BIOSTATISTIKA

L01

Kolik váží „typický“ savec?



Úvod do biostatistiky

cuni-natur-biostatistics.github.io/materialy.html

Ondřej Mottl



- 🌿 **Vegetační ekolog** na katedře botaniky
- 👨‍💻 vedoucí **Laboratoře kvantitativní ekologie**
- ♻️ zastánce **otevřené a reprodukovatelné vědy**

> Jak z velkého množství biologických dat “vyždímat” nové informace

Foto: Vladimír Šigut, Univerzita Karlova

Co mě zajímá?

Biodiverzita

napříč měřítky



Velká data

nové metody



Otevřená věda

reprodukovatelnost



Ilustrace vytvořené pomocí OpenAI, 2026

Co se od vás očekává?

Příprava → společné přemýšlení → vlastní práce s daty

Před výukou

Krátce projdete materiály (skripta).

90 minut přednášky

Aktivní účast na přednášce.

120 minut praktika

V jedné z pěti paralelních skupin budete analyzovat data v R.

Materiály najdete na webu kurzu (HUB)

cuni-natur-biostatistics.github.io
bit.ly/CUNI-MB120P163

Skripta

HTML pro interaktivní čtení
PDF pro tisk

Prezentace

HTML pro prohlížení v prohlížeči
PDF pro tisk

R skripty

Kód pro praktickou práci s daty v R.

SIS slouží pro zápis a termíny zkoušek. Moodle slouží pro testy, odevzdávání, výsledky a neveřejná oznámení.

Výuka je dobrovolná, kompetence povinné

Na přednášky a praktika chodit nemusíte.

Ve výuce

Získáte vysvětlení, podporu a zpětnou vazbu.

Při samostudiu

Všechny požadované znalosti najdete ve skriptech a cvičeních v R.

Povinně

Absolvujete praktický zápočet a závěrečnou zkoušku.

Učit se můžete prezenčně i doma — **stejné kompetence** však musí prokázat každý.

Zápočet a zkouška ověřují jiné dovednosti

PRAKTICKÝ ZÁPOČET

Dokážete samostatně provést a vysvětlit základní analýzu v R?

- malý dataset a biologická otázka
- částečně připravený R skript
- kontrola dat, graf, model a biologický závěr

ZÁVĚREČNÁ ZKOUŠKA

Rozumíte analytickým rozhodnutím a interpretaci výsledků?

- dodaný R kód, grafy a modelové výstupy
- interpretace, diagnostika a nejistota
- ne psaní syntaxe zpaměti

AI můžete používat při učení.
U zápočtu a zkoušky pracujete samostatně bez internetu a AI.

Pravidla hodnocení: web kurzu (HUB). Termíny zkoušek: SIS. Neveřejné pokyny a výsledky: Moodle.

Tři rychlé otázky na rozjezd

Odpovězte samostatně na tři krátké otázky. Společné odpovědi projdeme po hlasování.

Co se v tomto kurzu pokusíme z biologických dat „vyždímat“?

- Co nejvíce tabulek
- Nové informace
- Správnou odpověď za každou cenu

Správná odpověď: Nové informace
Cílem je získávat z biologických dat nové informace a umět přitom vysvětlit, jak jsme k závěru došli.

Kdy můžete v tomto kurzu používat AI?

- Nikdy
- Při učení a přípravě, ale ne u zápočtu a zkoušky
- Také u zápočtu a zkoušky, pokud použití AI uvedete

Správná odpověď: Při učení a přípravě, ale ne u zápočtu a zkoušky
AI můžete používat při učení. U praktického zápočtu a závěrečné zkoušky pracujete samostatně bez internetu a AI.

Kde najdete veřejné materiály ke každé lekci?

- Na webu kurzu (HUB)
- V Moodle
- V SIS

Správná odpověď: Na webu kurzu (HUB)
Veřejné materiály jsou na webu kurzu (HUB). Moodle slouží pro testy, odevzdávání, výsledky a neveřejná oznámení; SIS pro zápis a termíny zkoušek.

Kolik váží běžný savec?

Tabulka savců

druh	typ_potravy	rad	spanek_h	hmotnost_kg
Cheetah	masožravec	Carnivora	12.1	50.00
Owl monkey	všežravec	Primates	17.0	0.48
Mountain beaver	býložravec	Rodentia	14.4	1.35

Ukázka z tabulky `msleep` (celkem 83 řádků).

