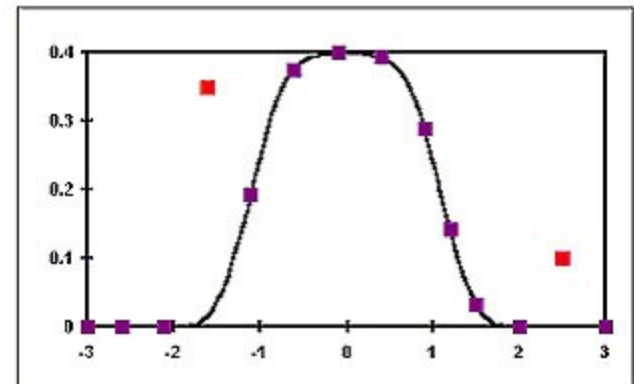
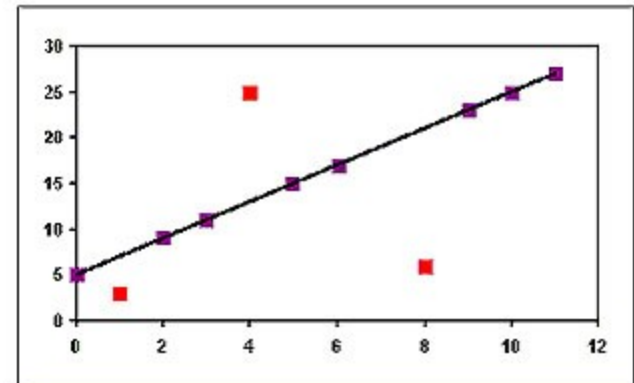
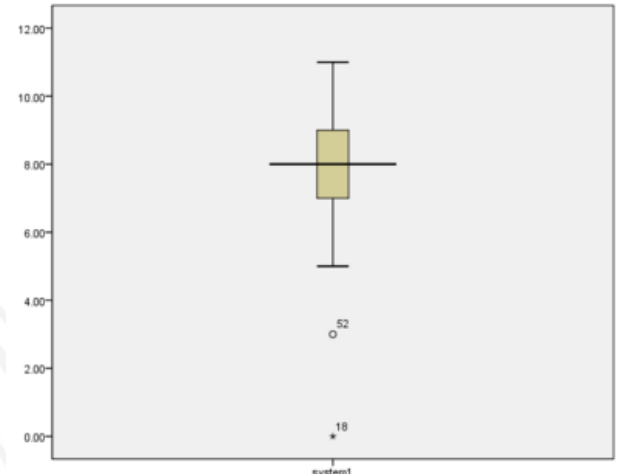
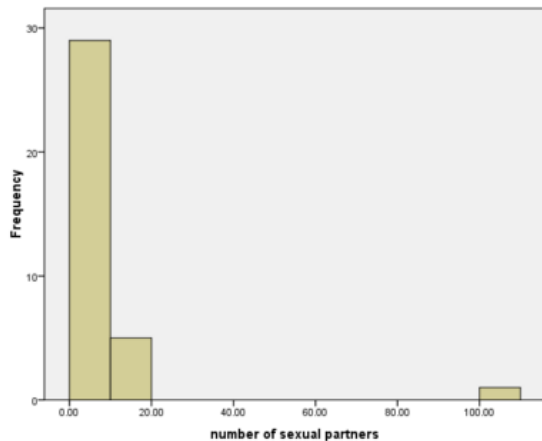
Koláčové, sloupcové grafy

Graf č. 4 – Požadované vzdělání knihovníků



Deskriptivní statistika a čištění dat

- První krok při každém zpracování dat
- „GIGO“ (*Garbage in, garbage out*)
- **Outliers** (extrémní hodnoty)
 - Podíváme se na nejvyšší a nejnižší hodnoty
 - (SPSS najde automaticky)



