KORELACE

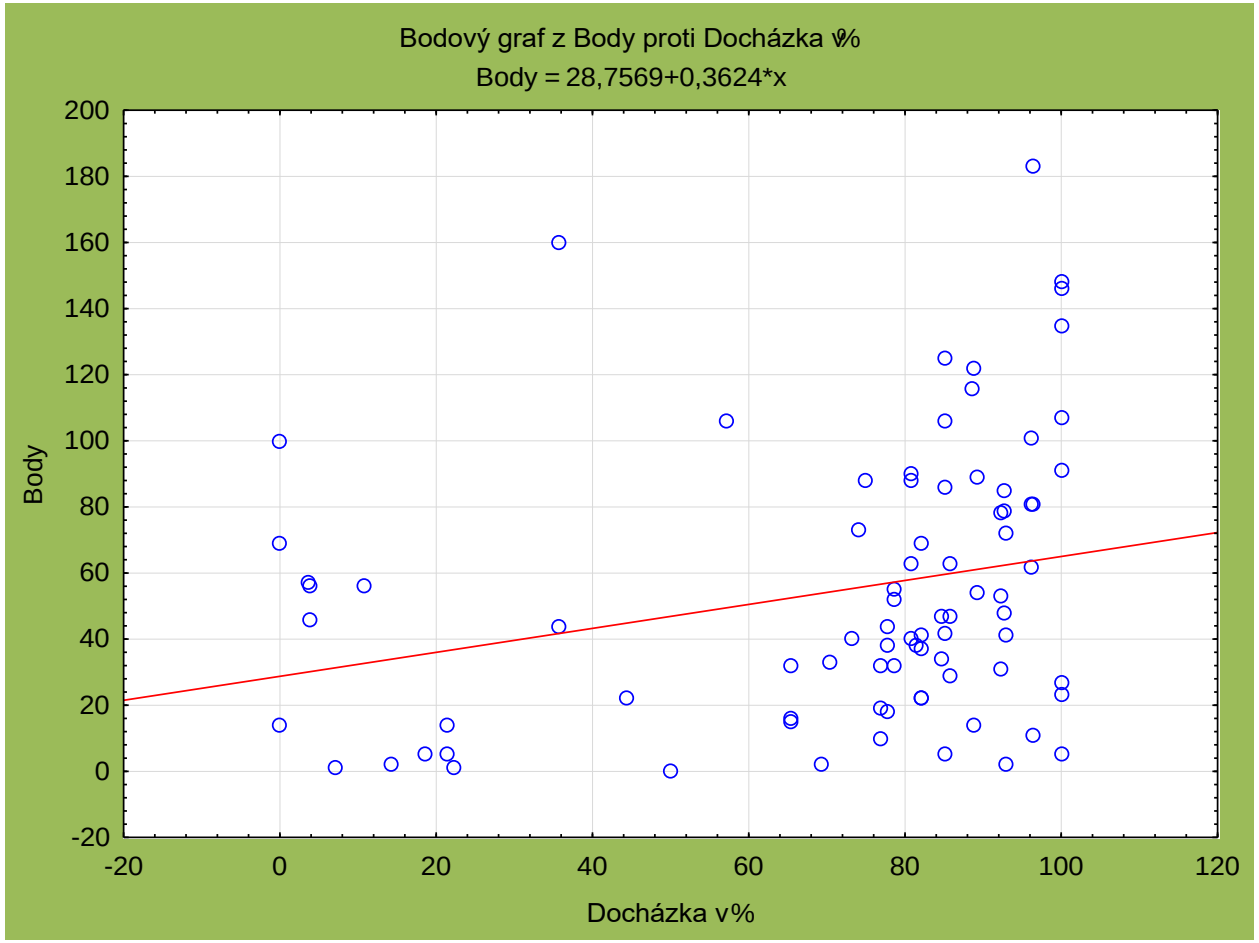
Komentované řešení pomocí programu *Statistica*

