Popisná statistika

Komentované řešení pomocí MS Excel

Vstupní data

	A	B	C
1	**Student(ka)**	**Body z 1. PP**	**Body z 2.PP**
2	1	43	65
3	2	37	43
4	3	27	7
5	4	33	70
6	5	41	19
7	6	55	34
760	759	36	37
761	760	59	58
762	761	89	88
763	762	44	28

- Máme k dispozici data o počtech bodů z 1. a 2. zápočtového testu z Matematiky I v zimním semestru 2015/2016 a to za všech 762 studentů, kteří psali oba testy.
- Ukázka části tabulky se vstupními daty viz. obrázek vlevo.
- V dalším budeme předpokládat, že tabulka se vstupními daty je umístěna v oblasti A1:C7633 (viz. obrázek) .

Vybrané kvantily

- Kvantily rozdělují uspořádaný soubor hodnot na dvě části dle předem zadaného poměru četností.
- Vzorce pro výpočet kvantilů pro 1.PP (sloupec B) jsou na obrázku vpravo, překopírováním do sloupce C dostaneme kvantily i pro 2.PP. Tímto způsobem budeme prezentovat vzorce a vypočítané hodnoty i na ostatních snímcích.

	A	B	C
765	**Vybrané kvantily**	**1. PP**	**2.PP**
766	Minimum	0	0
767	První decil ($x_{0,1}$)	16	8
768	Dolní kvartil ($x_{0,25}$)	30	21
769	Medián ($x_{0,5}$)	45	39
770	Horní kvartil ($x_{0,25}$)	65	56
771	Poslední decil ($x_{0,9}$)	78,9	71
772	Maximum	100	100

	B
765	**1. PP**
766	=MIN(B$2:B$763)
767	=PERCENTIL(B$2:B$763;0,1)
768	=QUARTIL(B$2:B$763;1)
769	=MEDIAN(B$2:B$763)
770	=QUARTIL(B$2:B$763;3)
771	=PERCENTIL(B$2:B$763;0,9)
772	=MAX(B$2:B$763)

- Kvantily použijeme jako vstupy pro výpočet různých charakteristik polohy a variability a pro konstrukci krabicového grafu (viz. další snímky).
- Kvantily nám však samy o sobě pomohou získat představu o poloze a variabilitě dat. Vidíme, že z druhého testu studenti dosahovali spíše méně bodů než v prvním testu, variabilita bodových výsledků je v obou testech srovnatelná. Lépe to však bude vidět na charakteristikách polohy a variability na dalších snímcích.

Míry polohy

- Pro posouzení polohy (středního množství počtu bodů) použijeme dva ukazatele
  - Aritmetický průměr (patří do třídy momentových charakteristik)
  - Medián (kvantilová charakteristika)

	A	B	C
775	**Míry polohy**	**1. PP**	**2.PP**
776	Prumer	46,57	39,45
777	Median	45	39

	B
775	**1. PP**
776	=PRŮMĚR(B$2:B$763)
777	=MEDIAN(B$2:B$763)

- Z hodnot průměrů i mediánů je patrné, že ve druhém testu dosahovali studenti spíše méně bodů.
- Střední počet bodů je u obou testů mírně pod 50 body.

Míry variability

- Variabilitu (měnlivost, míru rozptýlení) posoudíme pomocí těchto ukazatelů
  - Rozptyl (momentová charakteristika),
  - Směrodatná odchylka (momentová charakteristika),
  - Variační rozpětí (kvantilová charakteristika),
  - Decilové rozpětí (kvantilová charakteristika),
  - Kvartilové rozpětí (kvantilová charakteristika),
  - Variační koeficient (relativní míra variability.

	A	B	C
779	**Míry variability**	**1. PP**	**2.PP**
780	Rozptyl	539,0	552,9
781	Směrodatná odchylka	23,2	23,5
782	Variační rozpětí	100	100
783	Decilové rozpětí	62,9	63
784	Kvartilové rozpětí	35	35
785	Variační koeficient	0,50	0,60

	B
779	**1. PP**
780	=VAR(B$2:B$763)
781	=SMODCH(B$2:B$763)
782	=B$772-B$766
783	=B$771-B$767
784	=B$770-B$768
785	=B781/B776

- Momentové i kvantilové charakteristiky ukazují, že (absolutní) variabilita počtu bodů v obou testech je velmi podobná.
- V průměru se počet dosažených bodů od celkového průměru liší o cca 23 bodů (směrodatná odchylka).
- Bodové výsledky 80 % studentů se nachází v intervalu šířky 63 bodů (decilové rozpětí), počty bodů 50 % studentů jsou koncentrovány v intervalu šířky 35 bodů (kvartilové rozpětí).
- Relativní variabilita (směrodatná odchylka vztažena ku průměrnému počtu bodů) je u prvního testu trochu vyšší než u druhého testu, a to kvůli vyššímu průměru.