Charakteristiky rozložení proměnné: modus, medián, průměr

MODUS

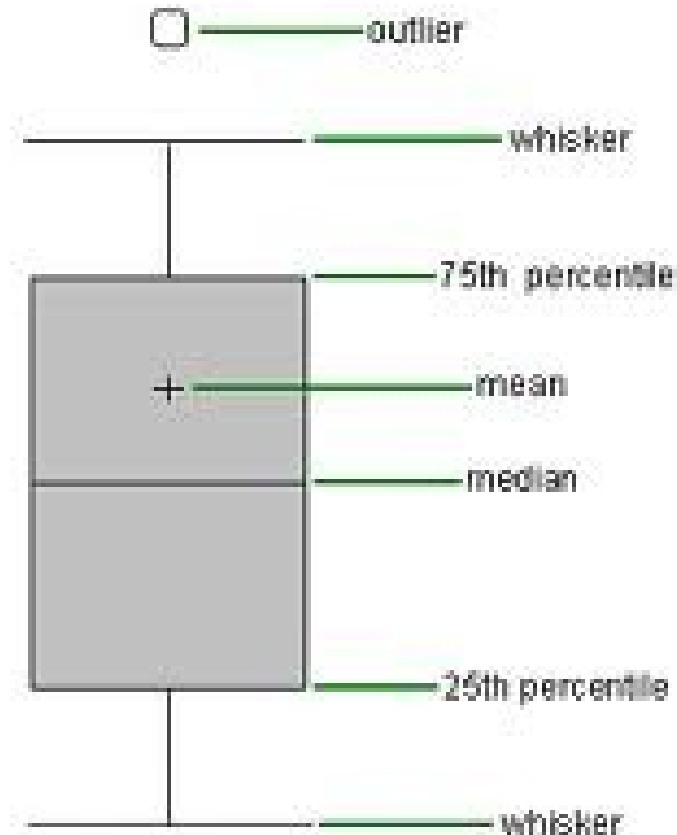
- U **nominálních proměnných**
- Nejčastěji obsazená kategorie/hodnota proměnné

MEDIÁN

- U **nominálních a ordinálních (pořadových) proměnných**
- Nejméně 50 % hodnot je menších nebo rovno mediánu a nejméně 50 % hodnot je větších nebo rovných mediánu
- Není ovlivněn extrémními hodnotami
- Pokud má soubor sudý počet prvků, dvě varianty (rozdílný výklad):
 - za medián označuje aritmetický průměr hodnot na místech $n/2$ a $n/2+1$
 - Medián nelze určit

Charakteristiky rozložení proměnné: modus, medián, průměr

- Medián = kvantil
- Kvartil
- Decil
- Percentil



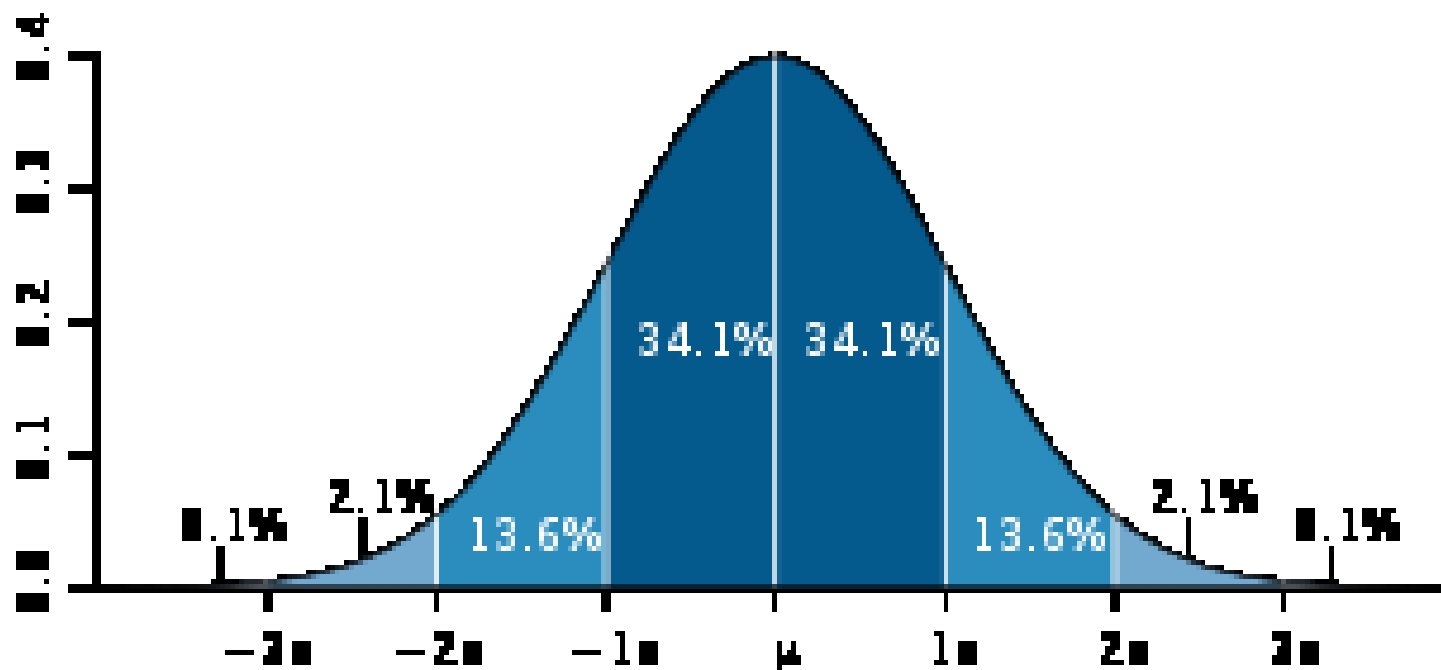
Charakteristiky rozložení proměnné: modus, medián, průměr

