

**Pravděpodobnost a statistika**

**Domácí úkoly 1S – 4S**

**Zadání 120**

|  |  |  |  |
| --- | --- | --- | --- |
| Jméno studentky/studenta: | | Vojtěch Prokop | |
| Osobní číslo: | | Pro 0255 | |
| Jméno cvičící/cvičícího: | | Mgr. Adéla Vrtková | |
|  | |  | |
|  | |  | |
|  | |  | |
|  | |  | |
|  | Datum odevzdání | | Hodnocení |
| Domácí úkol 1: |  | |  |
| Domácí úkol 2: |  | |  |
| Domácí úkol 3: |  | |  |
| Domácí úkol 4: |  | |  |
| Celkem: | --------------------- | |  |

**Ostrava, AR 2020/2021**

***Popis datového souboru***

*Běžné zářivky trpí efektem pomalého nabíhání, tedy plného výkonu dosáhnou až po jisté době provozu. Toto chování je ovlivněno okolní teplotou, což v praxi znamená, že v chladném prostředí může zářivkám trvat výrazně déle než dosáhnou maximálního výkonu.*

*Pro test náběhu zářivek na plný světelný výkon bylo vybráno celkem 350 zářivek od čtyř různých výrobců (Amber, Bright, Clear, Dim). Všechny zářivky měly deklarovaný maximální světelný tok 1000 lm. U každé zářivky byl změřen světelný tok po 30 sekundách od zapnutí, nejprve při teplotě 22 °C a poté při teplotě 5°C.*

*V souboru ukol\_X.xlsx jsou pro každou z testovaných zářivek uvedeny následující údaje:*

* *pořadové číslo zářivky,*
* *výrobce – Amber (A), Bright (B), Clear (C), Dim (D),*
* *naměřený světelný tok v lumenech při okolní teplotě 5°C,*
* *naměřený světelný tok v lumenech při okolní teplotě 22°C.*

***Obecné pokyny:***

* *Úkoly zpracujte dle obecně známých typografických pravidel.*
* *Všechny tabulky i obrázky musí být opatřeny titulkem.*
* *Do 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**

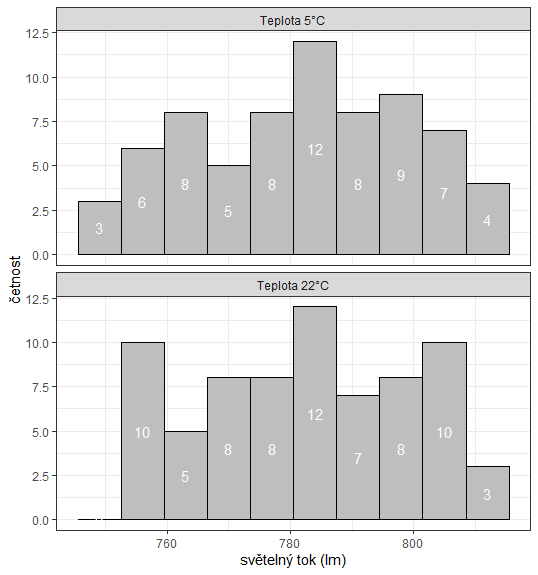
* *Pomocí nástrojů explorační analýzy analyzujte světelný tok zářivek výrobce Amber po 30 sekundách od zapnutí při teplotách 5°C a 22°C. Data vhodně graficky prezentujte (krabicový graf, histogram, q-q graf) a doplňte následující tabulky a text.*

Výsledky popisné statistiky lze vidět v Tab. 1 a jejich vizualizaci na Obr. 1, Obr. 2 a Obr. 3.

