🔗Úkoly

1. (1b) Z obou datových souborů načtěte texty k analýze. Pro každý text zvlášť odhadněte základní charakteristiky délek slov, tj. střední hodnotu a rozptyl. Graficky znázorněte rozdělení délek slov.
2. (1b) Pro každý text zvlášť odhadněte pravděpodobnosti písmen (symbolů mimo mezery), které se v textech vyskytují. Výsledné pravděpodobnosti graficky znázorněte.
3. (1.5b) Na hladině významnosti 5% otestujte hypotézu, že rozdělení délek slov nezávisí na tom, o který jde text. Určete také p-hodnotu testu.
4. (1.5b) Na hladině významnosti 5% otestujte hypotézu, že se střední délky slov v obou textech rovnají. Určete také p-hodnotu testu.
5. (1b) Na hladině významnosti 5% otestujte hypotézu, že rozdělení písmen nezávisí na tom, o který jde text. Určete také p-hodnotu testu.

Nápověda k bodům 3 a 5: Proveďte test nezávislosti v kontingenční tabulce.

Poznámky

Úlohy můžete řešit v libovolném softwaru umožnujícím provádět potřebné výpočty. Testové statistiky můžete počítat ručně nebo s využitím statistických balíčků a funkcí.

Vždy je potřeba vysvětlit postup a výsledek řádně interpretovat.

Software

Vhodný je např. Python[1] (+scipy stats), R, Mathematica, Matlab, Excel atd.