Míra (lineární) závislosti

- Sílu lineární závislosti posoudíme pomocí Pearsonova korelačního koeficientu, který je odvozen od druhých momentů obou proměnných

	A	B
787	**Míry linearní závislosti**	
788	Korelační koeficient	0,68

	B
788	=CORREL($B$2:$B$763;$C$2:$C$763)

- Kladná hodnota korelačního koeficientu svědčí o pozitivní lineární závislosti mezi počty bodů z prvního a druhého testu. Měl-li student nadprůměrný počet bodů v prvním testu, dá se očekávat, že měl i nadprůměrný počet bodů v druhém testu (a naopak).
- Nelze ovšem říci, že studenti se ve druhém testu nezhoršovali. Vzhledem k tomu, že průměrný počet bodů ve druhém testu je nižší než v prvním, mohli se studenti zhoršit i v případě pozitivní závislosti. Ta se totiž týká porovnání počtu bodů s průměrem.
- Posuzování intenzity závislosti podle velikosti korelačního koeficientu je arbitrární a v různých úlohách se může lišit. Často se však užívá obecné, empirické, pravidlo (viz. níže). Podle tohoto pravidla svědčí hodnota korelačního koeficientu 0,68 o relativně silné pozitivní závislosti, ne však perfektní závislosti. Znamená to tedy, že v některých případech se mohli studenti zhoršit i v porovnání s ostatními studenty (přesněji v porovnání s průměrem). Těchto případů však nebude mnoho.
- Síla i tvar závislosti dobře ilustruje bodový graf , který uvádíme na zvláštním snímku (viz. poslední snímek).

	E	F
787	**Hodnota korelačního koeficientu**	**Interpretace**
788	-1	perfektní negativní lineární závislost
789	-0,7	silná negativní lineární závislost
790	-0,5	mírná negativní lineární závislost
791	-0,3	slabá negativní lineární závislost
792	0	žádná lineární závislost
793	0,3	slabá pozitivní lineární závislost
794	0,5	mírná pozitivní lineární závislost
795	0,7	silná pozitivní lineární závislost
796	1	perfektní pozitivní lineární závislost

Krabicový graf (box plot) 1

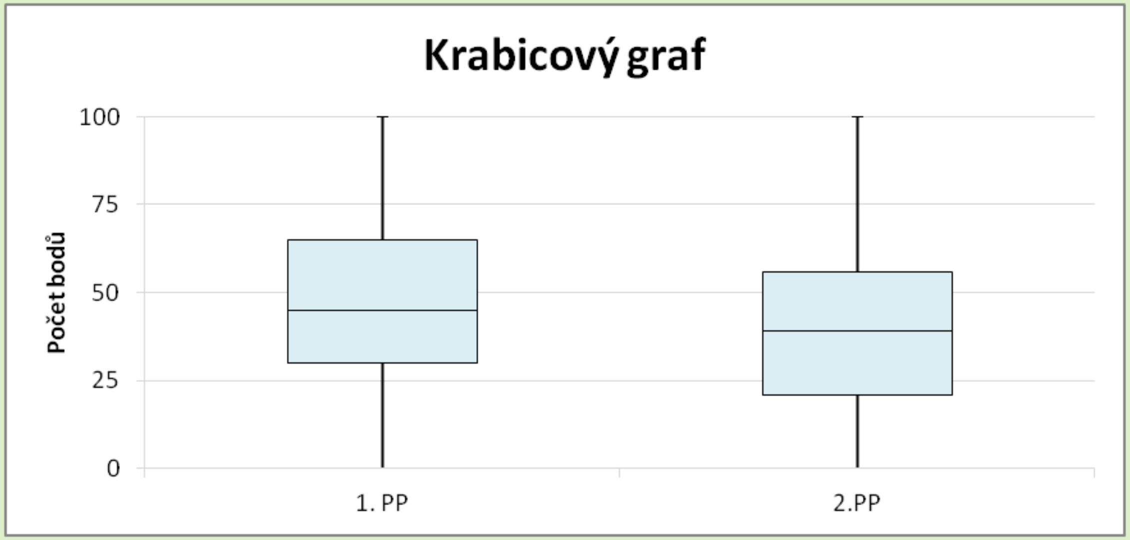
Postup:

