

**VŠB – Technická univerzita Ostrava**  
**Fakulta elektrotechniky a informatiky**

**PRAVDĚPODOBNOST A STATISTIKA**

**Zadání 12**

JMÉNO STUDENTKY/STUDENTA:

OSOBNÍ ČÍSLO:

JMÉNO CVIČÍCÍ/CVIČÍCÍHO:

	DATUM ODEVZDÁNÍ	HODNOCENÍ
DOMÁCÍ ÚKOL 1:		
DOMÁCÍ ÚKOL 2:		
DOMÁCÍ ÚKOL 3:		
DOMÁCÍ ÚKOL 4:		
CELKEM:	-----	

**Ostrava, AR 2016/2017**

## **Popis datového souboru**

**Steam** je platforma společnosti Valve Corporation určená k digitální distribuci her a softwaru a zajištění multiplayerového a komunikačního zázemí pro hráče [1]. Tato platforma je dostupná pro všechny rozšířené operační systémy (Windows, OSX, Linux), největší podíl mají různé verze OS Windows (dohromady 95,75%), poté OSX (dohromady 3,31%) a zbytek zabírají distribuce Linuxu [2].

V souboru ukol\_12.csv jsou uvedeny celkové doby hraní sledovaných hráčů pro rok 2015 a 2016 a používané verze operačních systémů. Soubor obsahuje položky ID hráče, celková doba hraní za rok 2015, celková doba hraní za rok 2016 a používaná verze operačního systému (Win 7, Win 8.1, Win 10, MacOS).

### **Obecné pokyny:**

- Domácí úkoly odevzdávejte vždy v termínu, který určil váš cvičící.
- Portfolio domácích úkolů budete odevzdávat postupně. Tj. nejdříve odevzdáte titulní stránku + úkol 1, následně doplníte úkol 2, atd.
- Domácí úkoly zpracujte dle obecně známých typografických pravidel.
- Všechny tabulky i obrázky musí být opatřeny titulkem.
- Do domácích úkolů nevkládejte tabulky a obrázky, na něž se v doprovodném textu nebudete odkazovat.
- Bude-li to potřeba, citujte zdroje dle mezinárodně platné citační normy ČSN ISO 690.

**Úkol 1**

- a) Pomocí nástrojů explorační analýzy analyzujte celkovou dobu hraní pro systém Windows 8.1 v roce 2015 a v roce 2016. Data vhodně graficky prezentujte (krabicový graf, histogram, q-q graf) a doplňte následující tabulky a text.

Celková doba hraní, Windows 8.1			Po odstranění odlehlých pozorování	
	Rok 2015	Rok 2016	Rok 2015	Rok 2016
rozsah souboru				
<b>Míry polohy</b>				
minimum				
dolní kvartil				
medián				
průměr				
horní kvartil				
maximum				
<b>Míry variability</b>				
směrodatná odchylka				
variační koeficient (%)				
<b>Míry šikmosti a špičatosti</b>				
šikmost				
špičatost				

Identifikace odlehlých pozorování – vnitřní hradby		
dolní mez		
horní mez		

Osobní číslo studentky/studenta:

Číslo domácího úkolu: 12

**Grafická prezentace (krabicový graf, histogram, q-q graf):**

**Analýza počtu odehraných hodin na systému Windows 8.1 v roce 2015**

Sledovali jsme ..... hráčů používajících operační systém Windows 8.1. Celková nahraná doba u jednotlivých hráčů se v roce 2015 pohybovala v rozmezí ..... až ..... hodin. Herní doby u hráčů ..... byly identifikovány jako odlehlá pozorování a nebudou zahrnuty do dalšího zpracování. Možné příčiny vzniku odlehlých pozorování jsou: ..... / Žádná z hodnot nebyla identifikována jako odlehlé pozorování. Dále uvedené výsledky tedy pocházejí z analýzy odehrané doby u ..... hráčů. Průměrná herní doba byla ..... hodin, směrodatná odchylka pak ..... hodin. U poloviny hráčů celková herní doba nepřekročila ..... hodin. U poloviny hráčů se odehraná doba pohybovala v rozmezí ..... až ..... hodin. Vzhledem k hodnotě variačního koeficientu (.....%) lze / nelze analyzovaný soubor považovat za homogenní.