Vstupní data I

	1 Docházka v %	2 Body
1	70,4	33
2	100	135
3	92,6	79
4	85,2	42
5	100	27
6	88,9	14
7	96,3	183
8	85,2	5
9	96,3	11
10	85,2	86
11	92,6	85
12	100	146
13	77,8	18
14	44,4	22
15	77,8	38
16	88,9	122
17	92,6	48
18	18,5	5
19	81,5	38
20	96,3	81
21	85,2	125
22	85,2	106
23	77,8	44
24	100	107
25	74,1	73
26	22,2	1

- Data umístěná v excelovském souboru překopírujeme do tabulky ve *Statistice* a pojmenujeme proměnné, viz prezentace k tématu **Popisná statistika**.
- Zajímá nás, zdali účast na cvičeních ovlivňuje výsledky zápočtových písemek (v součtu bodů za celý semestr).
- Prvotní představu o tvaru a síle závislosti docházky a počtu bodů nám poskytne bodový graf proložený grafem regresní funkce:
  - **Grafy** → **Bodové grafy** → **Proměnné** – na osu x zvolíme Docházku a na osu y Body → **OK** → **OK**
- Spočteme ještě koeficient determinace $R^2 = 0{,}0687$:
  - **Statistiky** → **Vícenásobná regrese** → **Proměnné** – nezáv. prom. je Docházka a záv. prom. je Body → **OK** → **OK**

Vstupní data II



- Z bodového grafu je patrné, že mezi docházkou a počtem bodů je pozitivní lineární závislost. Tato závislost je však dosti slabá, o čemž svědčí i velmi malý koeficient determinace $R^2 = 0{,}0687$. Pomocí docházky by se nám tedy podařilo vysvětlit pouze necelých 7 % variability počtu bodů. Proto jsou bodové rozdíly jednotlivých studentů z převážné části ovlivněny jinými faktory.
- Z bodového grafu je též patrné, že studenty lze rozdělit v zásadě do dvou skupin – na ty, kteří chodí pravidelně (účast nad 70 %), těch je většina, a na ty, kteří nechodí skoro vůbec (účast pod 20 %).