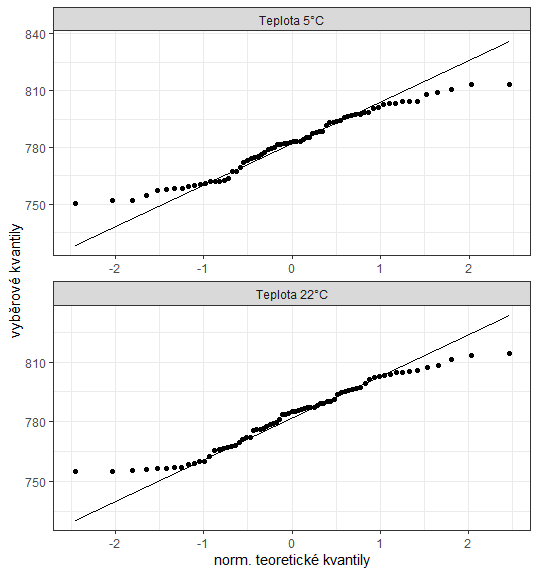
Tab. 1: Světelný tok (lm) zářivek Amber v závislosti na teplotě (souhrnné statistiky)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Světelný tok zářivek Amber (lm)** | | |  | **Po odstranění odlehlých pozorování** | | | |
|  | **5°C** | **22°C** |  | **5°C** | **22°C** | |
| **rozsah souboru** | **72** | **72** |  | 70 | 71 | |
| **minimum** | 652,1 | 658,3 |  | 750,3 | 754,9 | |
| **dolní kvartil** | 763,4 | 767,4 |  | 767,2 | 767,8 | |
| **Medián** | 782,5 | 784,6 |  | 782,8 | 785,1 | |
| **Průměr** | 778,6 | 781,0 |  | 782,1 | 782,7 | |
| **horní kvartil** | 796,7 | 795,9 |  | 797,0 | 796,0 | |
| **maximum** | 813,4 | 814,4 |  | 813,4 | 814,4 | |
| **směrodatná odchylka** | 27,1 | 22,5 |  | 17,5 | 17,1 | |
| **variační koeficient (%)** | 3,5 | 2,9 |  | 2,2 | 2,2 | |
| **šikmost** | -2,5 | -2,2 |  | -0,1 | -0,1 | |
| **špičatost** | 9,5 | 10,5 |  | -1,1 | -1,1 | |
| **Identifikace odlehlých pozorování – vnitřní hradby** | | |  |  |  |  |
| **dolní mez** | 713,59 | 724,6 |  |  |  |  |
| **horní mez** | 846,49 | 838,6 |  |  |  |  |

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

**** Obr. 1: Světelný tok zářivek (lm) výrobce Amber v závislosti na teplotě (krabicový graf, původní data)

Obr. 2: Světelný tok zářivek (lm) výrobce Amber v závislosti na teplotě (histogram, data po odstranění odl. pozorování)

****

Obr. 3: Světelný tok zářivek (lm) výrobce Amber v závislosti na teplotě (Q-Q graf, data po odstranění odl. pozorování)

**Analýza světelného toku zářivek výrobce Amber (po 30 sekundách od zapnutí, při teplotě 5°C)**

Během testu byl měřen světelný tok 72 kusů zářivek výrobce Amber. Naměřená světelný tok při teplotě 5°C se pohyboval v rozmezí 652,1 lm až 813,4 lm. Světelný tok zářivek č. 19, 29 byl na základě metody vnitřních hradeb identifikován jako odlehlé pozorování a nebude zahrnut do dalšího zpracování. Možné příčiny vzniku odlehlých pozorování jsou: nestandartně kvalitní zářivka. Dále uvedené výsledky tedy pocházejí z analýzy světelný toku 70 kusů zářivek. Jejich průměrný světelný tok byl 782,1 lm, směrodatná odchylka pak 17,5 lm. U poloviny testovaných zářivek světelný tok nepřekročil 782,8 lm. V polovině měření se světelný tok pohyboval v rozmezí 767,2 lm až 797,0 lm. Vzhledem k hodnotě variačního koeficientu (2.2 %) lze analyzovaný soubor považovat za homogenní.

**Analýza světelného toku zářivek výrobce Amber (po 30 sekundách od zapnutí, při teplotě 22°C)**

Během testu byl měřen světelný tok 72kusů zářivek výrobce Amber. Naměřená světelný tok při teplotě 22°C se pohyboval v rozmezí 658,3 lm až 814,4 lm. Světelný tok zářivek č. 29 byl na základě metody vnitřních hradeb identifikován jako odlehlé pozorování a nebude zahrnut do dalšího zpracování. Možné příčiny vzniku odlehlých pozorování jsou: nestandartně kvalitní zářivka. Dále uvedené výsledky tedy pocházejí z analýzy světelný toku 71 kusů zářivek. Jejich průměrný světelný tok byl 782,7 lm, směrodatná odchylka pak 17,1 lm. U poloviny testovaných zářivek světelný tok nepřekročil 785,1 lm. V polovině měření se světelný tok pohyboval v rozmezí 767,8 lm až 796,0 lm. Vzhledem k hodnotě variačního koeficientu (2.2 %) lze analyzovaný soubor považovat za homogenní.

