

Samostatné cvičení – ověřování normality dat

1. Načtete si do programu STATISTICA data *pacienti.sta*. **Přidejte** za proměnnou *váha* novou **proměnnou BMI** (body mass index – index tělesné hmotnosti), kterou vypočítáte z proměnné *výška* a *váha*.

Poznámka: V případě, že jste ze samostatného cvičení nepřepsali odlehlou hodnotu proměnné *výška*, učiňte tak nyní (hodnotu 100 přepište na hodnotu 144,1321).

2. Vypište zvlášť pro muže a ženy (proměnná *pohlaví*) **základní popisné statistiky** následujících proměnných: *váha*, *výška*, *BMI* (počet hodnot, průměr, medián, směrodatnou odchylku, minimum a maximum). Výsledek znázorněte v jedné tabulce (nápopvěda: změňte nastavení formy výstupů v sekci *By Group*).

3. Vykreslete kategorizované histogramy proměnných *výška*, *váha* a *BMI* pro muže a ženy zvlášť. Zkuste si proložit histogramy postupně **normálním rozdělením** a dalšími rozděleními ze záložky *Advanced* → *Fit types*.

4. Pro proměnné *výška*, *váha* a *BMI* (opět pro muže a ženy zvlášť) vykreslete **Q-Q graf**, **N-P graf** a **P-P graf**. Které proměnné dle těchto diagnostických grafů podle vás mají normální rozložení? Zapište svůj odhad do připravené tabulky.

5. Otestujte normalitu dat proměnných *výška*, *váha* a *BMI* pro muže a ženy zvlášť pomocí **Shapiro-Wilkova testu**. Zapište výsledek (p-hodnotu) do připravené tabulky. Srovnajte své odhady z diagnostických grafů s výsledky testů.

6. V případě, že se dle diagnostických grafů nebo S-W testu data řídí normálním rozdělením, jaký je v uvedených případech odhad **parametrů tohoto rozdělení** (střední hodnoty a rozptylu)? Hodnoty zaznamenejte do tabulky.

Tabulka: Vizuální a testové ověření normality.

Proměnná	Normalita dle Q-Q / N-P / P-P grafu (ano/ne)	p-hodnota Shapiro-Wilkova testu	Odhad střední hodnoty	Odhad rozptylu
Výška				
Muži	Ne/ne/ne	0.037		
Ženy	Ano/ano/ano	0.539	161.2	17.3
Váha				
Muži	Ne/ne/ne	0.004		
Ženy	Ano/ano/ano	0.784	65.9	25.1
BMI				
Muži	Ano/ano/ano	0.529	25.3	3.6
Ženy	Ano/ano/ano	0.200	25.4	4.3

Poznámky k nejčastějším chybám:

1. Parametry normální rozdělení jsou: **střední hodnota** a **rozptyl**. Nejlepším nestranným odhadem střední hodnoty u normálního rozdělení je **průměr** (nikoliv medián, ale měl by v případě normálního rozdělení stejný nebo podobný jako průměr), nejlepším nestranným odhadem rozptylu jako parametru je **výběrový rozptyl**.
2. Nepleťte si rozptyl a směrodatnou odchylku. **Směrodatná odchylka je odmocnina z rozptylu**. Na rozdíl od rozptylu je ve stejných jednotkách jako hodnocený parametr.

Další chyby:

1. Přehozené skupiny pohlaví (záměna žen a mužů).
2. Pořádně si přečtěte zadání! Odhad střední hodnoty a rozptylu měl být vyplněn pouze tam, kde jste pomocí testu nezamítli nulovou hypotézu o normalitě dat.
3. Tvrzení „u mužů nejde o rozdělení“, které se objevilo v jednom DÚ je naprosto chybné! Všechny náhodné veličiny se řídí nějakým rozložením (může a nemusí být modelové, jako např. normální rozložení). Správná interpretace např. výšky může být: „Pomocí Shapirova-Wilkova testu můžeme předpokládat, že se výška u žen v našem hodnoceném souboru řídí normálním rozdělením. U mužů jsme však nulovou hypotézu zamítli, tedy test prokázal, že výška u mužů nemá normální rozdělení.“