



Spolufinancované z
programu Európskej únie
Erasmus+



2. Štatistika

Oxford Learner's Dictionary definuje **štatistiku** ako súbor informácií reprezentovaný číslami. Štatistika je matematická oblasť, ktorá sa zaoberá metódami zberu údajov a ich analýzou. Veľmi jednoduchým príkladom takéhoto spracovania údajov je športová rubrika v novinách. Povedzme, že sa práve hrá Wimbledon a v novinách nájdeme tieto informácie:

| Final - Center Court | | Final |
|--|--|-------|
|  11 S. Williams | | 2 2 |
|  7 S. Halep | | 6 6 ◀ |

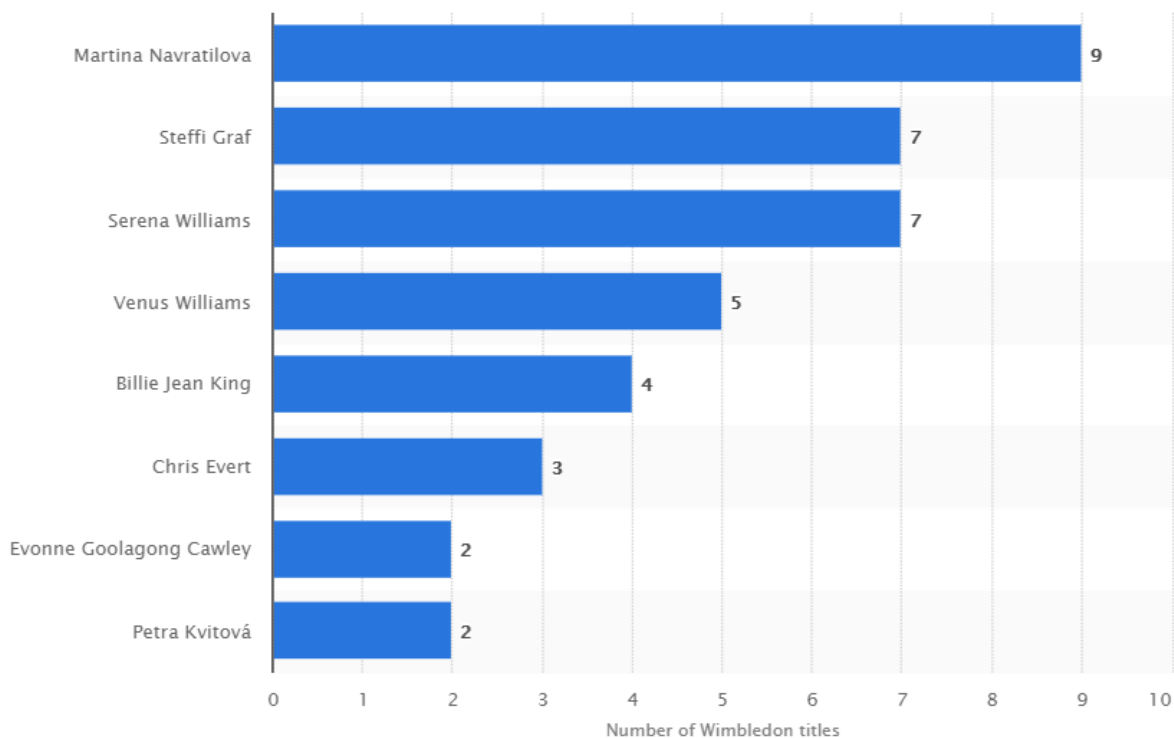
|  Serena Williams | |  Simona Halep |
|--|-------------------------|---|
| 2 | Aces | 1 |
| 1 | Double faults | 0 |
| 68% | First serve % | 76% |
| 59% | Win % on 1st serve | 83% |
| 50% | Win % on 2nd serve | 45% |
| 0/1 | Break points | 4/5 |
| 0 | Tiebreaks won | 0 |
| 12 | Receiving points won | 21 |
| 38 | Points won | 55 |
| 4 | Games won | 12 |
| 1 | Max games won in a row | 5 |
| 5 | Max points won in a row | 7 |