**Ověření normality světelného toku zářivek výrobce Amber po 30 sekundách od zapnutí při teplotě 5°C na základě explorační analýzy**

Na základě grafického zobrazení (viz Obr. 2 a Obr. 3) a výběrové šikmosti a špičatosti (výběrová šikmost i špičatost leží v intervalu lze předpokládat, že světelný tok zářivek výrobce Amber při teplotě 5°C má normální rozdělení. Dle pravidla 3 lze tedy očekávat, že přibližně 95 % zářivek bude mít světelný tok v rozmezí 747,3 lm až 817,0 lm.

**Ověření normality světelný toki zářivek výrobce Amber po 30 sekundách od zapnutí při teplotě 22 °C na základě explorační analýzy**

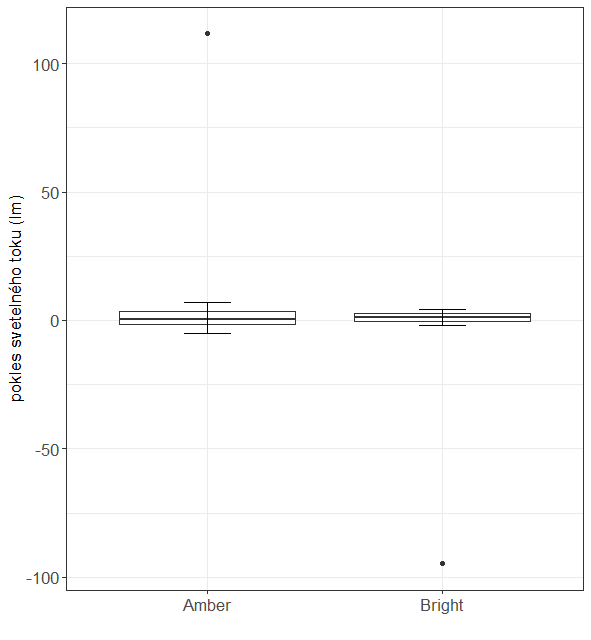
Na základě grafického zobrazení (viz Obr. 2 a Obr. 3) a výběrové šikmosti a špičatosti (výběrová šikmost i špičatost leží v intervalu lze předpokládat, že světelný tok zářivek výrobce Amber při teplotě 22°C má normální rozdělení. Dle pravidla 3 lze tedy očekávat, že přibližně 95 % zářivek bude mít světelný tok v rozmezí 748,6 lm až 816,9 lm.

**Úkol 2**

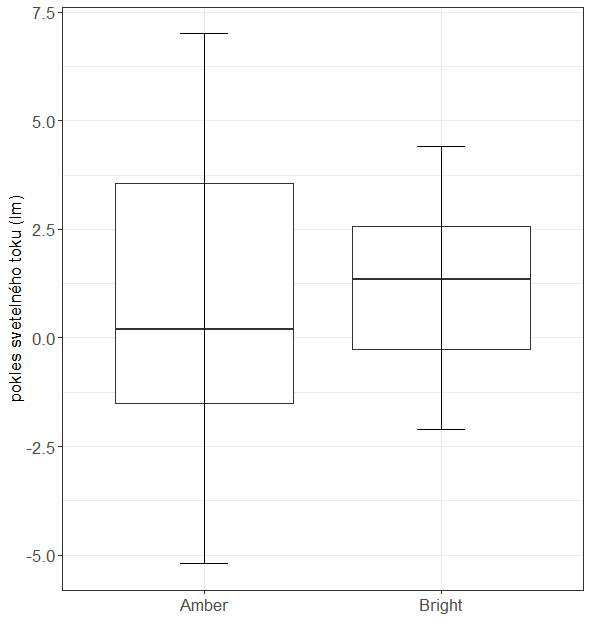
*Porovnejte pokles světelného toku po 30 sekundách od zapnutí při snížení okolní teploty z 22°C na 5°C u zářivek od výrobců Amber a Bright. Nezapomeňte, že použité metody mohou vyžadovat splnění určitých předpokladů. Pokud tomu tak bude, okomentujte splnění/nesplnění těchto předpokladů jak na základě explorační analýzy (např. s odkazem na histogram apod.), tak exaktně pomocí metod statistické indukce.*