ARITMETICKÝ PRŮMĚR

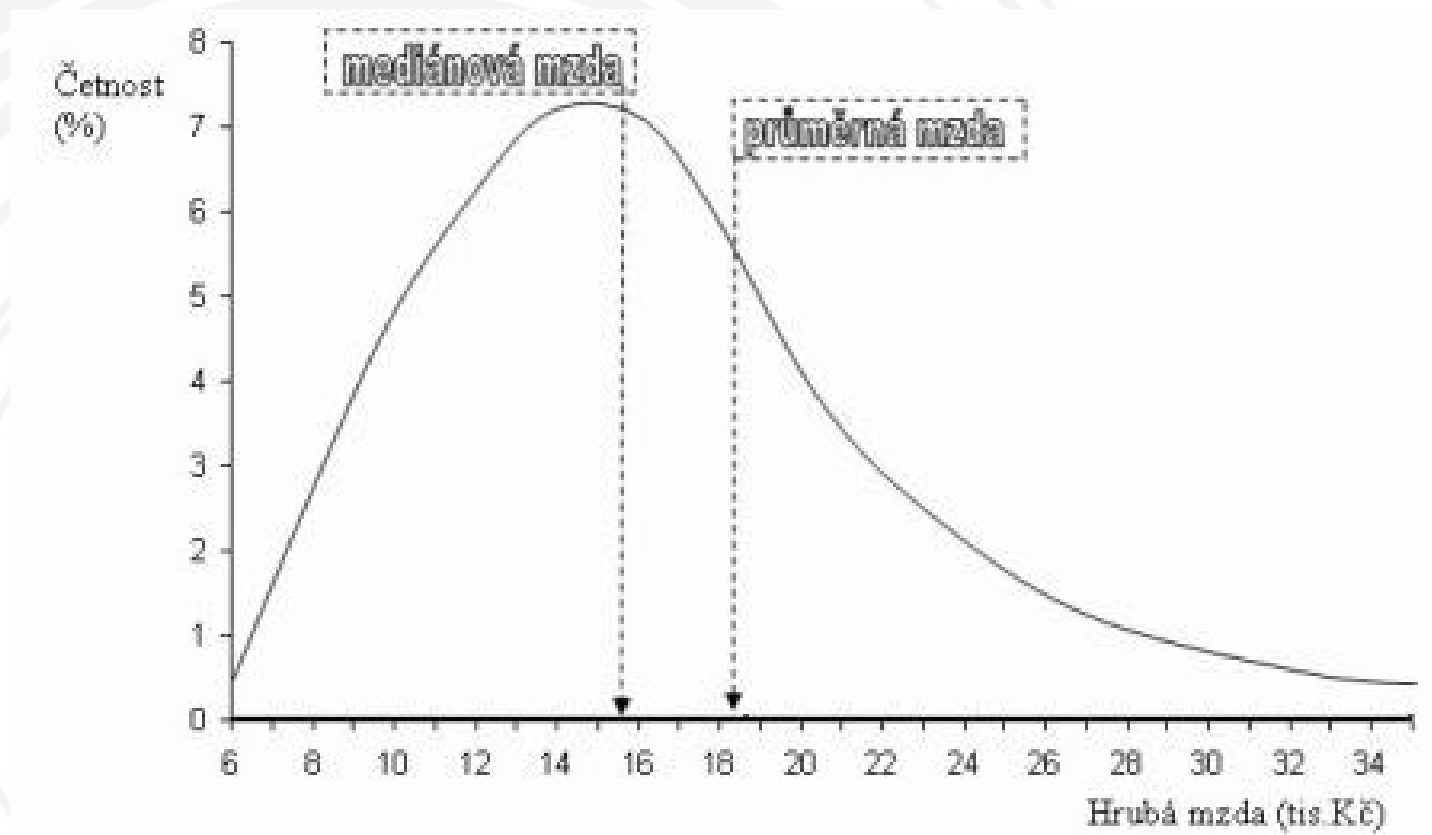
- Citlivý na extrémní hodnoty
- Aplikovatelná jen u **kardinálních znaků**
- Často udává hodnotu, která se v souboru vůbec nevyskytuje
 - (průměrný Čech navštíví knihovnu 1,12krát za rok)
- Kardinální znaky – nemá cenu vytvářet frekvenční tabulku nebo klasické grafy – využívá se **histogram**