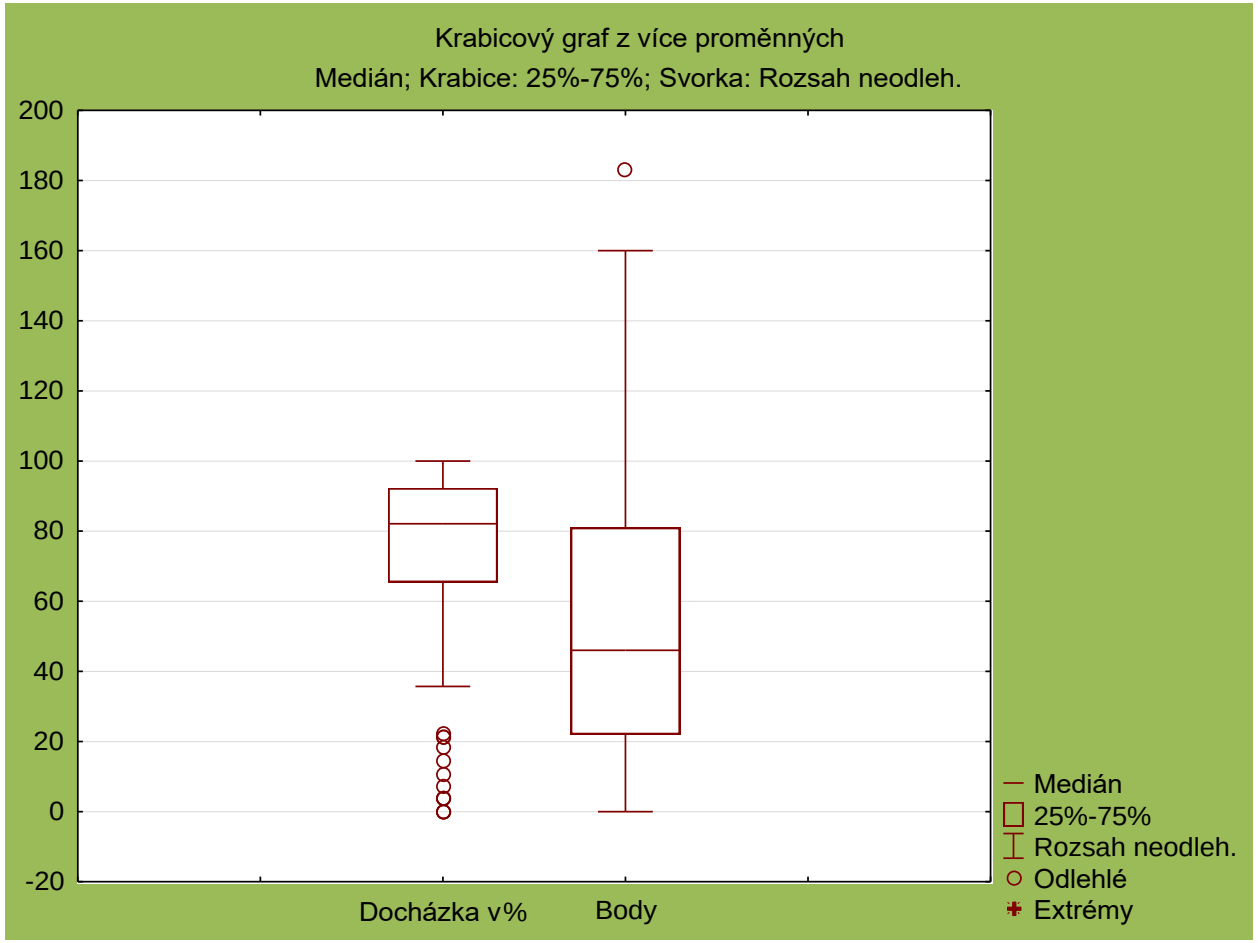
Základní statistiky

- Abychom získali základní představu o studovaných datech, spočítáme si základní charakteristiky dat
  - **Statistiky** → **Základní statistiky** → **Detailní výsledky** – zvolíme si, co nás zajímá → **Proměnné** → vybereme vše → **OK** → **Výpočet**

Proměnná	Popisné statistiky							
	N platných	Průměr	Medián	Minimum	Maximum	Dolní kvartil	Horní kvartil	Sm.odch.
Docházka v %	83	71,45663	82,10000	0,00	100,0000	65,40000	92,30000	29,95059
Body	83	54,65060	46,00000	0,00	183,0000	22,00000	81,00000	41,40621

- a dále např. krabicové grafy
  - **Grafy** → **Krabice** – zvolíme si úpravu, jakou si přejeme, např.