1. *Graficky prezentujte srovnání poklesů světelného toku zářivek výrobců Amber a Bright při snížení okolní teploty (vícenásobný krabicový graf, histogramy, q-q grafy).* *Srovnání okomentujte (včetně informace o případné manipulaci s datovým souborem).*

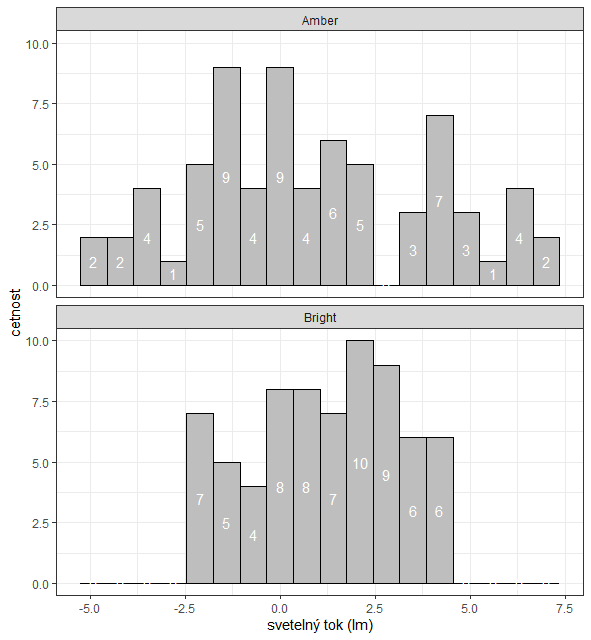
U výrobce Amber i Bright bylo pozorováno jedno odlehlé pozorování (viz Obr. 4), toto jsme se rozhodl z dalšího zpracování vypustit.

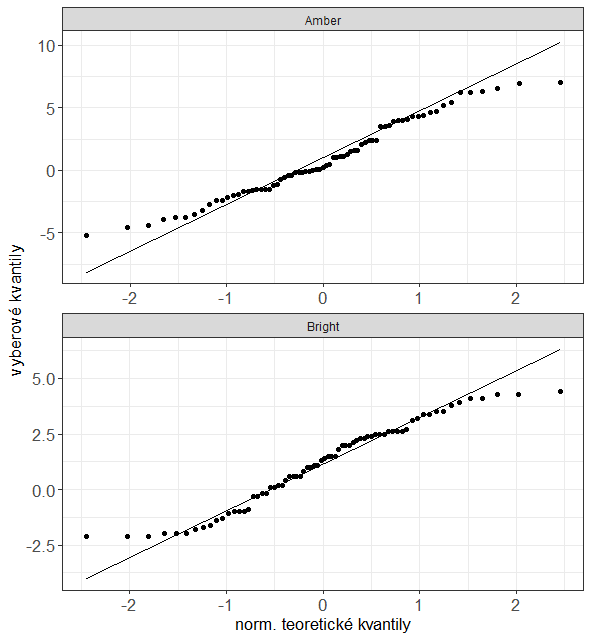
**U obou výrobců došlo spíše k poklesu světelného toku (viz. Obr. 5). Zároveň, ale dochází i k nárustu světelného toku. U výrobce Amber, lze pozorovat výraznější rozptyl hodnot. Zdá se, že u výrobce Bright oproti výrobce Amber dochází v průměru k výraznějším poklesům světelného toku.

Obr. 4: Srovnání poklesů světelného toku po 30 sekundách od zapnutí při snížení okolní teploty z 22°C na 5°C u zářivek od výrobců Amber a Bright (krabicový graf, původní data)



Obr. 5: Srovnání poklesů světelného toku po 30 sekundách od zapnutí při snížení okolní teploty z 22°C na 5°C u zářivek od výrobců Amber a Bright (krabicový graf, data po odstranění odlehlých pozorování)



Obr. 6: Srovnání poklesů světelného toku po 30 sekundách od zapnutí při snížení okolní teploty z 22°C na 5°C u zářivek od výrobců Amber a Bright (histogramy, data po odstranění odlehlých pozorování)

Obr. 7: Srovnání poklesů světelného toku po 30 sekundách od zapnutí při snížení okolní teploty z 22°C na 5°C u zářivek od výrobců Amber a Bright (Q-Q grafy, data po odstranění odlehlých pozorování)

