**ZUR357 Statistická analýza dat – Závěrečný test 14. prosince 2017 UČO:……………….**

*Do ISu odevzdáváte písemné odpovědi v tomto dokumentu* ***a*** *output s výpočty.*

1. **Jaké metody posouzení normality rozdělení znáte (3 body)?**

Matematicky:

- Kolmogorov-Smirnovův test

- Šikmost a špičatost

Graficky:

- Q-Q Plot

- Histogram

1. **Otestujte, zda jde proměnná *DCactNM* (počet online aktivit, které respondent vykonává) normálně rozložená. Vložte tabulku s výsledkem testu a okomentujte ji. (2 body)**

**Testujeme** a zamítáme nulovou hypotézu – rozložení počtu online aktivit není normální. Nazančuje to i histogram, k průměrným sedmi aktivitám počet postupně narůstá a naopak prudce klesá k sedmnácti – říkáme tomu rozložení zešikmené zprava.

|  |
| --- |
| **Tests of Normality** |
|  | Kolmogorov-Smirnova |
| Statistic | df | Sig. |
| DCactNM Number of online activities, TOTAL (17) | .076 | 25130 | .000 |
| a. Lilliefors Significance Correction |

1. **Vykonávají čeští (*country*) chlapci (*sexchild*) více online aktivit (*DCactNM*) než dívky?A můžeme takový výsledek očekávat i v populaci českých dětí? Vložte tabulky s výsledky testu a okomentujte je. (5 bodů)**

Nejprve je potřeba vybrat pouze české respondenty pomocí SELECT CASES, a poté provést (Independent Samples) t-test a testovat nulovou hypotézu o shodě populačních průměrů. Napovědět by Vám mělo, že nezávislá proměnná je dichotomická (nominální) a závislá proměnná kardinální.

Z výsledků vyplývá, že čeští chlapci vykonávají v průměru o 0,3 aktivity více než dívky, což není příliš velký rozdíl. Tabulku s výsledky t-testu jsem transponovala, aby se vešla na stránku. Výsledek čteme z druhého sloupce (Equal varinaces not assumed) a vidíme, že nelze nulovou hypotézu o shodě průměru chlapců a dívek zamítnout – tedy se populační průměry neliší.

|  |
| --- |
| **Group Statistics** |
|  | sexchild Child gender | N | Mean | Std. Deviation | Std. Error Mean |
| DCactNM Number of online activities, TOTAL (17) | 1.00 Male | 493 | 8.30 | 3.743 | .169 |
| 2.00 Female | 516 | 8.00 | 3.255 | .143 |

|  |
| --- |
| **Independent Samples Test** |
|  | DCactNM Number of online activities, TOTAL (17) |
| Equal variances assumed | Equal variances not assumed |
| Levene's Test for Equality of Variances | F | 16.660 |  |
| Sig. | .000 |  |
| t-test for Equality of Means | t | 1.362 | 1.358 |
| df | 1007 | 974.097 |
| Sig. (2-tailed) | .174 | .175 |
| Mean Difference | .300 | .300 |
| Std. Error Difference | .221 | .221 |
| 95% Confidence Interval of the Difference | Lower | -.132 | -.134 |
| Upper | .733 | .735 |

1. **Máme hypotézu, že děti z rodin s nižším vzděláním budou častěji problémové, resp. že se vzděláním rodičů bude klesat míra problémů dítěte. Otestujte ji, vložte tabulky s výsledky a okomentujte je. Pracujte s dětmi ze všech zemí. (5 bodů). *DCSDQcpMN*** – škála conduct problems, vyšší skóre znamená více problémů**; *DPEDUHH4 –*** nejvyšší dosažené vzdělání v domácnosti.

Nejprve je třeba vypnout filtr z předchozí úlohy, opět pomocí SELECT CASES. A poté pomocí ANOVY (nezávislá je kategorizovaná ordinální proměnná a má více než dvě kategorie, závislá proměnná je kardinální) opět testujeme hypotézu o shodě populačních průměrů.



Z tabulky Descriptives vidíme, že děti z rodin z nejnižším vzděláním mají nejvíce problémů a naopak, ale se nejedná o lineární vztah, neboť děti ze třetí vzdělanostní kategorie mají více problémů než děti z druhé, což bychom nejspíš čekali naopak. Můžeme se nicméně pomocí ANOVY podívat, jestli jsou rozdíly v průměrech statisticky významné.

Nejprve provedeme test homogenity rozptylu.



A poté přistoupíme k analýze. Z tabulky ANOVA vyplývá, že můžeme zamítnout nulovou hypotézu o shodě průměrů – mezi proměnnými je tedy vztah a některé vzdělanostní kategorie se budou zřejmě statisticky významně odlišovat. 

Abychom zjistili které, musíme provést post hoc test (ideálně Bonferonniho, který se nejsnáze intrpretuje, neboť označuje rozdíly hvězdičkami). Z tabulky jsou patrné statisticky významné rozdíly mezi všemi vzdělanostními kategoriemi s výjimkou první a třetí kategorie a druhé a třetí kategorie.