SMYŠLENÝ TITULEK A

Vědci spočítali:
„Průměrný savec váží **166 kg**.“

SMYŠLENÝ TITULEK B

Jiná analýza tvrdí:
„Běžný savec váží jen **1,67 kg**.“

Stejná tabulka. Dvě dramaticky odlišná čísla.

Výsledky učení

Po dnešní přednášce budete schopni:

- vysvětlit, co v tabulce představuje jeden řádek
- přečíst vhodný graf jedné proměnné
- vybrat smysluplnou dvojici středu a rozptýlení
- rozlišit čtyři typy proměnných a zvolit jejich první popis
- vysvětlit, proč jedno správně vypočtené číslo ještě nemusí být dobrý titulek

Jak budeme pracovat?

Otázka → data → graf → souhrn → (opatrná interpretace)

NA PŘEDNÁŠCE

Budete odhadovat, hlasovat, porovnávat grafy a obhajovat závěry.

V MATERIÁLECH

HTML nabízí záložky a rozbalovací otázky; PDF je vhodné pro souvislé čtení.

TITULEK A

„Průměrný savec váží **166 kg**.“

TITULEK B

„Běžný savec váží **1,67 kg**.“

Hlasování: který titulek je správný?

A) Pouze titulek A.

B) Pouze titulek B.

C) Oba mohou být správně.

D) Zatím nemáme dost informací.

Hlasujte samostatně a svou volbu pak obhajte sousedovi.

Přesvědčte souseda

Proč jste vybrali právě svoji možnost?

DŮKAZ, KTERÝ MÁTE K DISPOZICI

- stejná tabulka savců;
- 166 kg versus 1,67 kg

Odpověď zatím odložíme

Zjistíme, co bylo změřeno a jak hodnoty vypadají.

> Správný výpočet ještě nezaručuje správnou interpretaci.

Od organismu k datům

Začneme jedním savcem

Stejná tabulka obsahuje hmotnost i spánek. Na době spánku si ukážeme, jak data zobrazit a popsat. K otázce o hmotnosti se vrátíme s těmito nástroji.

Gepard (*Cheetah*)

Výzkumníci studovali spánkové návyky savců a u geparda zaznamenali **12.1 hodiny spánku za den**.



Ilustrace vytvořená pomocí AI.

> **PŘI ZÁZNAMU MĚŘENÍ**
> Výsledek pro jeden zastoupený druh zapíšeme do jednoho řádku tabulky.

První pozorování přidá první řádek

Druh (species)	Spánek za den (h)
Cheetah	12.1

> Jeden řádek zachovává spojení mezi **druhem** a jeho **hodnotou**.

Další druh přidá další řádek

Druh (species)	Spánek za den (h)
Cheetah	12.1
Owl monkey	17.0

> Tabulka roste **po řádcích**: jeden zastoupený druh po druhém.

Čtyři druhy, čtyři řádky

Druh (species)	Spánek za den (h)
Cheetah	12.1
Owl monkey	17.0
Mountain beaver	14.4
Greater short-tailed shrew	14.9

Na co přesně se vztahuje hodnota 12,1 hodiny v prvním řádku?

Co představuje jeden řádek?

A

Jednoho konkrétního geparda.

B

Jeden zastoupený druh savce.

C

Jednu hodinu měření.

Jeden řádek = jeden zastoupený druh

CO VÍME

Každá hodnota spánku patří jednomu **druhu**, který je v tabulce zastoupen.

CO Z TOHO NEPLYNE

Tabulka neobsahuje data o spánku všech jednotlivých savců na Zemi.

Jak uložíme jednu hodnotu v R?

```r
vec_spanek_ukazka <-
  c(
    12.1 # Cheetah
  )

vec_spanek_ukazka
```

```
[1] 12.1
```

VEKTOR (*VECTOR*)

Vektor je uspořádaná řada hodnot stejného typu.

Druhý řádek přidá druhou hodnotu

```r
vec_spanek_ukazka <-
  c(
    12.1, # Cheetah
    17.0  # Owl monkey
  )

vec_spanek_ukazka
```

```
[1] 12.1 17.0
```

CO ZNAMENÁ POŘADÍ?