Normální rozložení

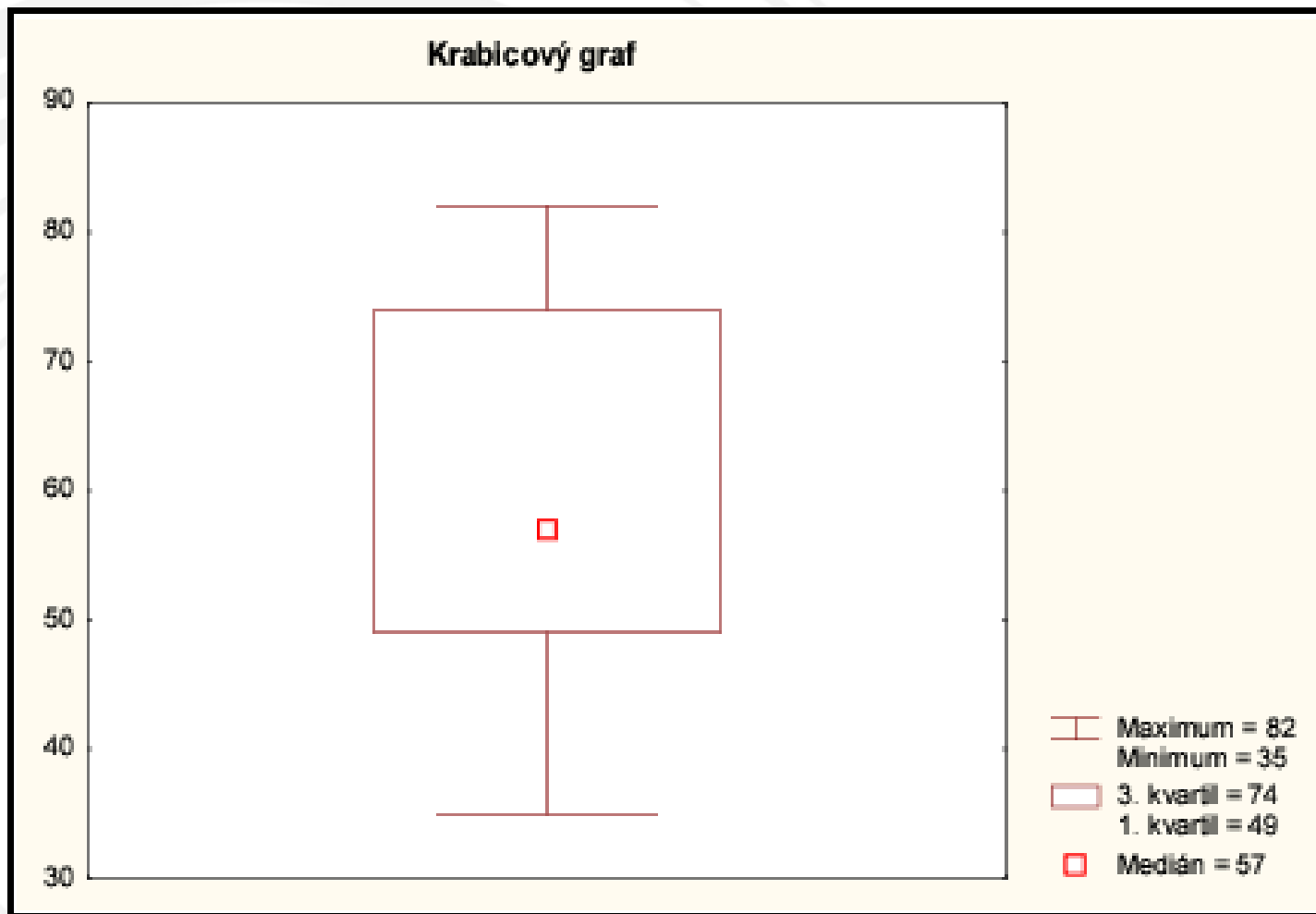
- Normální rozložení: modus = medián = průměr
- Asymetrie rozložení = šikmost



