Poznámky k řešením 1. domácí úlohy
(Místo oslavy státního svátku)

1. úloha Úkolem je porovnat spojitá rozdělení ve třech nezávislých výběrech.

- V popisných statistikách často chybí informace o variabilitě, např. alespoň výběrová směrodatná odchylka, nebo kvartilové rozpětí. Jen v jednom odevzdaném řešení jsou uvedeny rozsahy oněch výběrů, které v tabulce měly být také.
- K porovnání několika náhodných výběrů nemůže sloužit řada histogramů, které dokonce nemají stejné měřítko na vodorovné ose. Vhodnější je řada krabicových diagramů, které stejné měřítko mají.
- Ne všude byl uveden předpoklad nezávislosti tří výběrů. Občas sice uveden byl, ale bez vyjádření, zda (a proč) můžeme předpokládat jeho splnění.
- Podle mého názoru bylo možno v této úloze předpokládat normalitu i homoskedasticitu. Obojí bylo možno ověřit testy, které bez problému bezpečně projdou. Nicméně přijal jsem i řešení úlohy pomocí Kruskalova-Walisova testu.
- V žádném případě nelze předpokládat, že oněch 96 údajů o koncentraci železa tvoří **náhodný výběr**. Pokud by tomu tak bylo, jednoduché třídění by nemělo smysl, protože náhodný výběr předpokládá kromě nezávislosti také stejné rozdělení, takže by nebylo co testovat

2. úloha Jde o test homogenity tří binomických rozdělení.

- Když informace o roku vypěstování brambor **nebyla** v 1. úloze chápána jako realizace náhodné veličiny, nevidím důvod, proč stejný údaj chápat ve 2. úloze jako náhodnou veličinu. Interpretace kontingenční tabulky závisí na způsobu získání dat, na tom, které marginální četnosti jsou pevné (nenáhodné) a které náhodné. Jsou-li obojí náhodné, půjde o testování (stochastické) nezávislosti. V našich dvou úlohách jsou jedny z marginálních četností (year, site) spíše pevné, takže půjde o test homogenity multinomických rozdělení. Binomické rozdělení je speciálním případem multinomického. Hypotézu, že odpovídající si pravděpodobnosti jsou u všech multinomických rozdělení stejné, lze formulovat i jako nezávislost oněch rozdělení (jejich pravděpodobností) na příslušnosti k populacím, které srovnáváme. Jedná se však o nezávislost v běžném smyslu, nikoliv o nezávislost stochastickou (v pravděpodobnostním smyslu). Je pravda, že testová statistika je v obou případech stejná. Proto jsem pouze komentoval, když proměnná site či year byla chápána jako náhodná. Moje hodnocení řešení úlohy taková interpretace neovlivnila.

3. úloha Jde o test homogenity (shody pravděpodobností) dvou binomických rozdělení.

- Interpretace místa pěstování brambor jako náhodné veličiny je opět podivná. Výraz závislost v zadání bych raději chápal jako závislost v běžném, nikoliv stochastickém smyslu. Proto bych dal přednost intervalu spolehlivosti pro rozdíl dvou pravděpodobností, případně pro poměr mezi dvojími šancemi. Nicméně ani test hodnotící výběrový korelační koeficient dvou nula-jedničkových veličin nehodnotím jako chybu.

Všeobecně Jazyk, termíny.

- Je třeba držet se zavedené terminologie. V několika případech byl výběrový průměr označen jako střední hodnota, což skutečnost zatemňuje. Je třeba rozlišovat parametry rozdělení od jejich odhadů.
- Zpravidla je zbytečné uvádět hodnoty statistik na tolik desetinných míst, jak je poskytne počítač.
- U většiny testů je ekvivalentní porovnat hodnotu testové statistiky s příslušným kvantilem jejího rozdělení za platnosti nulové hypotézy na jedné straně a porovnáním p-hodnoty se zvolenou hladinou na straně druhé. Není důvod uvádět ve zprávě oba postupy.
- Výrok formulující výsledek testu musí být srozumitelný. Uvědomte si, co může a co nemůže být v klasické statistické indukci náhodné. Náhodná jsou data a vše, co je od nich odvozeno (rozhodnutí, interval spolehlivosti, p-hodnota). Skutečnost, že hypotéza platí, však náhodná není. Náhodné jsou meze intervalu spolehlivosti, nikoliv odhadovaný parametr.
- Často se rozepisujete formou *... kdyby nebylo normální rozdělení, pak bychom...* nebo bez dalšího zdůvodnění *... předpokládejme, že data pocházejí z normálního rozdělení ...* Vaším úkolem bylo rozhodnout se, zda ono normální rozdělení můžete předpokládat a popsat proč jej můžete předpokládat. Případně, proč nevadí, když data tak úplně předpokladu normálního rozdělení možná nevyhovují, Nepíšete písemku u zkoušky, abyste vyjmenovali všechny možnosti, které by mohly nastat. Zákazníka, pro kterého statistiku děláte, to zpravidla nezajímá, spíš je to mate.
- V zadání se připouští možnost kombinace češtiny a slovenštiny. Dovedu si představit, že jedna dílčí úloha bude řešena v češtině, jiná slovensky. Nelíbí se mi, když jsou oba jazyky smíchány v jednom odstavci nebo dokonce v jedné gramatické větě.
- Je třeba použít spisovnou češtinu, nikoliv češtinu hovorovou. Chápu, že se jazyk vyvíjí, ale některá díla byla psána až příliš hovorově.

Problémy sazby Platí pro sazbu v TeXu (LaTeXu) i Wordu.

- Bohužel ne všichni vědí, že v českém textu (snad i ve slovenském) se každý ze zápisů 5 % a 5% čte jinak. První znamená pět procent, kdežto druhý se čte pětiprocentní. Určitě nelze psát například 5%-ní.
- Popis tabulky (`\caption{}`) se zpravidla umisťuje nad tabulku, kdežto u obrázku pod něj.
- V budoucí diplomce, pokud ji budete psát česky, si pohlídejte, aby jednoslabičné předložky nezůstávaly osamocené na konci řádků. Pomoci může pomůcka nazvaná vlna či vlnka.

30. října 2014

Karel Zvára