1. *Na hladině významnosti 5 % rozhodněte, zda jsou střední poklesy (popř. mediány poklesů) světelného toku zářivek výrobců Amber a Bright statisticky významné. K řešení využijte bodové a intervalové odhady i testování hypotéz. Výsledky okomentujte.*

Tab. 2: Nástroje pro ověření předpokladů symetrie a normality poklesu světelného toku v závislosti na teplotě

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Výrobce*** | ***Šikmost*** | ***Špičatost*** | ***Shapirův-Wilkův test (p-hodnota)*** | ***Test symetrie (p-hodnota)*** |
| Amber | 0,2 | -0,8 | 0,103 | 0,028 |
| Bright | -0,2 | -1,1 | 0,012 | 0,279 |

Dle prezentovaných grafů (viz Obr. 6 a Obr. 7) lze usuzovat, že poklesy světelného toku jdou modelovat normálním rozdělením. Šikmost a špičatost (viz Tab. 2) jsou v přípustném intervalu.

Dle Shapirova-Wilkova testu nelze na hladině významnosti 0,05 pokles světelného toku výrobce Bright modelovat normálním rozdělením (viz Tab. 2).

U obou výrobců nebylo prokázáno normální rozdělení, a proto nelze přistoupit k určení odhadů střední hodnoty.

Následujeme tedy provedením testu symetrie, kde zase podle výsledků můžeme usoudit, že symetrie byla porušena u výrobce Amber, a proto využijeme znaménkový test.

V datové sadě, máme pokles reprezentovaný jako kladné číslo. Při intervalových odhadech se proto přikláníme k levostranné verzi.

Tab. 3: Odhad mediánů poklesu světelného toku (lm) dle výrobce a test významnosti poklesu

|  |  |  |  |
| --- | --- | --- | --- |
| ***Výrobce*** | ***Bodový odhad (lm)*** | ***95% levostranný intervalový odhad (lm)*** | ***Znaménkový levostranný test (p-hodnota)*** |
| Amber | 0,20 | (-0,20; ∞) | 0,275 |
| Bright | 1,35 | (0,60; ∞) | < 0,001 |

U výrobce Amber, lze očekávat, že polovina zářivek bude vykazovat pokles světelného toku menší než 0,20 lm. 95% levostranný intervalový odhad mediánu poklesu světelného toku u výrobce Amber je (-0,20; ∞) lm. Společně se znaménkovým levostranným testem, lze pozorovat, že medián poklesu světelného toku je statisticky nevýznamný (na hladině významnosti 5%).

U výrobce Bright, lze očekávat, že polovina zářivek bude vykazovat pokles světelného toku menší než 1,35 lm. 95% levostranný intervalový odhad mediánu poklesu světelného toku u výrobce Amber je (0,60; ∞) lm. Společně se znaménkovým levostranným testem, lze pozorovat, že medián poklesu světelného toku je statisticky významný (na hladině významnosti 5%).

1. *Na hladině významnosti 5 % rozhodněte, zda je rozdíl středních hodnot (mediánů) poklesů světelných toků zářivek výrobců Amber a Bright (při snížení okolní teploty) statisticky významný. K řešení využijte bodový a intervalový odhad i čistý test významnosti. Výsledky okomentujte.*

Vzhledem k předchozímu příkladu, jsme zamítli normalitu u výrobce zářivek Bright (viz Tab. 2), takže budeme pokračovat kontrolou stejného tvaru rozdělení.

Z histogramů, lze pozorovat srovnatelný tvar rozdělení, takže jsme schopni využít Mannův-Whitneyho test.

Tab. 4: Srovnání mediánů poklesu světelného toku (lm) výrobce Bright ( a výrobce Amber (