1. **Změřte sílu vztahu mezi následujícími dvojicemi proměnných, vyberte správný koeficient asociace, výsledek okomentujte s ohledem na těsnost (sílu) vztahu, zobecnitelnost na základní soubor a smysluplnost takové úvahy. (12 bodů)**
2. Čas denně strávený online (*DCtimeuse*) a počet online dovedností (*DCskillsNM*).
3. Pohlaví dítěte (*sexchild*) a přesvědčení, že je zkušenější uživatel/ka internetu než rodiče (*DCwebableA*).
4. Vzdělání rodiče, který poskytl rozhovor (*DPPTEDUA4),* a jeho sebejistota při používání internetu (*QP218*).
5. Počet přátel na sociálních sítích (QC316) a míra problémů s vrstevníky (*DCSDQppMN*).

V této otázce šlo ve všech případech o koeficienty asociace – to je jediný způsob, jak změřit sílu vztahu.

1. Obě proměnné jsou kardinální, použijeme proto Pearsonův korelační koeficient, vypočteme i Spearmana, abychom získali představu o linearitě vztahu. Oba koeficienty se příliš neliší, není tedy třeba pochybovat o linearitě vztahu. Hodnota r=0,32, (p=0,000) naznačuje, že s časem stráveným online rostou i dovednosti, naopak by to moc nedávalo smysl, hodnota r naznačuje středně těsný vztah (na sociální vědy dobrý výsledek), který lze navíc podle hodnoty p/sig. Očekávat i v populaci.

 

1. Závislá proměnná je ordinální (3 kategorie) a nezávislá proměnná je nominální (ale dichotomická) – můžeme tedy použít koeficienty pro ordinální proměnné. Použijeme Somersovo D (pro asymetrické vztahy. Výsledek čteme z druhého řádku (znalost ve srovnání s rodiči je závislou proměnnou) – záporná hodnota naznačuje, že dívky si budou důvěřovat méně než chlapci, zároveň se však hodnota blíží nule, tedy vztah mezi proměnnými je zanedbatelný (byť zobecnitelný, což nám ale nijak nepomůže). 
2. Obě pro měnné jsou ordinální, podobně jako v předchozím případě můžeme použít Somersovo D – opět čteme výsledek ze druhého řádku – se vzděláním roste i jistota rodiče při používání internetu, ačkoli vztah není příliš silný (na sebejistotu mají totiž vliv i mnohé jiné faktory), můžeme jej očekávat i v základním souboru.
3. Počet přátel měříme na ordinální škále, problémy s vrstevníky na kardinální – opět tedy použijeme koeficient pro ordinální proměnné. Tentokrát však není tak jisté, která proměnná je závislá a která nezávislá (nízký počet online přátel může být výsledkem konfliktů, konflikty ale také mohou být způsobeny právě vyloučením z kolektivu) – použijeme tedy koeficient pro symetrické vztahy: buď Kendallovo tauc nebo symetrického Somerse. Oba koeficienty jsou podobně nízké, naznačují zanedbatelnou souvislost, byť zobecnitelnou. 
4. **Pro vztah mezi vzděláním rodiče, který poskytl rozhovor (*DPPTEDUA4*), a jeho sebejistotou při používání internetu (*QP218*) vytvořte kontingenční tabulku a upravte ji pro publikování (musíte se správně rozhodnout pro řádková/sloupcová/total procenta). Otestujte, zda je mezi proměnnými vztah. Výsledek okomentujte. (8 bodů)**

Nezávislou proměnnou (vzdělání) máme v řádcích, proto použijeme řádková procenta – ta se pak interpretují ve sloupcích. Vidíme například, že se vzděláním jednoznačně roste podíl rodičů, kteří jsou si velmi jistí. Test chí-kvadrát naznačuje existenci vztahu (zamítáme nulovou hypotézu). Níže najdete výstup z SPSS a úplně na konci, jak by měla v ideálním případě vypadat tabulka pro publikování ve Vaší bakalářské práci (musí být jasné, kde je 100%, velikost vzorku, zdroj a výsledek statistického testu, tabulka je v češtině, očíslovaná a opatřená titulkem).



**Tabulka 1. Sebejistota rodičů při používání internetu podle vzdělání v % (Zdroj: EU Kids Online).**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Vůbec nejistý | Ne moc jistý | Docela jistý | Velmi jistý | Celkem | N |
| Základní a nižší vzdělání | 7.0% | 30.9% | 44.8% | 17.3% | 100.0% | 1219 |
| Střední bez maturity | 4.7% | 22.3% | 51.7% | 21.3% | 100.0% | 3784 |
| Střední s maturitou | 3.1% | 20.3% | 52.8% | 23.8% | 100.0% | 9535 |
| Vyšší odborné nebo vysokoškolské vzdělání | 1.6% | 9.8% | 49.6% | 39.0% | 100.0% | 5589 |
| Celkem | 3.2% | 18.4% | 51.2% | 27.2% | 100.0% | 20127 |

\* X2(9, N=20127) = 933,85, p<.01