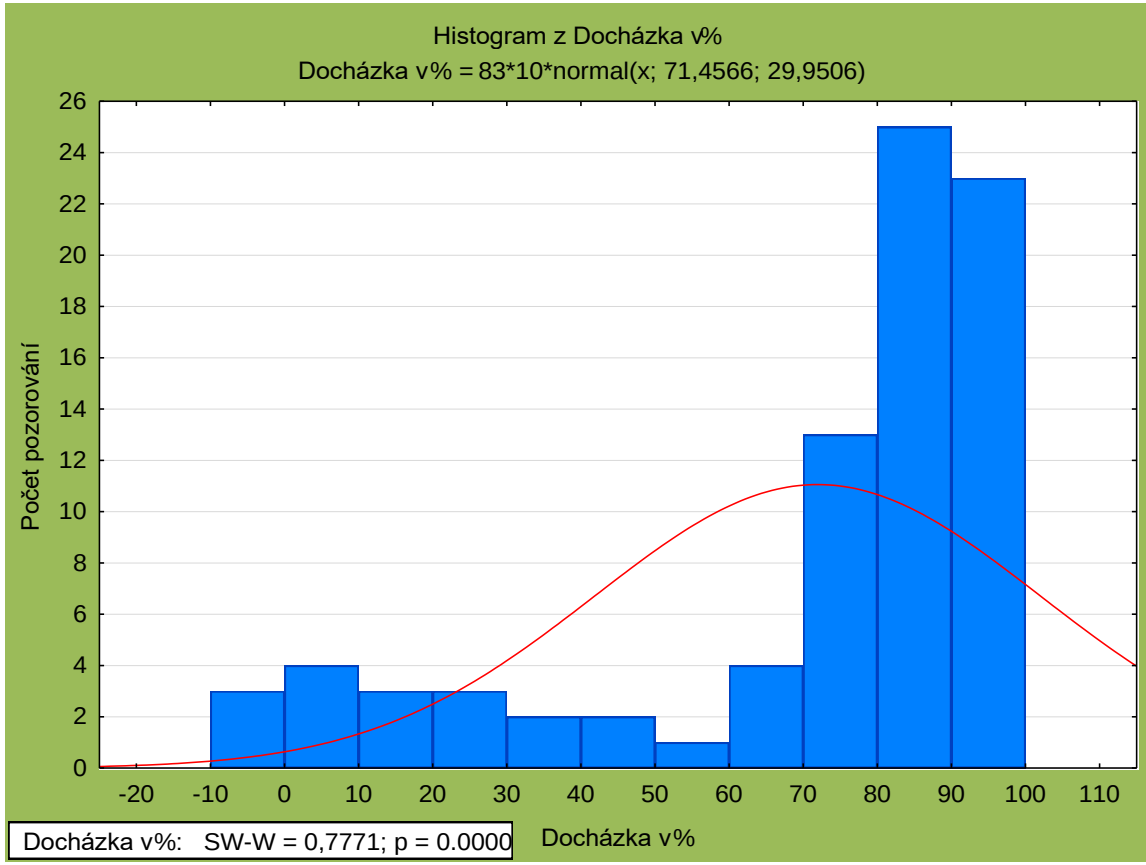
Krabicové grafy

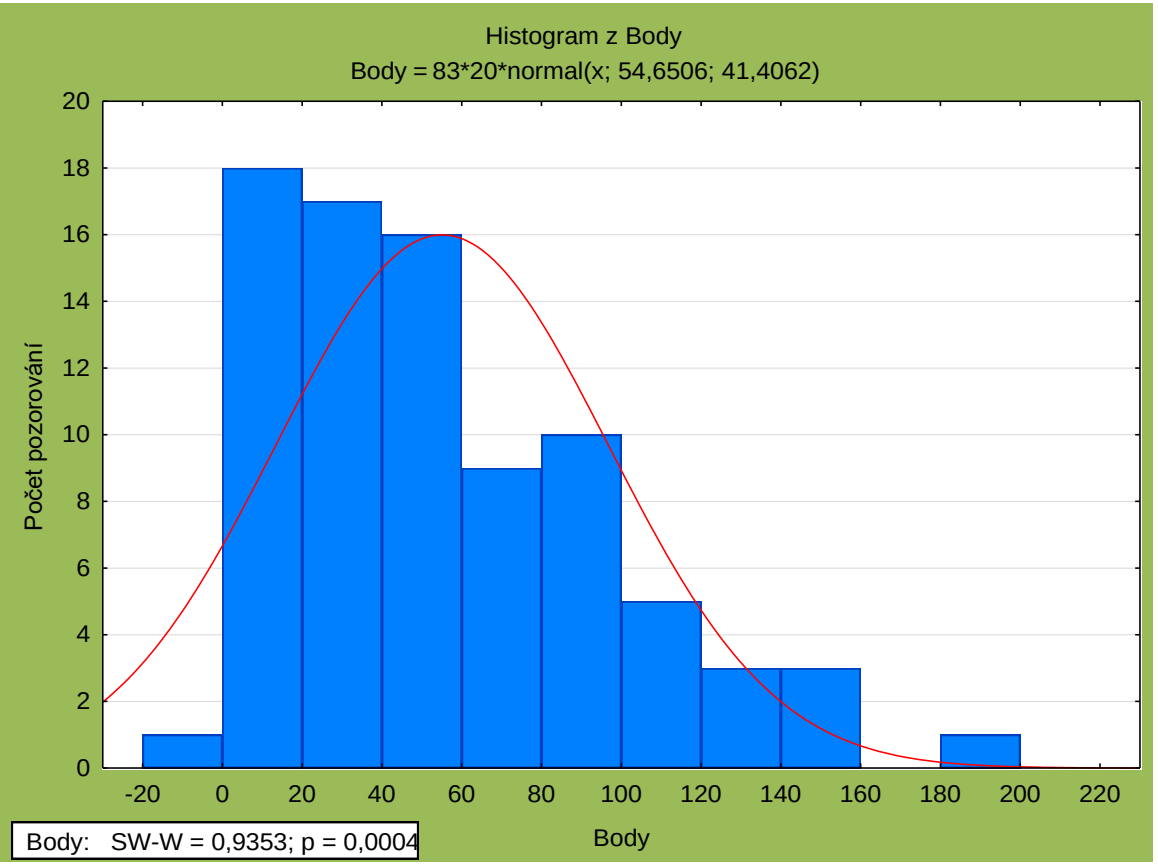


Grafické výstupy – histogram, Q – Q plot

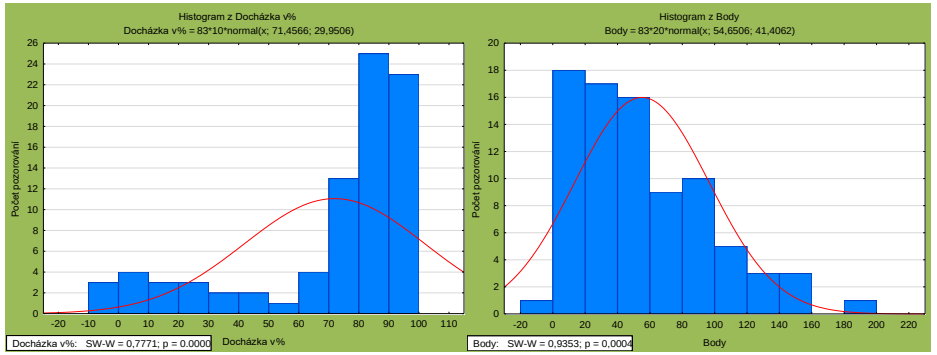
- Chceme-li použít test nulovosti korelačního koeficientu, musíme nejdříve ověřit normalitu obou proměnných (ve skutečnosti bychom měli ověřit dvourozměrnou normalitu, to je však obtížné). K ověření (jednorozměrné) normality použijeme histogram a Q – Q plot.
- **Grafy → Histogram**
  - nastavení **Proměnné** Docházka → **OK** → **OK**, stejně pro Body
  - dále **Detaily → Shapiro – Wilkův test** (což je test normality dat), abychom se přesvědčili o našich úsudcích početně, nejen na základě grafických výstupů
- **Grafy → Grafy vstupních dat → Pravděpodobnostní graf** „Body" (či Docházka, podle toho, v které buňce tabulky s daty se nachází kurzor) → **Normál. pravděpodobnost**

