## Korelační analýza – řešení.

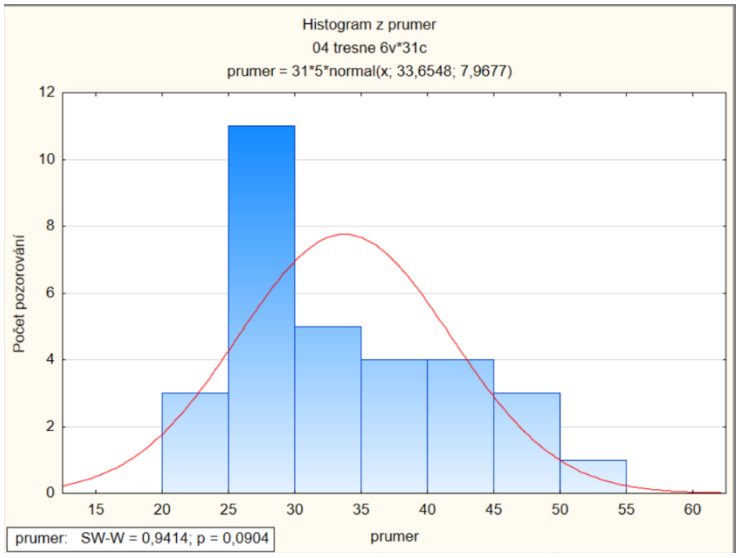
### Datový soubor KMENY.

Studujeme třešňové stromy (31 stromů) a máme údaje o průměru kmene v prsní výšce [cm], odhadu výšky stromu [metry] a odhadu dřevní hmoty [$m^3$]. Zajímá nás, jak silné jsou závislosti mezi naměřenými charakteristikami.

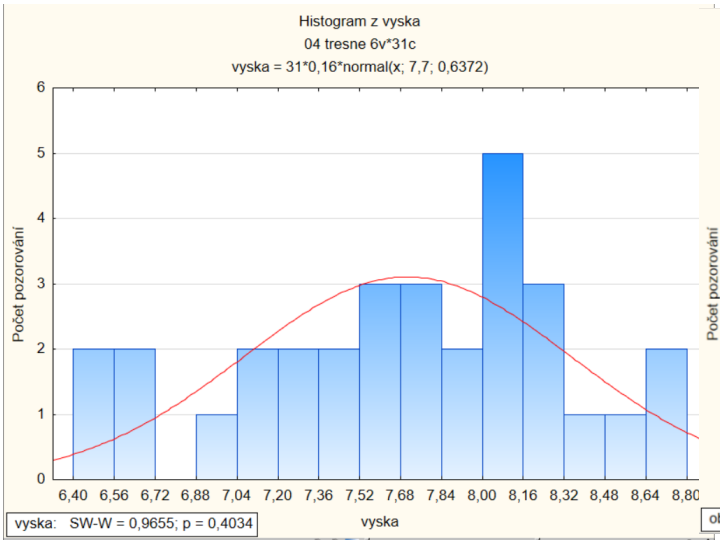
Sílu vztahů dvou proměnných popisujeme pomocí korelačního koeficientu a testu hypotézy, že skutečný (populační) korelační koeficient je roven nule a tedy obě studované charakteristiky jsou vzájemně nezávislé.

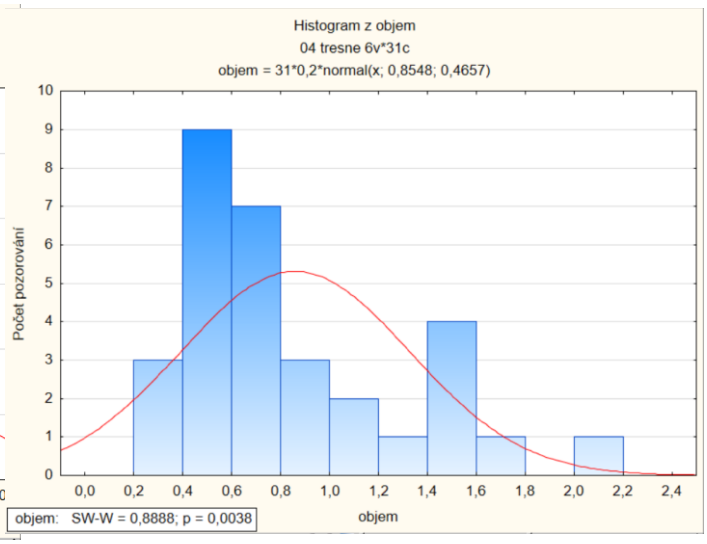
**Korelační koeficient** můžeme počítat parametricky (Pearsonův k.k.), pokud obě charakteristiky mají normální rozdělení, nebo neparametricky (Spearmanův k.k.).