První hodnota vektoru patří prvnímu řádku tabulky, druhá druhému řádku.

Stejné čtyři hodnoty zapíšeme do vektoru

```r
vec_spanek_ukazka <-
  c(
    12.1, # Cheetah
    17.0, # Owl monkey
    14.4, # Mountain beaver
    14.9  # Greater short-tailed shrew
  )
```

> Tabulka uchovává i jména druhů. Samotný vektor uchovává jen hodnoty a jejich pořadí.

Základní operace

```
1 length(
2   x = vec_spanek_ukazka
3 )
```

```
[1] 4
```

`length()` říká, kolik hodnot máme.

```
1 head(
2   x = vec_spanek_ukazka,
3   n = 2
4 )
```

```
[1] 12.1 17.0
```

`head()` ukáže první hodnoty (2 hodnoty v našem případě).

Celá tabulka přidává další proměnné

DATASET `msleep`

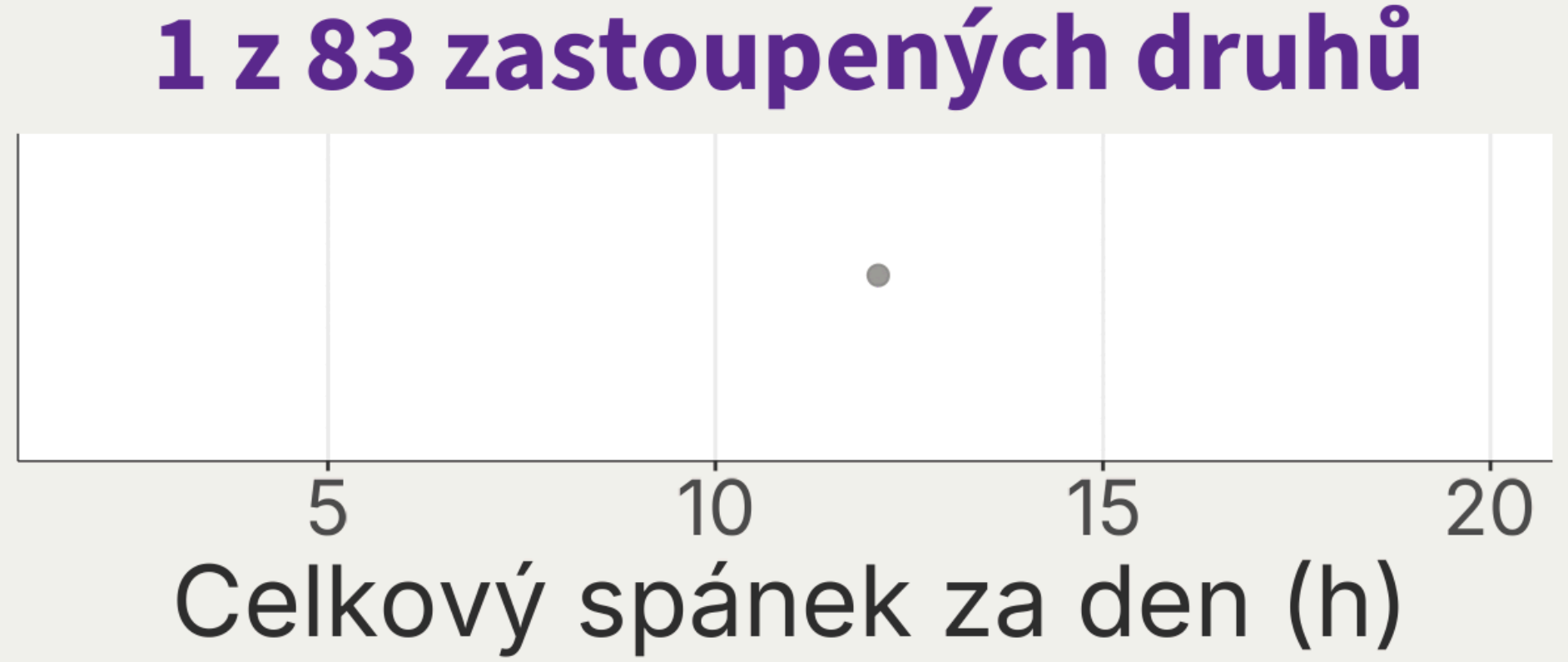
- **83 řádků:** zastoupené druhy savců;
- **5 sloupců:** například spánek, hmotnost, potrava nebo taxonomický řád;

druh	typ_potravy	rad	spanek_h	hmotnost_kg
Cheetah	masožravec	Carnivora	12.1	50.00
Owl monkey	všežravec	Primates	17.0	0.48
Mountain beaver	býložravec	Rodentia	14.4	1.35