- Excel nenabízí krabicový graf v základní nabídce grafů. Krabicový graf ale můžeme vytvořit vhodným nakombinováním sloupcových grafů.
- Nejdříve si spočítáme pomocné hodnoty pro tvorbu grafu (viz. vedlejší tabulka).
- Označíme oblast B800:C803 → karta „Vložení“ → „Sloupcový“ (panel „Grafy“) → „Skládaný sloupcový“.
- Karta „Návrh“ → „Přepnout řádek či sloupec“.
- Nyní přidáme dolní fousy: označíme v grafu spodní sloupce („Skrytý sloupec“) → karta „Rozložení“→“Chybové úsečky“ →“Další možnosti chybových úseček“→ „Svislé chybové úsečky“→ vybereme Směr: „Minus“ a Typ chybové hodnoty: „Vlastní“ a stiskneme „Zadat hodnotu“ → do pole „Záporná chybová hodnota vložíme oblast s délkou dolního fousu B808:C808.
- Podobně se přidají horní fousy: v grafu se označí horní sloupce („sloupec nad mediánem“) a dále se postupuje analogicky, akorát se zadá Směr: „Plus“ a do pole „Kladná chybová hodnota“ se zadá oblast s délkou horního fousu B809:C809.
- Označíme spodní sloupce → „Formát datové řady“→ Výplň: „Bez výplně“.
- Podobným způsobem upravíme výplně ostatních dvou sloupců (zadáme stejnou barvu) a přidáme ohraničení, zadáme Barva ohraničení: „Plná čára“ a Barva: Černá.
- Pro větší přehlednost:
  - Odstraníme legendu,
  - Upravíme Minimum, Maximum a Hlavní jednotku,
  - Přidáme název grafu a název svislé osy.

	A	B	C
800	**Krabicový graf**	**1. PP**	**2.PP**
801	skrytý sloupec	30	21
802	sloupec pod mediánem	15	18
803	sloupec nad mediánem	20	17
804	mez pro dolní fous	-23	-32
805	mez pro horní fous	118	109
806	hranice dolního fousu	0	0
807	hranice horního fousu	100	100
808	délka dolního fousu	30	21
809	délka horního fousu	35	44

	B
800	**1. PP**
801	=B768
802	=+B769-B768
803	=B770-B769
804	=B768-1,5*B784
805	=B770+1,5*B784
806	=+B766
807	=B772
808	=+B801-B806
809	=B807-SUMA(B801:B803)

Krabicový graf (box plot) 2



Interpretace výsledků

- Krabicové grafy potvrzují naše předchozí úsudky na základě spočítaných charakteristik – v prvním testu dosahovali studenti o trochu více bodů, variabilita je v obou testech srovnatelná
- Došlo také k malé změně rozdělení počtu bodů (viz. posun mediánu uvnitř krabice u druhého testu)
- Vzhledem k tomu, že počet bodů je shora omezen 100 a zespoda 0, nejsou v datech žádná odlehlá pozorování.

Histogram 1

Postup:

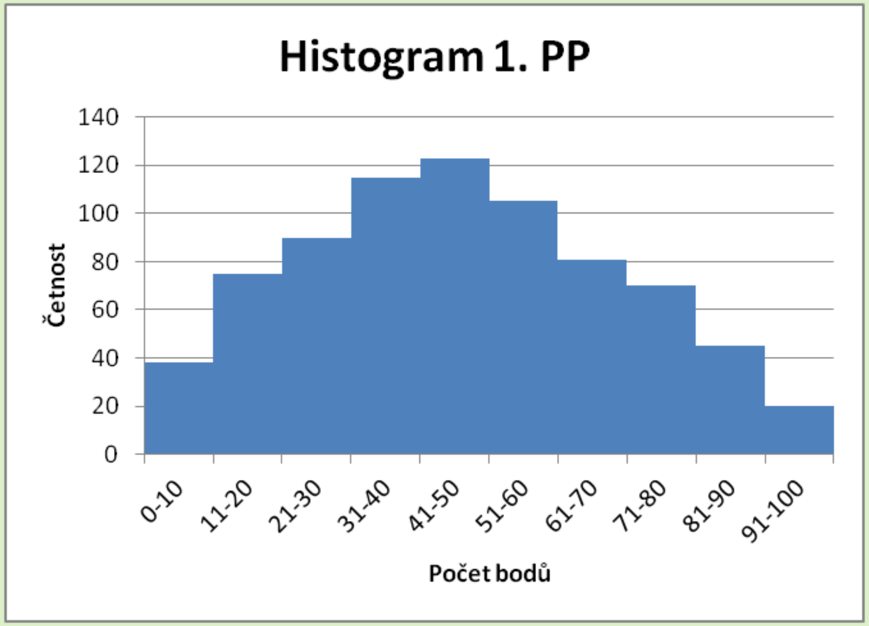
- Nejdříve musíme určit intervaly (arbitrárně nebo pomocí nějakého empirického pravidla). Pro přehlednost a díky dostatečnému počtu pozorování zvolíme intervaly šířky 10 bodů.
- Dále spočítáme četnost pozorování v jednotlivých intervalech (viz. tabulka vpravo).