Na obrázku vidíme štatistiku finálového stretnutia medzi Simonom Halepovou z Rumunska a Serenou Williamsovou z USA. Každé číslo v tejto štatistike má svoj vlastný „príbeh“ a niečo vyjadruje. Vpravo hore na obrázku je tabuľka, ktorej prvý a druhý stĺpec obsahuje rovnaké hodnoty a to 2 a 6. Každý stĺpec predstavuje jeden set stretnutia a každý riadok uvádza koľko hier (gemov) v ňom každá hráčka získala. V druhej časti obrázka sú ďalšie informácie o stretnutí. Tieto čísla môžeme interpretovať ako odpovede na konkrétne otázky o zápase:

- Koľko es dosiahla Serena Williamsová?
 - Odpoveď: 2
- Koľko dvojčieb urobila Simona Halepová?
 - Odpoveď: 0
- Aká bola úspešnosť prvého podania Simony Halepovej?

- Odpoveď: 76%
- Koľko hier (gemov) vyhrala Serena Williamsová?
 - Odpoveď: 4 (2 v prvom setu a 2 v druhom setu, čož je uvedeno v pravém horním rohu)

Práve uvedený príklad je ukážkou spracovania údajov získaných počas tenisového stretnutia. Môžeme si však položiť aj iné otázky, napríklad ktorá tenistka získala vo Wimbledone najviac singlových titulov rokoch 1968 až 2019. Na zodpovedanie tejto otázky potrebujeme iné údaje:



© Statista

Tentoraz musíme získať údaje o tom, kto vyhral finálové stretnutie vo Wimbledone v rokoch 1968 až 2019.

Bežné kroky, ktoré vykonáva štatistik pri svojej práci sú:

1. Zber údajov:
 - Súbor víťaziek ženských finálových stretnutí v rokoch 1968 až 2018.
2. Organizácia údajov do tabuliek a grafov:
 - Usporiadanie hráčok podľa počtu získaných singlových titulov.
3. Vyvodzovanie záverov z analýzy údajov:
 - Z vyššie uvedeného grafu vyplýva, že najviac titulov získala Martina Navrátilová.

Ak sú tieto tri kroky súčasťou štatistického spracovania, tak hovoríme o **popisnej štatistike**. Nespracované údaje majú zväčša formu zoznamu, súboru, či databázy.

Ako zberať nespracované dáta?

Ako je z príkladov zrejmé, údaje s ktorými pracujeme môžu byť rôzneho typu. Na údaje sa môžeme pozeráť ako na výsledok:

1. **Dotazníku**

- Súbor otázok (položiek) v ktorých môžeme buď odpoveď napísať alebo si vyberieme niektorú z ponúkaných možností.
- Príklad: sčítanie obyvateľstva, voľby

2. **Evidencie**

- Alebo denník do ktorého si pravidelne zaznamenávame informácie.
- Príklad: nemocničné záznamy (lieky podávané každých 8 rokov), vonkajšia teplota (teplota vzduchu každý deň o 7:00 ráno), víťazstvo vo Wimbledone (víťazka ženského finále, raz ročne).

Sčítanie obyvateľstva je dobrým príkladom toho, ako zo sledovanej populácie zbierame informácie. Na základe údajov získaných zo sčítania obyvateľstva môže vedenie štátu prijímať rozhodnutia napríklad o tom ako prerozdeľovať peniaze do viac, či menej obývaných oblastí a podobne.

Avšak vo väčšine prípadov nedokážeme získať informácie od celej populácie (základného súboru). Dobrým príkladom je reklamná agentúra, ktorú zaujíma, ktorý televízny program je najsledovanejší. S touto informáciou môže spoločnosť nakúpiť reklamný čas v tomto programe a maximalizovať dopad reklamnej kampane v populácii. Určite by bolo časovo aj finančne nákladné postupovať rovnako ako v prípade sčítania obyvateľstva a teda snažiť sa osloviť celú cieľovú skupinu. Praktickejší postup získania tejto informácie vyžaduje to, čo sa nazýva **výberový súbor**. Výberový súbor je podmnožinou (časťou) nami zvolenej populácie.

Voľba výberového súboru musí spĺňať určité kritériá:

Nicméně ve většině případů nejsme schopni oslovit celou zájmovou populaci. Uvažujme např. reklamní agenturu, kterou zajímá, který televizní pořad je nejsledovanější. S touto informací může agentura koupit reklamní čas v daných pořadech a maximalizovat dopad na populaci. Určitě by bylo velmi nákladné a časově náročné imitovat sčítání lidu a snažit se oslovit každého v publiku. Tato studie vyžaduje to, co nazýváme **vzorek**. Vzorek je část „populace“, kterou uvažujeme.

