Statistická analýza váhy policistů

Vaše jméno

January 29, 2025

1 Úvod

Tento dokument obsahuje statistickou analýzu kvantitativní proměnné – váhy policistů. Použité metody zahrnují výpočet deskriptivních statistik, měření variability a vizualizaci dat pomocí histogramu a boxplotu. Každá statistika je doplněna o vzorce a podrobnou interpretaci.

2 Data

Data obsahují váhy policistů v kilogramech. Proměnná `weight` je numerická a spojitá.

3 Deskriptivní statistiky

3.1 Aritmetický průměr

Aritmetický průměr je definován jako součet všech hodnot dělený počtem hodnot:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Výsledek:

$$\bar{x} = 78.448\,\mathrm{kg}$$

Interpretace: Průměrná váha policistů je 78.448 kg. Tato hodnota je citlivá na extrémní hodnoty (outliery).

3.2 Vážený průměr

Vážený průměr je definován jako:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

kde w_i jsou váhy jednotlivých hodnot. V našem případě jsou všechny váhy stejné, proto je vážený průměr shodný s aritmetickým průměrem:

$$\bar{x}_w = 78.448\,\text{kg}$$

3.3 Upravený průměr

Upravený průměr (trimmed mean) odstraní určité procento nejnižších a nejvyšších hodnot, aby se omezil vliv outlierů. Pro $\alpha = 0.10$:

$$\bar{x}_{\text{trim}} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}$$

kde $k = \lfloor \alpha n \rfloor$. Výsledek:

$$\bar{x}_{\text{trim}} = 77.9025\,\text{kg}$$

Interpretace: Po odstranění 10 % nejnižších a nejvyšších hodnot je průměrná váha 77.9025 kg. To naznačuje, že extrémní hodnoty mírně zvyšují aritmetický průměr.

3.4 Medián

Medián je prostřední hodnota seřazených dat:

$$\text{Medián} = x_{(n+1)/2} \quad \text{(pro lichý počet dat)}$$

Výsledek:

$$\text{Medián} = 77.95\,\text{kg}$$

Interpretace: Polovina policistů má váhu nižší než 77.95 kg a polovina vyšší. Medián je robustní vůči extrémním hodnotám.

3.5 Huberův odhad

Huberův odhad je robustní metoda kombinující prvky aritmetického průměru a mediánu. Pro výpočet se používá iterativní algoritmus, který minimalizuje kombinaci kvadratické a lineární ztrátové funkce. Výsledek:

$$\text{Huberův odhad} = 78.173\,15\,\text{kg}$$

Interpretace: Tento odhad je méně ovlivněn extrémy než aritmetický průměr, ale stále zohledňuje informace z celého rozsahu dat.

3.6 Tukeyho čísla

Tukeyho čísla zahrnují minimum, první kvartil (Q1), medián (Q2), třetí kvartil (Q3) a maximum:

$$\text{Tukeyho čísla} = (55.10\,\text{kg}, 69.60\,\text{kg}, 77.95\,\text{kg}, 87.00\,\text{kg}, 102.20\,\text{kg})$$

Interpretace:

- Minimální váha: 55.10 kg
- 25 % policistů má váhu nižší než 69.60 kg
- Medián: 77.95 kg
- 75 % policistů má váhu nižší než 87.00 kg
- Maximální váha: 102.20 kg

4 Měření variability

4.1 Interquartile Range (IQR)

IQR je rozdíl mezi třetím a prvním kvartilem:

$$\mathrm{IQR} = Q3 - Q1 = 87.00\,\mathrm{kg} - 69.60\,\mathrm{kg} = 17.15\,\mathrm{kg}$$

Interpretace: Střední 50 % dat má rozptyl 17.15 kg.

4.2 Rozptyl

Rozptyl je průměrný kvadratický rozdíl od aritmetického průměru:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Výsledek:

$$\sigma^2 = 131.1711\,\mathrm{kg}^2$$

Interpretace: Průměrný kvadratický rozdíl od průměru je 131.1711 kg^2. Vyšší rozptyl indikuje větší variabilitu dat.

4.3 Směrodatná odchylka

Směrodatná odchylka je odmocnina rozptylu:

$$\sigma = \sqrt{\sigma^2} = 11.453\,\mathrm{kg}$$

Interpretace: Průměrná odchylka od průměru je 11.453 kg. Vyšší hodnota znamená větší rozptyl dat.

4.4 Variační koeficient

Variační koeficient je poměr směrodatné odchylky k průměru:

$$\mathrm{CV} = \left(\frac{\sigma}{\bar{x}}\right) \times 100 = 14.599\,47\,\%$$

Interpretace: Relativní variabilita dat je 14.599 47 %. Tato hodnota umožňuje srovnání variability mezi různými soubory dat.

4.5 Mediánová absolutní odchylka (MAD)

MAD je robustní měření variability:

$$\text{MAD} = \text{med}(|x_i - \text{med}(x)|) = 12.750\,36\,\text{kg}$$

Interpretace: Průměrná absolutní odchylka od mediánu je 12.750 36 kg. Tato metrika je méně citlivá na extrémy.

5 Šikmost a špičatost

5.1 Šikmost

Šikmost měří asymetrii rozdělení:

$$\text{Šikmost} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3 = 0.2932968$$

Interpretace: Pozitivní šikmost naznačuje, že data jsou mírně nakloněna k nižším hodnotám.

5.2 Špičatost

Špičatost měří koncentraci hodnot kolem modu a výskyt extrémních hodnot:

$$\text{Špičatost} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^4 - 3 = -0.8435781$$

Interpretace: Negativní špičatost naznačuje, že distribuce má "lehké konce" (nižší výskyt extrémních hodnot).

6 Vizualizace

6.1 Histogram

Histogram ukazuje frekvenční rozdělení váhy policistů. Většina hodnot se nachází mezi 70 kg a 90 kg, což potvrzuje mírně pozitivní šikmost.

histogram.png

Figure 1: Histogram váhy policistů

6.2 Boxplot

Boxplot zobrazuje Tukeyho čísla a potvrzuje, že střední 50 % dat je mezi 69.60 kg a 87.00 kg. Nejsou zde viditelné extrémní odlehlé hodnoty.

boxplot.png

Figure 2: Boxplot pro váhu policistů

7 Závěr

Analýza ukazuje, že váhy policistů jsou přibližně normálně rozdělené s mírnou pozitivní šikmostí. Variabilita dat je střední, s relativně nízkým výskytem extrémních hodnot. Robustní metody (medián, upravený průměr, Huberův odhad) potvrzují stabilitu dat.