**Analýza počtu odehraných hodin na systému Windows 8.1 v roce 2016**

Sledovali jsme ..... hráčů používajících operační systém Windows 8.1. Celková nahraná doba u těchto hráčů se v roce 2016 pohybovala v rozmezí ..... až ..... hodin. Herní doby u těchto hráčů ..... byly identifikovány jako odlehlá pozorování a nebudou zahrnuty do dalšího zpracování. Možné příčiny vzniku odlehlých pozorování jsou: ..... / Žádná z hodnot nebyla identifikována jako odlehlé pozorování. Dále uvedené výsledky tedy pocházejí z analýzy odehrané doby u ..... hráčů. Průměrná herní doba byla ..... hodin, směrodatná odchylka pak ..... hodin. U poloviny hráčů celková herní doba nepřekročila ..... hodin. U poloviny hráčů se odehraná doba pohybovala v rozmezí ..... až ..... hodin. Vzhledem k hodnotě variačního koeficientu (.....%) lze / nelze analyzovaný soubor považovat za homogenní.

**Ověření normality počtu odehraných hodin na systému Windows 8.1 v roce 2015 na základě explorační analýzy**

Na základě grafického zobrazení (viz ..... ) a výběrové šikmosti a špičatosti (výběrová šikmost i špičatost leží / neleží v intervalu  $(-2; 2)$ ) lze / nelze předpokládat, že odehraná doba na systému Windows 8.1 v roce 2015 má normální rozdělení. Dle pravidla  $3\sigma$  / Čebyševovy nerovnosti lze tedy očekávat, že přibližně 95 % / více než 75 % hráčů bude mít odehráno mezi ..... až ..... hodinami.

**Ověření normality počtu odehraných hodin na systému Windows 8.1 v roce 2016 na základě explorační analýzy**

Na základě grafického zobrazení (viz ..... ) a výběrové šikmosti a špičatosti (výběrová šikmost i špičatost leží / neleží v intervalu  $(-2; 2)$ ) lze / nelze předpokládat, že odehraná doba na systému Windows 8.1 v roce 2016 má normální rozdělení. Dle pravidla  $3\sigma$  / Čebyševovy nerovnosti lze tedy očekávat, že přibližně 95 % / více než 75 % hráčů bude mít odehráno mezi ..... až ..... hodinami.

## Úkol 2

Porovnejte počet odehraných hodin u systému Windows 8.1 v roce 2015 a 2016. Nezapomeňte, že použité metody mohou vyžadovat splnění určitých předpokladů. Pokud tomu tak bude, okomentujte splnění/nesplnění těchto předpokladů jak na základě explorační analýzy (např. s odkazem na histogram apod.), tak i exaktně pomocí metod statistické indukce.

a) Graficky prezentujte srovnání počtu odehraných hodin v roce 2015 a 2016 pro hráče používající operační systém Win 8.1 (vícenásobný krabicový graf, histogramy, q-q grafy).

b) Určete bodové a 95% intervalové odhady pro střední odehranou dobu (resp. medián odehrané doby) hráčů používajících operační systém Win 8.1 v roce 2015 a 2016.

c) Určete bodový a 95% intervalový odhad střední hodnoty (resp. mediánu) rozdílu odehraných hodin v roce 2015 a 2016 u systému Win 8.1.

d) Na hladině významnosti 5 % rozhodněte, došlo-li v roce 2016 ke statisticky významnému zvýšení počtu odehraných hodin oproti roku 2015 u hráčů používajících operační systém Win 8.1.

e) V roce 2014 ve Spojených státech nahráli hráči v průměru 22 hodin za týden [3]. Na hladině významnosti 5 % rozhodněte, můžeme-li považovat počty odehraných hodin u hráčů používajících operační systém Win 8.1 v roce 2015 za nadprůměrné oproti uvedené hodnotě z roku 2014. (Uvažujte, že rok má 52 týdnů).

### Úkol 3

Na hladině významnosti 5 % rozhodněte, zda se střední roční odehraná doba, resp. medián roční odehrané doby, (pro rok 2016) liší v závislosti na operačním systému. Posouzení proveďte nejprve na základě explorační analýzy a následně pomocí vhodného statistického testu, včetně ověření potřebných předpokladů. V případě, že je mezi operačními systémy statisticky významný rozdíl, určete, pořadí operačních systémů dle oblíbenosti pro hraní her.

a) Daný problém vhodným způsobem graficky prezentujte (vícenásobný krabicový graf, histogramy, q-q grafy).

b) Ověřte normalitu počtu odehraných hodin v roce 2016 u všech čtyř operačních systémů (empiricky i exaktně).