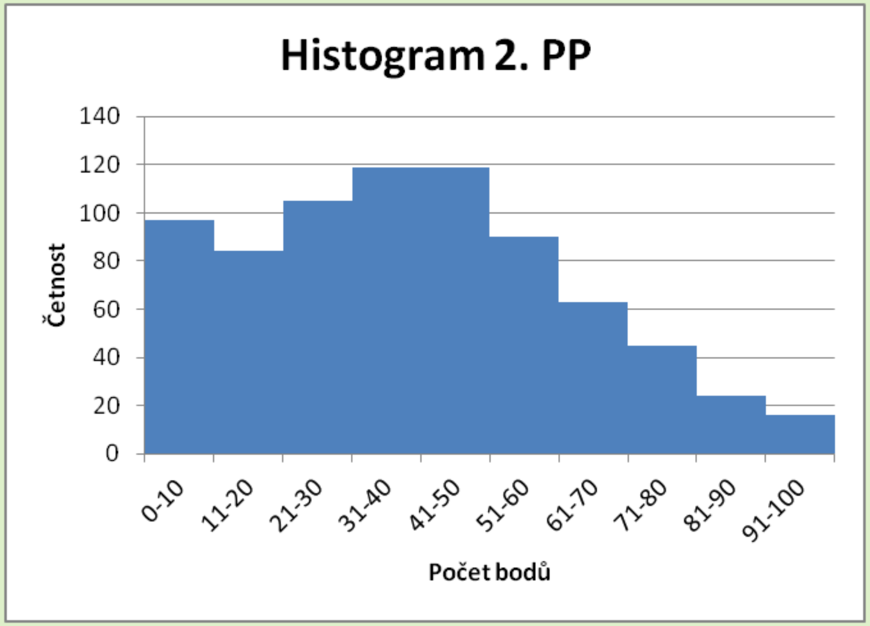
	A	B	C	D	E
812	**Histogram**				
813	**Interval**			**Četnost**	
814	**Dolní mez**	**Horní mez**	**Popis**	**1. PP**	**2.PP**
815	0	10	0-10	38	97
816	11	20	11-20	75	84
817	21	30	21-30	90	105
818	31	40	31-40	115	119
819	41	50	41-50	123	119
820	51	60	51-60	105	90
821	61	70	61-70	81	63
822	71	80	71-80	70	45
823	81	90	81-90	45	24
824	91	100	91-100	20	16
825	Celkem			762	762

	D
812	
813	**Četnost**
814	**1. PP**
815	=COUNTIFS(B$2:B$763;">="&$A815;B$2:B$763;"<="&$B815)
816	=COUNTIFS(B$2:B$763;">="&$A816;B$2:B$763;"<="&$B816)
817	=COUNTIFS(B$2:B$763;">="&$A817;B$2:B$763;"<="&$B817)
818	=COUNTIFS(B$2:B$763;">="&$A818;B$2:B$763;"<="&$B818)
819	=COUNTIFS(B$2:B$763;">="&$A819;B$2:B$763;"<="&$B819)
820	=COUNTIFS(B$2:B$763;">="&$A820;B$2:B$763;"<="&$B820)
821	=COUNTIFS(B$2:B$763;">="&$A821;B$2:B$763;"<="&$B821)
822	=COUNTIFS(B$2:B$763;">="&$A822;B$2:B$763;"<="&$B822)
823	=COUNTIFS(B$2:B$763;">="&$A823;B$2:B$763;"<="&$B823)
824	=COUNTIFS(B$2:B$763;">="&$A824;B$2:B$763;"<="&$B824)
825	=SUMA(D815:D824)

- Pro histogram bodů z prvního testu označíme oblast D814:D824 → karta „Vložení“ → „Sloupcový“ (panel „Grafy“) → „Skupinový sloupcový“.
- Označíme sloupce v grafu (datovou řadu) → „Formát datové řady“→ v záložce Možnosti řady upravíme Šířka mezery: 0% (Bez mezery)
- Přidáme popisky x-ové osy: Karta „Návrh“→ „Vybrat data“ → ve sloupci Popisky vodorovné osy stiskneme „Upravit“→ vybereme oblast C815:C824
- Pro větší přehlednost odstraníme legendu, přidáme název grafu a názvy os
- Analogicky vytvoříme histogram pro druhý test