1. Výberový súbor musí **reprezentovať celú skúmanú populáciu**,
2. Výberový súbor musí obsahovať musí byť **dostatočne veľkýozumný počet**,
3. Mali by sa použiť vhodné **spôsoby výberu vzorky a náhodný výběr**.

NÁJDITE NEJAKÝ VHODNÝ PRÍKLAD VÝBEROVÉHO SÚBORU a HLASOVANIA (MÔŽETE POUŽIŤ INTERNETOVÚ STRÁNKU YUGOV) a JEDNODUCHO JU VYSVETLITE!!!!

Reklamná agentúra si majme priskumnú agentúru YouGov, aby im pomohla pochopiť v akom vysielacom čase by bolo v Spojenom Kráľovstve vhodné inzerovať určitý produkt. YouGov vytvorí online dotazník, ktorý následne zašle svojim registrovaným užívateľom, ktorí reprezentujú všetky vekové skupiny, socio-ekonomické skupiny a ďalšie demografické skupiny. Užívateľskú skupinu tvorí v Spojenom Kráľovstve v súčasnosti približne 1 milión užívateľov.

Ako vyzerajú nespracované údaje?

Spôsob ako organizovať údaje sme uviedli v úvode kapitoly. Použili sme dva príklady prezentácie údajov na základe výsledkov ženských finálových stretnutí vo Wimbledone.

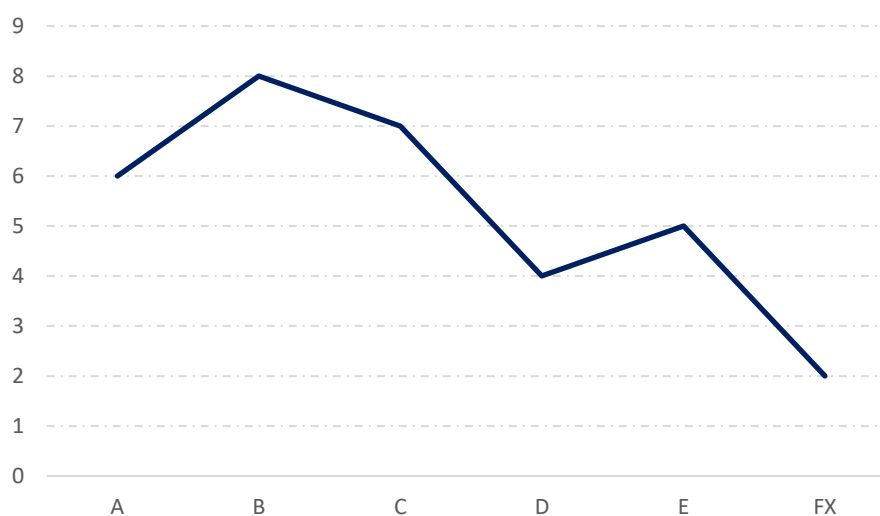
Pozrime sa teraz na tradičnejší príklad. Trieda 32 študentov napísala test s nasledovnými výsledkami: 82, 90, 61, 80, 88, 96, 58, 74, 69, 82, 53, 78, 72, 58, 54, 95, 99, 86, 80, 95, 99, 88, 83, 51, 65, 77, 28, 82, 70, 73, 100, 19

Hranicou pre úspešné absolvovanie testu je 50 bodov. Každý študent, ktorý dosiahne viac ako 50 bodov predmet úspešne absolvuje a dosiahne hodnotenie podľa kľúča uvedeného v tabuľke. Nižšie uvádzame nástroje organizácie a reprezentácie údajov:

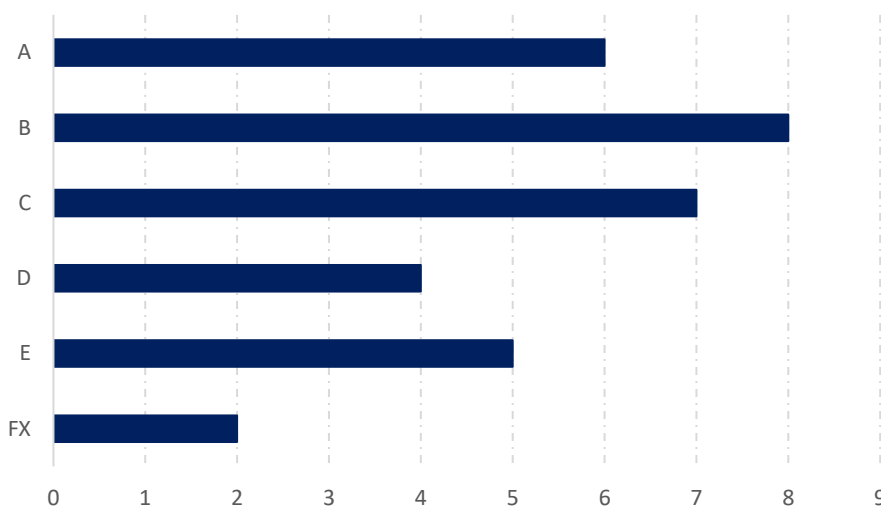
1. Tabuľka:

| Hodnotenie | Skóre | # študentov | % študentov |
|------------|----------|-------------|-------------|
| A | 100 - 91 | 6 | 19% |
| B | 90 - 81 | 8 | 25% |
| C | 80 - 71 | 7 | 22% |
| D | 70 - 61 | 4 | 13% |
| E | 60 - 51 | 5 | 16% |
| FX | 50 - 0 | 2 | 6% |

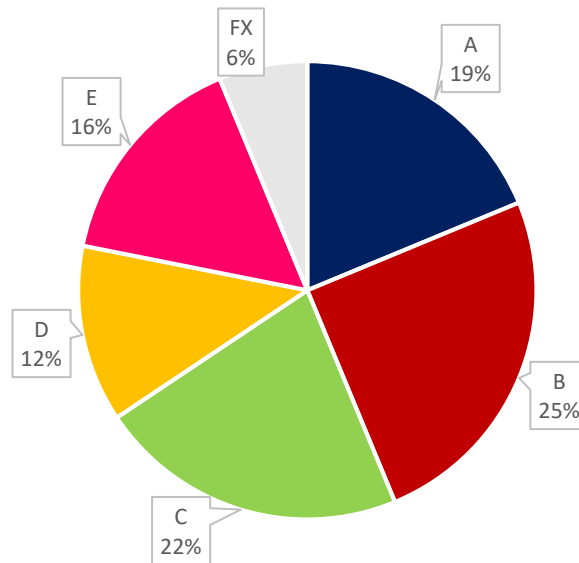
2. Spojnicový graf:



3. Stĺpcový graf:



4. Koláčový graf:



Čo vyjadruje priemer, medián a modus?

Ako údaje zberať, organizovať a prezentovať sme si už povedali. Púhym pohľadom na grafy vieme zistiť informáciu o najčastejšie sa vyskytujúcom hodnotení. K tomu, aby sme porozumeli i ďalším informáciám, si najskôr musíme vysvetliť pojmy ako priemer, modus a medián.

1. Priemer:

- V štatistike sa aritmetický priemer nazýva **priemer**.
- V našom prípade nás môže zaujímať, aký bol priemerný počet získaných bodov. v teste alebo to, aký bol priemerný počet bodov na každom stupni hodnotenia.
- **Priemer množiny N čísel vypočítame tak, že všetky čísla sčítame a vydělíme ich ich počtom.**
- Prvý priemer tak bude vyžadovať, aby sme počty bodov všetkých študentov, ktorí písali test = 2385) a tento súčet vydělíme ich počtom (= 32); priemer počtu bodov je **74,53**.
- Priemerný počet bodov na jednotlivých stupňoch hodnotenia bude vyžadovať jeden krok navyše.
- V tomto kroku musíme identifikovať počty bodov, ktoré dosiahli študenti s daným hodnotením (pre hodnotenie A: 96, 95, 99, 95, 99 a 100).
- Jakmile tak učiníme, postupujeme stejným způsobem jako v předchozím případě; sečteme skóre a vydělíme počtem studentů s daným hodnocením (tj. $(96+95+99+95+99+100)/6 = 97,33$).

| Hodnocení | Průměr |
|-----------|--------|
| A | 97,33 |
| B | 85,13 |
| C | 76,29 |
| D | 66,25 |
| E | 54,80 |
| FX | 23,50 |
| Celkem | 74,53 |

2. Medián:

- Často je výhodné zistiť, ktorá hodnota sa nachádza uprostred údajov.