**Ověření normality** všech tří proměnných:
Vizuální kontrola histogramů prozrazuje data sešikmená doprava pro průměr a pro objem. Pochybnosti potvrzují i testy normality (Shapiro-Wilkův test v rámečku na „dně“ grafu), které zamítají hypotézu, že data pocházejí z normálních rozdělení.
Histogram výšek stromů je tvarově sporný, Shapiro-Wilkův test však normalitu nevyvrací.




**Máme dvě možnosti**: buď zkusíme data transformovat pomocí logaritmické funkce, nebo zvolíme rovnou Spearmanův neparametrický korelační koeficient.




**Logaritmická transformace pomůže** (musím vytvořit nové proměnné, do kterých nechám napočítat logaritmy původních hodnot), Shapiro-Wilkovy testy již nezamítají hypotézu o normalitě proměnných: LOG(průměr) → p-hodnota = 0,32; LOG(výška) → p = 0,20; LOG(objem) → p = 0,38. Data o odhadu výšek stromů bychom transformovat nemuseli, ale bývá zvykem podrobit všechny „jednotkově podobné“ proměnné stejné úpravě…

**Pearsonův korelační koeficient r**.

P-hodnota se vztahuje k testu hypotézy, že skutečný korelační koeficient je nulový a charakteristiky jsou nezávislé (v testované dvojici). Test hypotézy provádíme t-testem. Testové statistiky najdeme v jiné tabulce zde: záložka *Možnosti*: *Formát zobrazení* → *Zobrazit detailní tabulku výsledků*, pak tlačítko VÝPOČET.

Korelace (04 tresne)
Označ. korelace jsou významné na hlad. p < ,050
N=31 (Celé případy vynechány u ChD)

| Proměnná | LOGprumer | LOGvyska | LOGobjem |
|---|---|---|---|
| **LOGprumer** | 1,0000 | ,5305 | ,9766 |
| | p= --- | p=,002 | p=0,00 |
| LOGvyska | ,5305 | 1,0000 | ,6492 |
| | p=,002 | p= --- | p=,000 |
| LOGobjem | ,9766 | ,6492 | 1,0000 |
| | p=0,00 | p=,000 | p= --- |

Statistiky → Základní statistiky → Korelační matice

**LOG(průměr) X LOG(výška): r = 0,53, p = 0,002, t = 3,37**

**LOG(průměr) X LOG(objem): r = 0,98, p < 0,001, t = 24,45**

**LOG(výška) X LOG(objem): r = 0,65, p < 0,001, t = 4,60**.

**Spearmanův korelační koeficient**:

(Ve STATISTICA zadám všechny tři proměnné do obou seznamů. Z nabídky *VYTVOŘIT* vyberu „*čtvercová matice*“ pro tuto přehlednou tabulku a potom „*detailní report*“ pro testové statistiky a p-hodnoty.)