Histogram 2





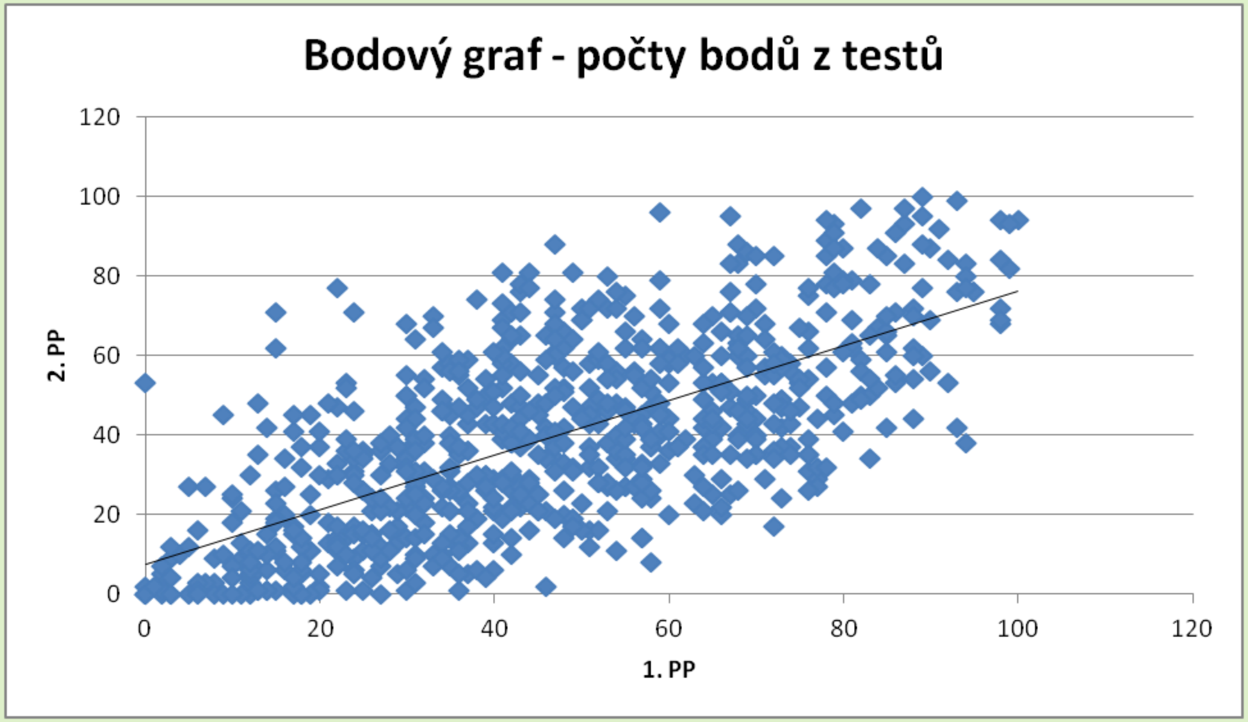
<u>Interpretace výsledků</u>

- Rozdělení počtu bodů v prvním testu je dosti symetrické kolem středu (lze pozorovat jen velmi mírné pozitivní zešikmení) a neliší se příliš od Gaussovy křivky.
- Rozdělení počtu bodů v druhém testu je více pozitivně zešikmené. Nalevo od střední hodnoty jsou body rozděleny mnohem rovnoměrněji než napravo, kde četnosti klesají podobně jako Gaussova křivka.
- Oproti prvnímu testu můžeme pozorovat u druhého testu výrazné navýšení počtu studentů s velmi malým počtem bodů (0-10), to je také jeden z hlavních důvodů poklesu průměrného počtu bodů.

Bodový graf

Postup

- Označíme oblast s individuálními daty B1:C763→ karta „Vložení“ → „Bodový“ (panel „Grafy“) → „Bodový pouze se značkami“.
- Označíme body na grafu (datovou řadu) → „Přidat spojnici trendu“→ vybereme Typ trendu a regrese: „Lineární“.
- Pro větší přehlednost odstraníme legendu, přidáme název grafu a názvy os.



Interpretace výsledků

- Bodový graf potvrzuje existenci relativně silné pozitivní lineární závislosti mezi počty bodů z prvního a druhého testu. Závislost však není perfektní – někteří studenti se dosti zlepšili ve druhém testu (body vlevo nahoře), někteří zase zhoršili (body vlevo dole). Většina studentů je však soustředěna kolem regresní přímky.