Histogramy I





Histogramy II

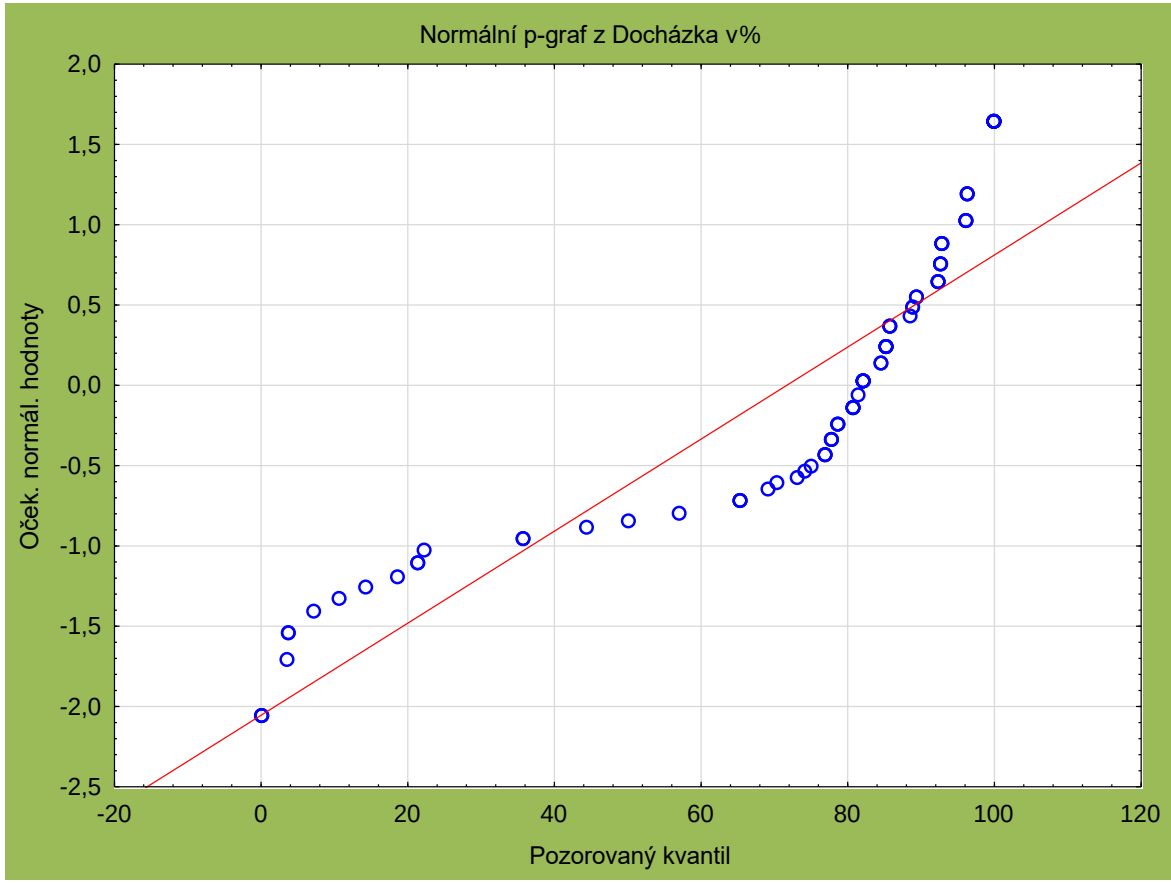


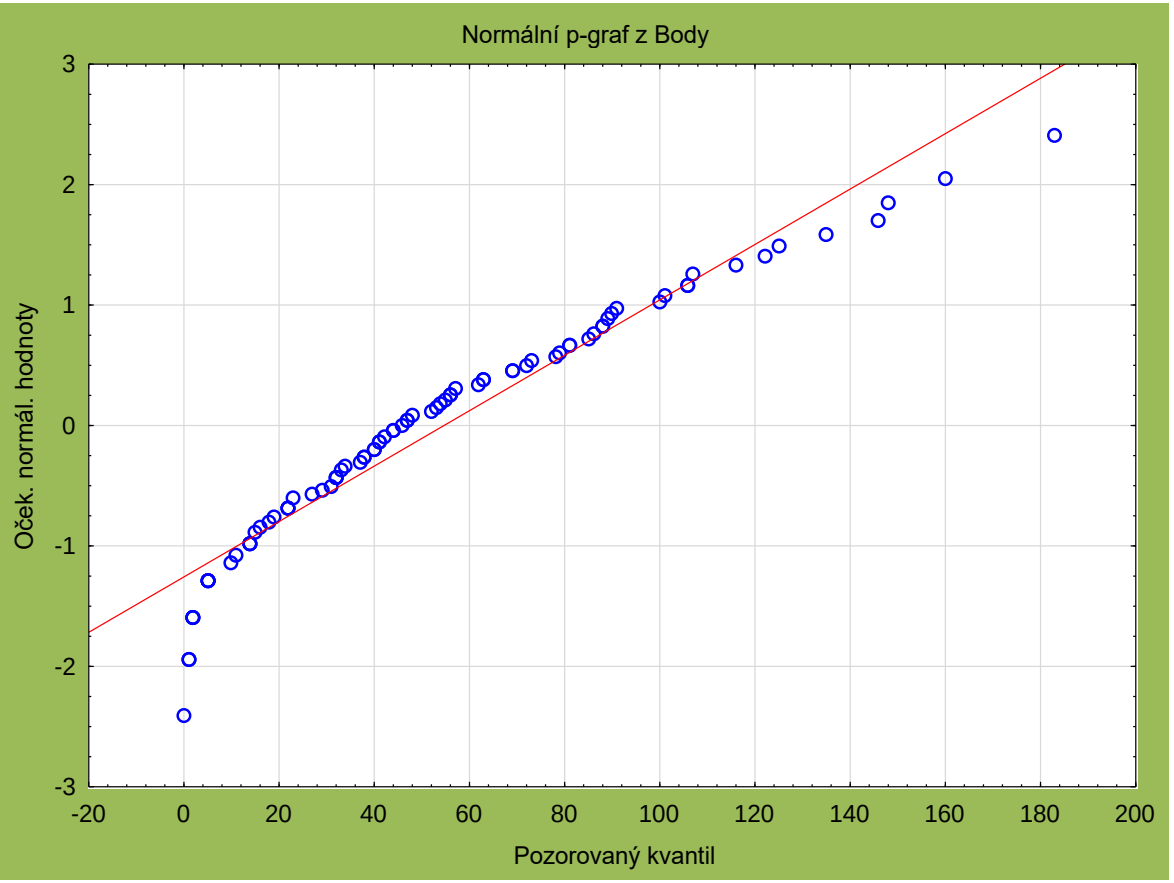
- **Docházka**
  - Z histogramu je patrné, že empirické rozdělení docházky se výrazně liší od normálního rozdělení (jeho hustota je vyznačena červeně).
  - Empirické rozdělení je **bimodální**, čímž se potvrdila naše domněnka o existenci dvou skupin studentů, s pravidelnou účastí a s minimální účastí.
  - Z výše uvedeného vyplývá, že **normalitu v případě docházky předpokládat nemůžeme**, což nám ostatně potvrzuje i *Shapirova – Wilkův test* (na výstupu bílý rámeček vlevo dole), jehož p – hodnota je blízká nule.

- **Body**
  - Ve srovnání s docházkou je rozdělení počtu bodů blíže normálnímu rozdělení. Ale i v případě počtu bodů jsou patrné jisté rozdíly oproti normálnímu rozdělení. Hlavní rozdíl je zjevné zešikmení empirického rozdělení, zatímco normální rozdělení je symetrické.
  - *Shapirův – Wilkův test* normalitu bodů zamítá na hladině 0,04 %.

Q – Q plot

- Q – Q grafy potvrzují naše předchozí zjištění. Rozdělení docházky se výrazně liší od normálního.
- Rozdělení bodů se více podobá normálnímu rozdělení, ale i tak je zde patrný systematický rozdíl.





Asymptotický test nezávislosti I

- Předpoklad normality tedy přijmout nemůžeme a nadále musíme s testem nulovosti korelačního koeficientu pracovat jako s asymptotickým testem. Výsledky pak musíme interpretovat opatrněji.
- **Statistiky → Základní statistiky → Korelační matice → 1 seznam proměn. → Vybrat vše → OK → Výpočet**

Korelace
Označ. korelace jsou významné na hlad. p < ,05000
N=83 (Celé případy vynechány u ChD)

Proměnná	**Průměry**	Sm.odch.	Docházka v %	Body
Docházka v %	71,45663	29,95059	1,000000	0,262115
Body	54,65060	41,40621	0,262115	1,000000

- Chceme-li znát p – hodnota testu, v **Možnosti** zvolíme **Zobrazit r, p-hodnoty a N**