Spearmanovy korelace (04 tresne)
ChD vynechány párově
Označ. korelace jsou významné na hl. p <,050

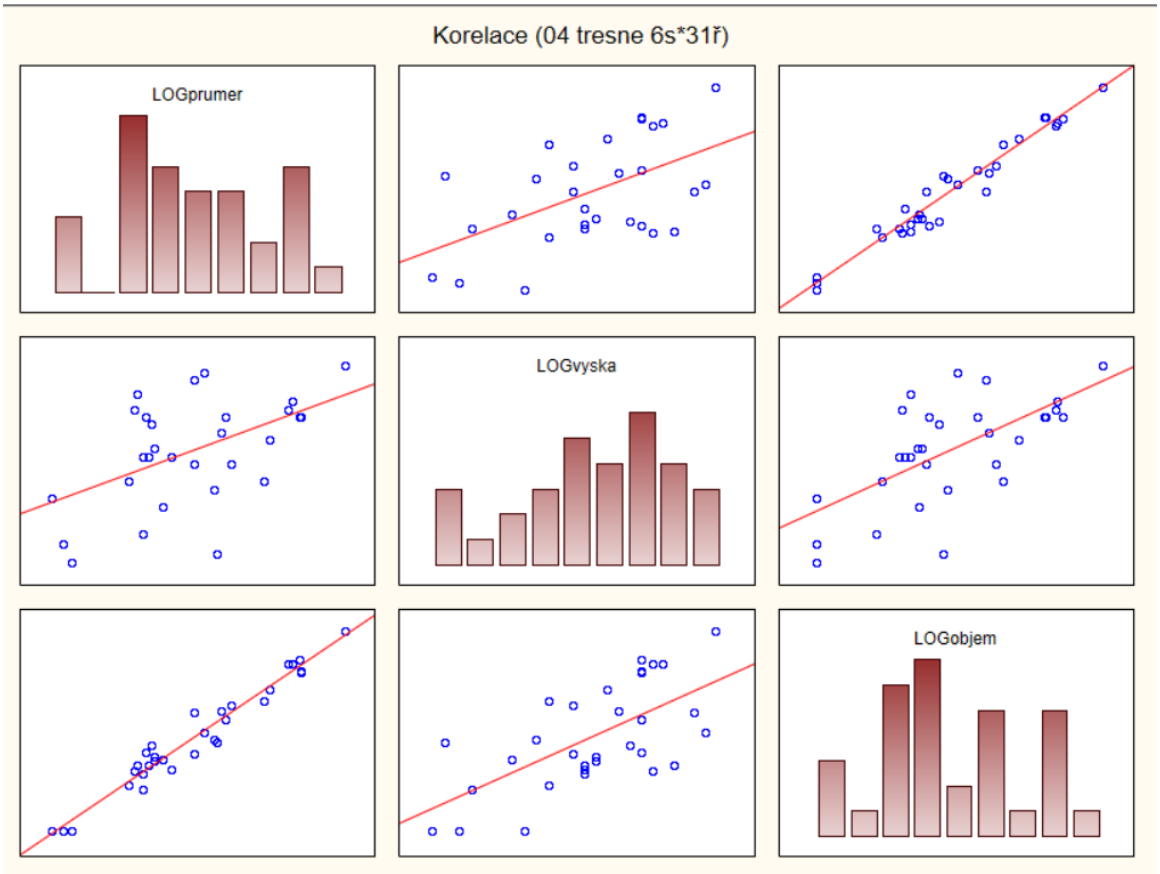
| Proměnná | prumer | vyska | objem |
|---|---|---|---|
| **prumer** | 1,000000 | 0,440839 | 0,954899 |
| vyska | 0,440839 | 1,000000 | 0,580562 |
| objem | 0,954899 | 0,580562 | 1,000000 |

**průměr X výška: r = 0,44, p = 0,013, t = 2,64**

**průměr X objem: r = 0,95, p < 0,001, t = 17,32**

**výška X objem: r = 0,58, p < 0,001, t = 3,84**.

Nejsilnější vztah je tedy mezi průměrem kmene a objemem dřevní hmoty. Jejich bodový graf je také uspořádán kolem teoretické regresní přímky „nejúžeji“, s nejmenším rozptylem.



**<u>Předpověď objemu dřevní hmoty</u>**: toto umí regresní analýza. Má-li být závislá proměnná OBJEM, potom jako nezávislou proměnnou vybírám tu nejvíce korelovanou charakteristiku. V našem příkladu to je PRŮMĚR.

V záhlaví bodového grafu z modulu „Korelační analýza“ máme rovnou uvedenu rovnici regresního lineárního modelu.