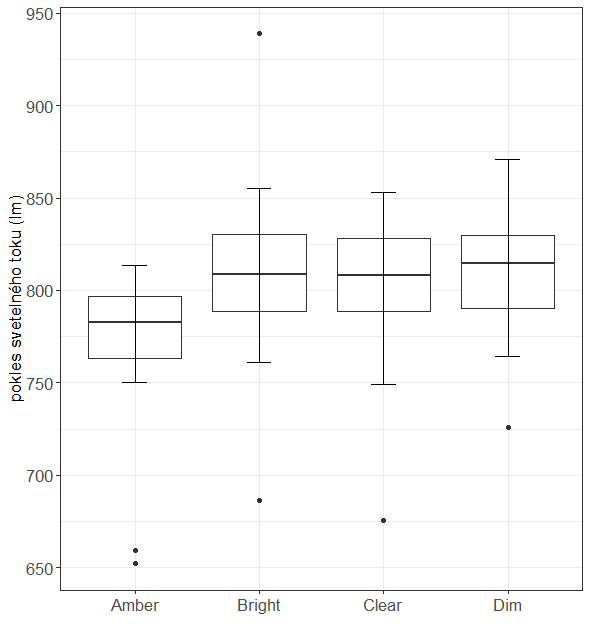
|  |  |
| --- | --- |
| ***Bodový odhad - (lm)*** | 1,15 |
| ***95% levostranný intervalový odhad - (lm)*** | (-0,49; 1,50) |
| ***Mannův-Whitneyho levostranný test (p-hodnota)*** | 0,345 |

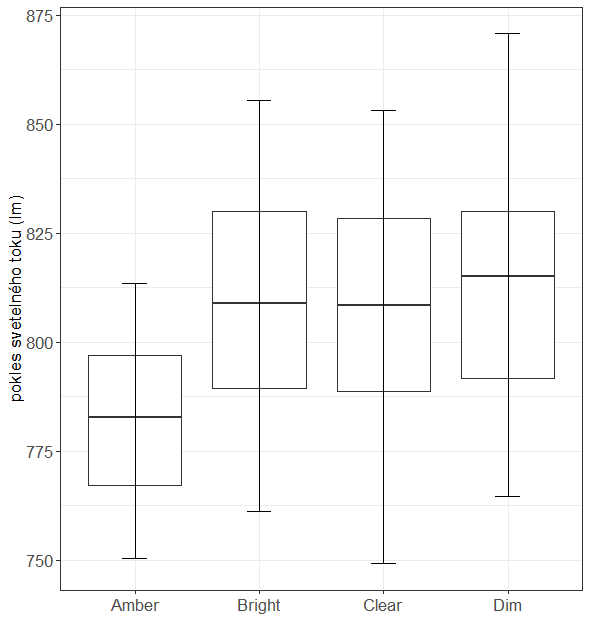
Data neposkytují dostatek důkazů pro podpoření tvrzení, že by se zářivky vyrobené od Bright signifikantně lišily od zářivek vyrobených od Amber. Na zadané hladině významnosti 5%, tedy rozdíl mediánů nelze považovat za statisticky významný (viz výsledky Tab. 4).

**Úkol 3**

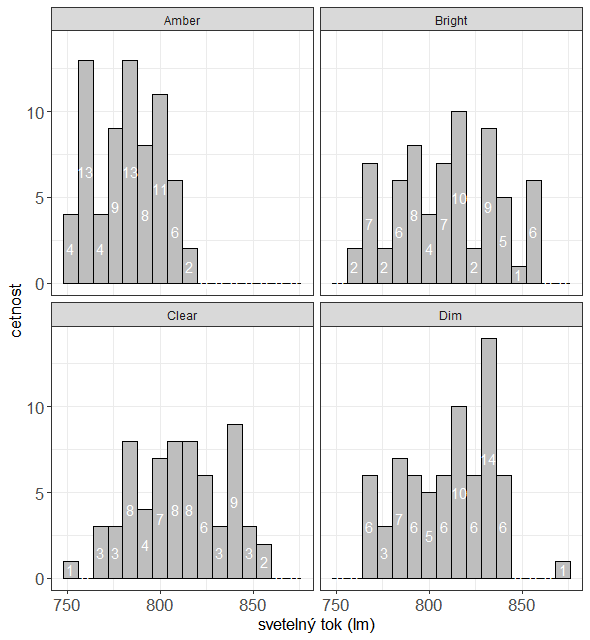
*Na hladině významnosti 5 % rozhodněte, zda se světelný tok zářivek při teplotě 5 °C liší v závislosti na tom, od kterého výrobce pocházejí. Posouzení proveďte nejprve na základě explorační analýzy a následně pomocí vhodného statistického testu, včetně ověření potřebných předpokladů. V případě, že se světelný tok zářivek jednotlivých výrobců statisticky významně liší, určete pořadí výrobců dle středního světelného toku (popř. mediánu světelného toku) zářivek při 5°C.*