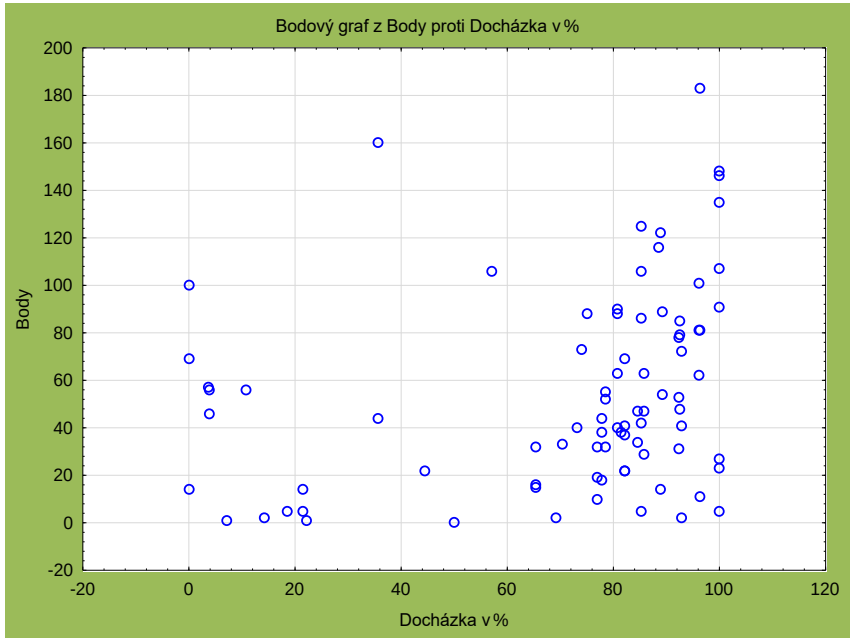
Korelace
Označ. korelace jsou významné na hlad. p < ,05000
N=83 (Celé případy vynechány u ChD)

Proměnná	**Docházka v %**	Body
Docházka v %	1,0000	,2621
	p= ---	p=,017
Body	,2621	1,0000
	p=,017	p= ---

Asymptotický test nezávislosti II

- **Úloha A)**
  - Na základě p – hodnoty 0,017 můžeme zamítnout na hladině významnosti 5 % nulovou hypotézu o nezávislosti ve prospěch oboustranné alternativy. **Prokázali jsme existenci lineární závislosti**. Připomeňme ovšem, že na hladině významnosti 1 % už bychom nulovou hypotézu o nezávislosti nezamítali. Navíc spočítaná p – hodnota je jen přibližná, protože se jedná o asymptotický test. Závěr o existenci lineární závislosti je tedy na hranici prokazatelnosti.
- **Úloha B)**
  - V případě jednostranné alternativy je p – hodnota 0,0085, což je polovina p – hodnoty pro oboustrannou alternativu. Tvrzení o **existenci pozitivní lineární závislosti** docházky a počtu bodů je tedy **prokázáno i na hladině významnosti 1 %.** I v tomto případě se však jedná pouze o přibližnou p – hodnotu. Prokázali jsme tedy, že dobrá docházka má pozitivní vliv na výsledky studentů v testech. Tento vliv je však relativně slabý.

Spermanův test nezávislosti



- Jak je vidět, data nevypadají, že by byla lineárně závislá. Přesto bychom přirozeně čekali, že by vysoká docházka mohla mít pozitivní vliv na výsledky testu. Neboli, že vztah mezi body a docházkou je (až na náhodný faktor) rostoucí funkce. Tímto monotónním typem závislosti se zabývá **Spearmanův korelační koeficient**.
- **Statistiky → Neparametrické statistiky → Korelace → Vytvořit – Detailní report → Proměnné – 1. seznam:** Docházka, **2. seznam:** Body **→ OK → Spearman. R**

Dvojice proměnných	Spearmanovy korelace ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000 **Počet plat.**	Spearman R	t(N-2)	p-hodn.
Docházka v % & Body	83	0,345707	3,315812	0,001369

Ježto je p – hodnota testu menší než 1 %, je to pro nás dostačující údaj, abychom zamítli nekorelovanost ve prospěch alternativy, že vztah mezi body a docházkou je rostoucí funkce.