Ukázka šikmého rozložení



Ukázka šikmého rozložení



Rozložení u kardinálních dat

Rozpětí: rozdíl mezi nejmenší a nejvyšší hodnotou

Rozptyl: vypovídá o rozložení hodnot kolem aritmetického průměru

(průměrná čtvercová chyba (ve čtvercích jednotek původní proměnné) – součet druhých mocnin odchylek všech jednotlivých hodnot od průměru dělený rozsahem souboru

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Směrodatná odchylka:

- Druhá odmocnina rozptylu
- ukazuje homogenitu/variabilitu souboru
- čím menší SO je, tím více můžeme věřit aritmetickému průměru

Průměr a standardní odchylka

Zajímavost předmětu	není vůbec zajímavý	..*** (X) ***..	je velmi zajímavý
Přínosnost předmětu	není vůbec přínosné	*** X(*) **..	je velmi přínosné
Obtížnost obsahu	velmi snadný (.) ** X **	velmi obtížný
Náročnost na přípravu	velmi snadný (.) ** X **	velmi obtížný
Dostupnost studijních zdrojů	velmi špatně dostupné (.) ... X	velmi dobře dostupné
Jak učitel učí	velmi špatný	..*** (X) ***..	vynikající
Učitel jako odborník	není odborníkem (.) ... X	je odborníkem

Zajímavost předmětu	není vůbec zajímavý (.) ** X **	je velmi zajímavý
Přínosnost předmětu	není vůbec přínosné (.) ... X *	je velmi přínosné
Obtížnost obsahu	velmi snadný (*) ** X **	velmi obtížný
Náročnost na přípravu	velmi snadný (.) ** X **	velmi obtížný
Dostupnost studijních zdrojů	velmi špatně dostupné (.) ** X **	velmi dobře dostupné
Jak učitel učí	velmi špatný (.) ... X *	vynikající
Učitel jako odborník	není odborníkem (.) ... X *	je odborníkem

Transformace dat a proměnných

- Kategorizace spojitých proměnných (CATEGORIZE) → vytvoření intervalů

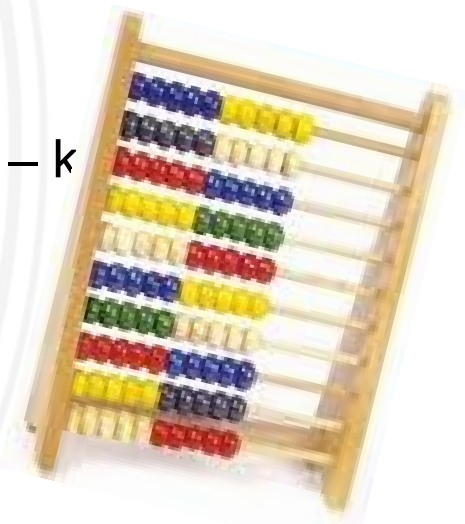
Otázka: Proč je důležité rekatégorizovat proměnné?