1. *Daný problém vhodným způsobem graficky prezentujte (vícenásobný krabicový graf, histogramy, q-q grafy). Srovnání okomentujte (včetně informace o případné manipulaci s datovým souborem).*

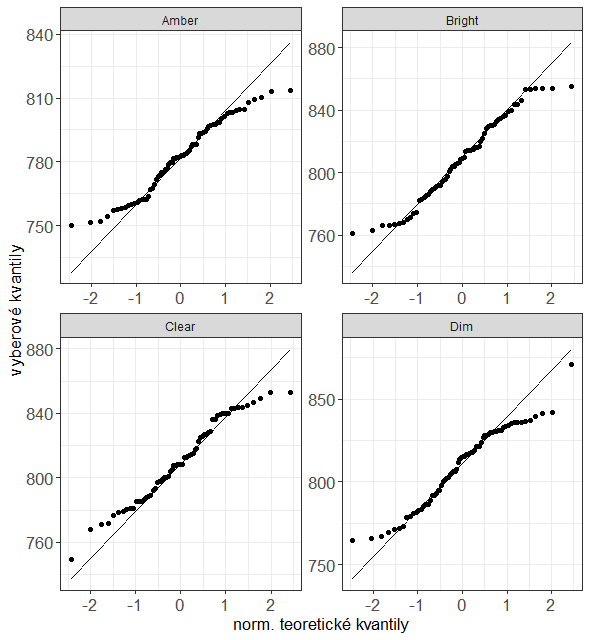


Obr. 8: Světelný tok zářivek (lm) všech výrobců v závislosti na teplotě 5°C (krabicový graf, původní data)

Obr. 9: Světelný tok zářivek (lm) všech výrobců v závislosti na teplotě 5°C (krabicový graf, data po odstranění odlehlých pozorování)



Obr. 10: Světelný tok zářivek (lm) všech výrobců v závislosti na teplotě 5°C (histogram, data po odstranění odlehlých pozorování)



Obr. 11: Světelný tok zářivek (lm) všech výrobců v závislosti na teplotě 5°C (Q-Q graf, data po odstranění odlehlých pozorování)

1. *Ověřte normalitu a symetrii světelného toku zářivek při teplotě 5°C u všech čtyř výrobců (empiricky i exaktně).*

Tab. 5: Nástroje pro ověření předpokladů symetrie a normality dle výrobce

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Výrobce*** | ***Šikmost*** | ***Špičatost*** | ***Shapirův-Wilkův test***  ***(p-hodnota)*** | ***Test symetrie***  ***(p-hodnota)*** |
| Amber | -0,1 | -1,1 | 0,032 | 0,677 |
| Bright | 0,0 | -1,0 | 0,046 | 0,985 |
| Clear | -0,1 | -0,8 | 0,121 | 0,413 |
| Dim | -0,2 | -0,7 | 0,023 | 0,017 |

Dle prezentovaných grafů na Obr. 10 a Obr. 11 nelze zcela usoudit, zda světelný tok lze modelovat pomocí normálního rozdělení. Například na histogramu u výrobce Dim, lze pozorovat záporné zešikmení, u výrobce Amber naopak špičatost. V závislosti na Q-Q grafech, lze pozorovat podobné anomálie, přičemž si můžeme povšimnout statistických jednotek, které vypadají jako odlehlá pozorování, i když tato byla korektně odstraněna v předchozích krocích.

Dle empiricky vypočtených hodnot šikmosti a špičatosti (viz. Tab. 5) data odpovídají normálnímu rozdělení.

Dle Shapirova-Wilkova testu lze na hladině významnosti 0,05 předpoklad normality pro výrobce Amber, Bright a Dim zamítnout.

Dle testu symetrie (viz. Tab. 5), grafů (viz. Obr. 10), lze zamítnout symetričnost u výrobce Dim.