c) Ověřte homoskedasticitu (shodu rozptylů) počtu odehraných hodin v roce 2016 jednotlivých operačních systémů (empiricky i exaktně).



d) Určete bodové a 95% intervalové odhady střední odehrané doby (resp. mediánu doby) v roce 2016 pro všechny srovnávané operační systémy. (Nezapomeňte na ověření předpokladů pro použití příslušných intervalových odhadů.)

e) Čistým testem významnosti ověřte, zda existuje statisticky významný rozdíl v oblíbenosti testovaných operačních systémů pro hraní her. Pro posouzení srovnajte střední odehrané doby (resp. mediány dob) v roce 2016 u všech operačních systémů. Pokud existuje statisticky významný rozdíl (na hladině významnosti 5%), zjistěte, zda lze některé skupiny operačních systémů označit (z daného hlediska) za homogenní, tj. stanovte pořadí operačních systémů dle oblíbenosti pro hraní her. (Nezapomeňte na ověření předpokladů pro použití zvoleného testu.)

#### Úkol 4

Vývojáři her se především soustřeďují na skupinu hráčů, kteří hraním tráví alespoň 5 hodin týdně. Rozdělte hráče na dvě skupiny, kde první skupina odehrála v průměru méně než 5 hodin týdně („občasní hráči“) a druhá alespoň 5 hodin týdně („náruživí hráči“) a následně posuďte skladbu hráčů v závislosti na typu používaného operačního systému (za rok 2016).

a) Srovnajte skladbu hráčů pro operační systémy Win 7, Win 8.1, Win 10 a MacOS. Výsledky prezentujte pomocí kontingenční tabulky, vhodného grafu a vhodné míry kontingence.

b) Určete bodový i 95% intervalový odhad pravděpodobnosti, že hráč hrající na systému Win 8.1 patří do skupiny „náruživých hráčů“.

c) Určete bodový i 95% intervalový odhad relativního rizika, že hráč počítačových her bude patřit do skupiny „občasných hráčů“ pro systém MacOS vzhledem ke (sloučeným) systémům Windows. Výsledky slovně interpretujte.

- d) Určete bodový i 95% intervalový odhad poměru šancí, že hráč počítačových her bude patřit do skupiny „občasných hráčů“ pro systém MacOS vzhledem ke (sloučeným) systémům Windows. Výsledky slovně interpretujte.
- e) Pomocí chí-kvadrát testu nezávislosti rozhodněte, jestli pravděpodobnost, že hráč počítačových her bude patřit do skupiny „občasných hráčů“ závisí statisticky významně na tom, na kterém operačním systému hráč paří.

## Literatura

- [1] Steam. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2017-01-27]. Dostupné z: <https://cs.wikipedia.org/wiki/Steam>
- [2] Steam Hardware & Software Survey: December 2016. *Steam* [online]. Bellevue (Washington), c2017 [cit. 2017-01-27]. Dostupné z: <http://store.steampowered.com/hwsurvey>
- [3] This is how much time the average gamer spends playing games every week. *BGR* [online]. New York, 2014 [cit. 2017-01-27]. Dostupné z: <http://bgr.com/2014/05/14/time-spent-playing-video-games/>

## **Jak identifikovat, zda jsou v datech odlehlá pozorování?**

### Empirické posouzení:

- použití vnitřních (vnějších) hradeb, resp.  $z$  – souřadnice, resp. mediánová suřadnice,
- vizuální posouzení krabicového grafu.

### Exaktní posouzení:

- Grubbsův test (parametrický test - vyžaduje normalitu dat)
- Deanův - Dixonův test (neparametrický test)

Jak naložit s odlehlými hodnotami by měl definovat hlavně zadavatel analýzy (expert na danou problematiku).

## **Jak ověřit normalitu dat?**

### Empirické posouzení:

- vizuální posouzení histogramu,
- vizuální posouzení grafu odhadu hustoty pravděpodobnosti,
- Q-Q graf,
- P-P graf,
- posouzení výběrové šikmosti a výběrové špičatosti.

### Exaktní posouzení:

- testy normality (např. Shapirův – Wilkův test, Andersonův-Darlingův test, Lillieforsův test, ...)

## **Jak ověřit homoskedasticitu (shodu rozptylů)?**

### Empirické posouzení:

- poměr největší a nejmenší směrodatné odchylky,
- vizuální posouzení krabicového grafu.

### Exaktní posouzení:

- $F$  – test (parametrický dvouvýběrový test),
- Bartlettův test (parametrický vícevýběrový test),
- Leveneův test (neparametrický test).