- Slučování kategorií (spíše spokojen – velice spokojen = spokojen)

Otázka: Kdy je vhodné slučovat proměnné?

Otázka: Lze slučovat i nominální proměnné?

- COUNT – vytváří novou proměnnou (pro sady otázek – k nabízených možností respondent zvolil)



Připomeňme si...

Hypotéza ➡ proměnné ➡ otázky v dotazníku

Hypotéza: Lidé s vyšším vzděláním navštěvují knihovny častěji, než lidé s nižším vzděláním.

Proměnné: vzdělání, frekvence návštěv knihovny

Otázky:

- *Jaké je Vaše nejvyšší ukončené vzdělání?*
- *Jak často navštěvujete knihovnu?*

Třídění druhého stupně

- Porovnání rozložení znaku v podsouborech populace (dle jiného znaku)
- Hypotézy nás vedou v tom, jaké vlastnosti a jejich souvislosti sledovat

Kdy to má smysl:

- Jedná-li se o reprezentativní výběrový soubor (ideálně náhodný výběr)
- Jde-li o nezávislý výběr

Jak statistika vypovídá o základním souboru?

Hlavní roli hraje **směrodatná odchylna / výběrová chyba**:

S 95% jistotou (5% riziko chyby) můžeme tvrdit, že:
průměr základního souboru (parametr)
=
průměr výběrového souboru (statistika)
 ± 2 směrodatné chyby

S 99% jistotou (1% riziko chyby) můžeme tvrdit, že:
průměr základního souboru (parametr)
=
průměr výběrového souboru (statistika)
 ± 3 směrodatné chyby

Statistické testování hypotéz

1. *Testování nulové hypotézy o neexistenci vztahu mezi proměnnými*
2. *Hypotéza zamítnuta → testování alternativní hypotézy*

Příklad nulové hypotézy:

Rozložení četností hodnot proměnné (vlastností jednotky), např. příjmu, věku, míry anomie, spokojenosti v životě (atd.) ve výběrovém souboru odpovídá rozložení proměnné v populaci.

Mezi vzděláním a výší příjmu není žádný vztah.

Testy pro statistické testování nulových hypotéz

- T-test o shodě dvou průměrů (parametrický test)
- Man-Whitney test (neparametrický test)

Zlaté pravidlo pro induktivní statistiku:

vysoká hodnota testu signifikance (tj. $\alpha > 0,05$) → držíme nulovou hypotézu

nízká hodnota testu signifikance (tj. $\alpha \leq 0,05$) → zamítáme nulovou hypotézu

Porovnávání průměrů

Report

Jsem spokojen/a s tematickou šíří nabídky povinně volitelných (B) kurzů.

V jakém stupni stu...	Mean	N	Std. Deviation
bakalářské studium	2,42	229	,821
magisterské studium	2,37	210	,766
Total	2,39	439	,795

Směrodatná odchylka u normálního rozložení:

- 68 % případů < 1 směrodatná odchylka
- 95 % případů < 2 směrodatné odchylky
- 99 % případů < 3 směrodatné odchylky

Tabulky rozložení

Vnímáte studium na KISK jako přínosné? * V jaké formě a stupni studia? Crosstabulation

			V jaké formě a stupni studia?				Total
			bakalářské prezenční	bakalářské kombinované	magisterské prezenční	magisterské kombinované	
Vnímáte studium na KISK jako přínosné?	velmi přínosné	Count	59	56	42	56	213
		% within V jaké formě a stupni studia?	45,7%	53,8%	41,2%	51,9%	48,1%
	spíše přínosné	Count	70	48	60	52	230
		% within V jaké formě a stupni studia?	54,3%	46,2%	58,8%	48,1%	51,9%
Total		Count	129	104	102	108	443
		% within V jaké formě a stupni studia?	100,0%	100,0%	100,0%	100,0%	100,0%

Grafy

Máte pocit, že při studiu můžete uplatnit to, co umíte nejlépe, nebo to, co vás nejvíce baví?