- **Medián je prostredná hodnota v množine usporiadaných hodnôt; v prípade, že postupnosť má párny počet členov, je za prostrednú hodnotu považovaný aritmetický priemer dvoch prostredných hodnôt.**
- Prečo by nás mala zaujímať práve táto hodnota?
- Predstavme si päť tanierov so sušienkami; na štyroch z nich je po päť sušienok a piaty je prázdny.
- Do kuchyne príde päť detí, každé si vezme jeden tanier a zje všetko, čo je na ňom.
- Priemer nám hovorí, že na jedno dieťa pripadli $(5+5+5+5+0)/5 = 4$ sušienky.
- My však vieme, že jednému z detí sa neušla žiadna sušienka a určite by nesúhlasilo s tvrdením, že v priemere zjedlo 4 sušienky.
- Medián počtu sušienok na tanieroch je prostredná hodnota postupnosti : 0, 5, 5, 5, 5 => medián = 5.
- Predstavme si päť talířů sušenek; na čtyřech talířích je po pěti sušenkách a pátý je prázdny.
- Do kuchyně přijde pět dětí, každé si vezme jeden talíř a sní vše, co je na něm.
- Průměr nám říká, že jsme měli $(5+5+5+5+0)/5 = 4$ sušenky na dítě.
- Nicméně my víme, že jedno dítě nemělo ani sušenku a určitě by nesouhlasilo s tvrzením, že v průměru snědlo 4 sušenky.
- Medián těchto talířů bude střední hodnota této posloupnosti: 0, 5, 5, 5, 5 => medián = 5.
- Priemer nám prezrádza, že na jednom alebo viacerých tanieroch bolo menej ako päť sušienek a niektoré deti môžu byť stále hladné.
- Na druhej strane, môžeme mať štyri taniere prázdne a na jedenj päť sušienok; priemer hovorí, že každé dieťa zjedlo jednu sušienku $(0+0+0+0+5)/5 = 1$, ale medián ponúka realistickejší pohľad.
- V tomto prípade je medián rovný nule a preto by sme sa mali detailnejšie pozrieť na to, ako boli sušienky rozdelené a či náhodou niektoré z detí nie sú hladné.
- V tabuľke nižšie uvádzame medián počtu bodov z testu pre každý stupeň hodnotenia. Čo vyplýva z výsledkov?
- V tomto prípade je medián roven 0 a preto bychom sa mali bližšie podívať na to, jak byly sušenky rozděleny a kolik dětí má ještě hlad.
- Níže uvádíme medián pro každý stupeň hodnocení a celkový medián; co z výsledků vyčteme?

| Hodnotenie | Medián |
|------------|--------|
| A | 97.50 |
| B | 84.50 |
| C | 77.00 |
| D | 67.00 |
| E | 54.00 |
| FX | 23.50 |
| Celkovo | 79.00 |

3. Modus:

- Pri analýze údajov nás môže zaujímať, ktorá hodnota sa vyskytuje najčastejšie; takáto hodnota sa nazýva **modus**.
- Čísla opäť zoradíme do neklesajúcej postupnosti a pozrieme sa, ktorá hodnota sa v nej vyskytuje najčastejšie.

- V našom prípade je to číslo 82 **82**; traja študenti získali 82 bodov, po dvaja študenti dosiahli 58, 80, 88, 95 a 99 bodov.
- Čo by sa stalo, keby traja študenti dosiahli 82 bodov a rovnakotraja študenti 58 bodov?
- V takom prípade by sme museli konštatovať, že množina má ***dva modusy*** a je teda ***bimodálna***.
- A čo keď žiadny dvaja študenti nedosiahnu rovnaký počet bodov? (tj. Každá z hodnôt v postupnosti sa nachádza práve raz).

Zdroje:

<https://www.statista.com/statistics/280393/womens-tennis-players-with-the-most-victories-at-wimbledon/>

<https://yougov.co.uk/about/panel-methodology/>

Dressler and Keenan: Integrated Mathematics (Second Edition)

(https://books.google.sk/books/about/Integrated_Mathematics_Course_1.html?id=h4oiXcjGT3oC&redir_esc=y)