1. *Ověřte homoskedasticitu (shodu rozptylů) světelného toku zářivek při teplotě 5 °C jednotlivých výrobců (empiricky i exaktně).*

Tab. 6: Rozptyly světelného toku v závislosti na teplotě 5 °C dle výrobce zářivek

|  |  |
| --- | --- |
| ***Výrobce*** | ***Rozptyl*** |
| Amber | 303,6 |
| Bright | 717,2 |
| Clear | 616,8 |
| Dim | 552,0 |

Tab. 7: Nástroje pro ověření předpokladů shody rozptylů

|  |  |
| --- | --- |
| ***Empiricky (největší ku nejmenšímu) < 2*** | (717.1908 / 303.6150) ≅ 2,4 |
| ***Exaktně (*Leveneho test*) p-hodnota*** | 0,004 |

Dle Tab. 6 a Tab. 7 lze říct, že rozptyly světelného toku při teplotě 5 °C nejsou srovnatelné. Tento úsudek můžeme prezentovat v závislosti na empirickém ověření největšího rozptylu ku nejmenšímu. Tento poměr vyšel 2,4 což přesahuje hodnotu 2. Homoskedacitia je zároveň zamítnuta v závislosti na Leveneho testu, kde nám p-hodnota vychází nižší než hladina významnosti.

1. *Určete bodové a 95% intervalové odhady střední hodnoty (popř. mediánu) světelného toku zářivek při teplotě 5°C pro všechny srovnávané výrobce. (Nezapomeňte na ověření předpokladů pro použití příslušných intervalových odhadů.)*
2. *Čistým testem významnosti ověřte, zda je pozorovaný rozdíl středních hodnot (popř. mediánů) světelného toku zářivek při teplotě 5°C statisticky významný na hladině významnosti 5 %. Pokud ano, zjistěte, zda lze některé skupiny výrobců označit (z hlediska světelného toku zářivek po 30 sekundách od zapnutí, při teplotě 5°C) za homogenní, tj. určete pořadí výrobců dle středních hodnot (popř. mediánů) světelného toku zářivek při 5°C. (Nezapomeňte na ověření předpokladů pro použití zvoleného testu.)*

**Úkol 4**

*Všichni čtyři výrobci udávají, že jejich zářivky dosáhnou při 5°C po 30 sekundách od zapnutí alespoň osmdesáti procent deklarovaného maximálního světelného toku (tj. 80 % z 1 000 lm). Definujte si novou dichotomickou proměnnou Splnění požadavku na deklarovaný světelný tok po 30 s (při 5°C), která bude nabývat hodnot {ANO, NE}. Poznámka: Pracujte s původními daty, nikoliv s daty po odstranění odlehlých pozorování.*

1. *Srovnejte zářivky jednotlivých výrobců dle toho, zda při teplotě 5°C splňují deklarovaný světelný tok po 30 s od zapnutí pro jednotlivé výrobce (Amber, Bright, Clear, Dim). Výsledky prezentujte pomocí kontingenční tabulky, vhodného grafu a vhodné míry kontingence. Vaše úsudky komentujte.*
2. *V případě výrobce Bright určete bodový i 95% intervalový odhad pravděpodobnosti, že při teplotě 5°C zářivka nedosáhne po 30 sekundách požadovaného světelného toku (80 % deklarovaného maximálního světelného toku). Nezapomeňte na ověření předpokladů pro použití intervalového odhadu.*
3. *Určete bodový i 95% intervalový odhad relativního rizika, že zářivka při teplotě 5°C nedosáhne po 30 sekundách požadovaného světelného toku (80 % deklarovaného maximálního světelného toku), pro „nejhoršího“ výrobce (vzhledem k „nejlepšímu“ výrobci). Výsledky slovně interpretujte.*
4. *Určete bodový i 95% intervalový odhad poměru šancí, že zářivka při teplotě 5°C nedosáhne po 30 sekundách požadovaného světelného toku (80 % deklarovaného maximálního světelného toku), pro „nejhoršího“ výrobce (vzhledem k „nejlepšímu“ výrobci). Výsledky slovně interpretujte.*
5. *Pomocí chí-kvadrát testu nezávislosti rozhodněte, jestli to, že zářivka při teplotě 5°C nedosáhne po 30 sekundách požadovaného světelného toku (80 % deklarovaného maximálního světelného toku), závisí statisticky významně na tom, od kterého výrobce zářivka pochází. Výsledky okomentujte.*

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

*Emiprické posouzení:*

* *použití vnitřních (vnějších) hradeb,*
* *vizuální posouzení krabicového grafu.*

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

## Jak ověřit normalitu dat?

*Emiprické posouzení:*

* *vizuální posouzení histogramu,*
* *vizuální posouzení grafu odhadu hustoty pravděpodobnosti,*
* *Q-Q 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)?

*Emiprické posouzení:*

* *poměr největšího a nejmenšího rozptylu,*
* *vizuální posouzení krabicového grafu.*

*Exaktní posouzení:*

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