Než popisujeme čísla, vždy si ujasníme **jednotku pozorování**, význam proměnné a jednotky měření.

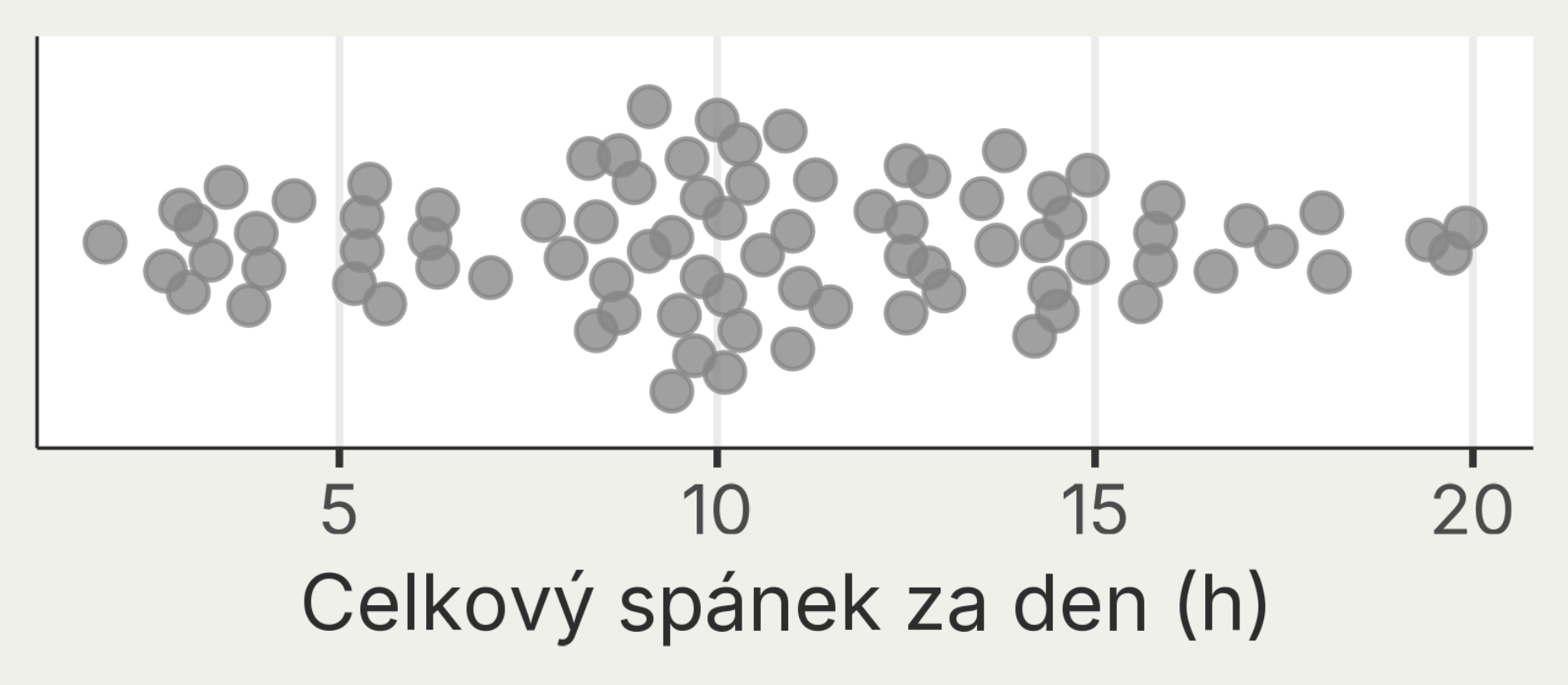
Od tabulky ke grafu

Data přibývají druh po druhu



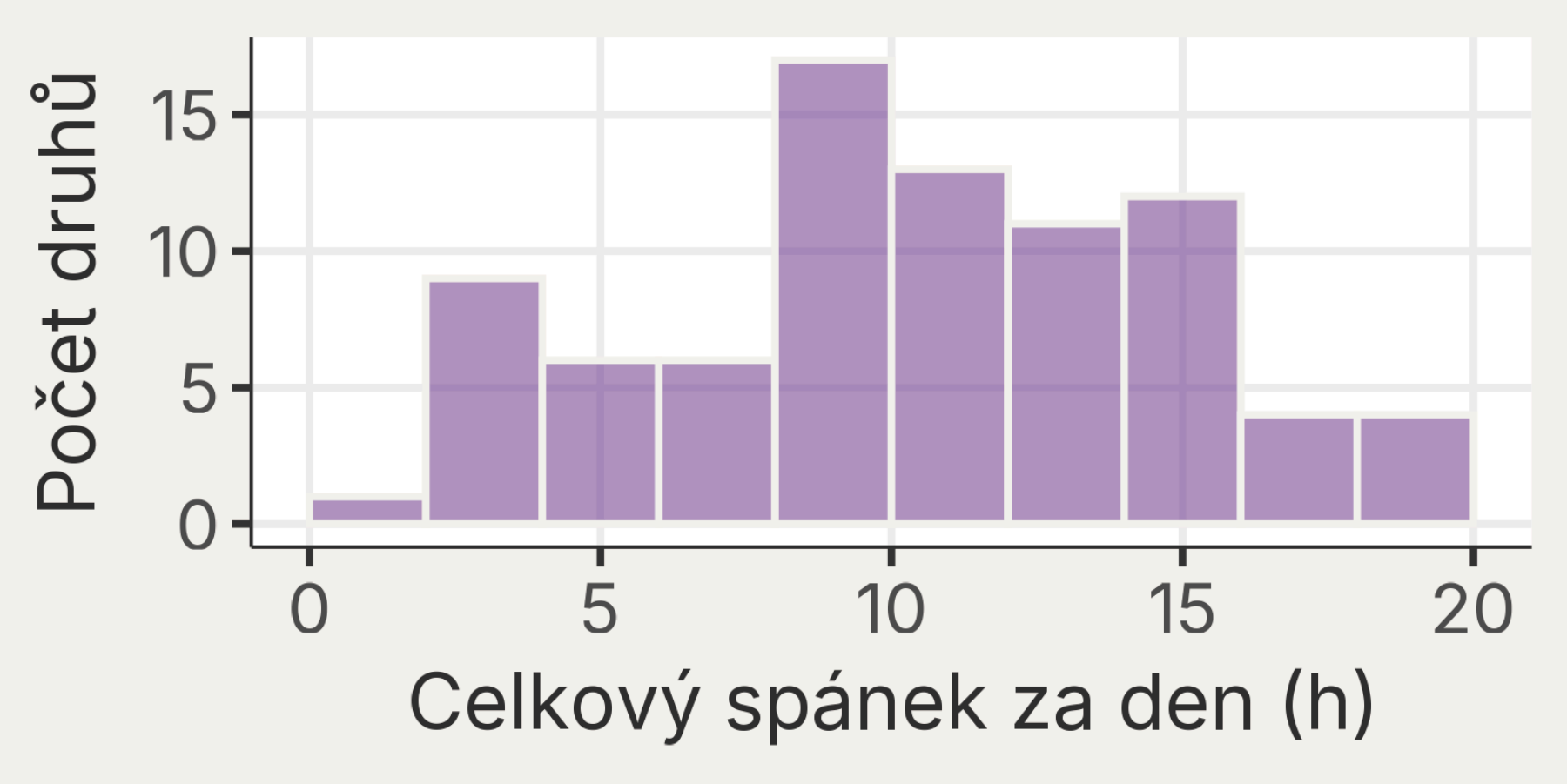
Graf nevzniká místo dat. Vzniká postupným zobrazením hodnot, které jsme vložili do tabulky a vektoru. Svislá poloha teček jen brání překryvu; osa y zde neznamená další měření.

Jak vypadá spánek všech zastoupených druhů?



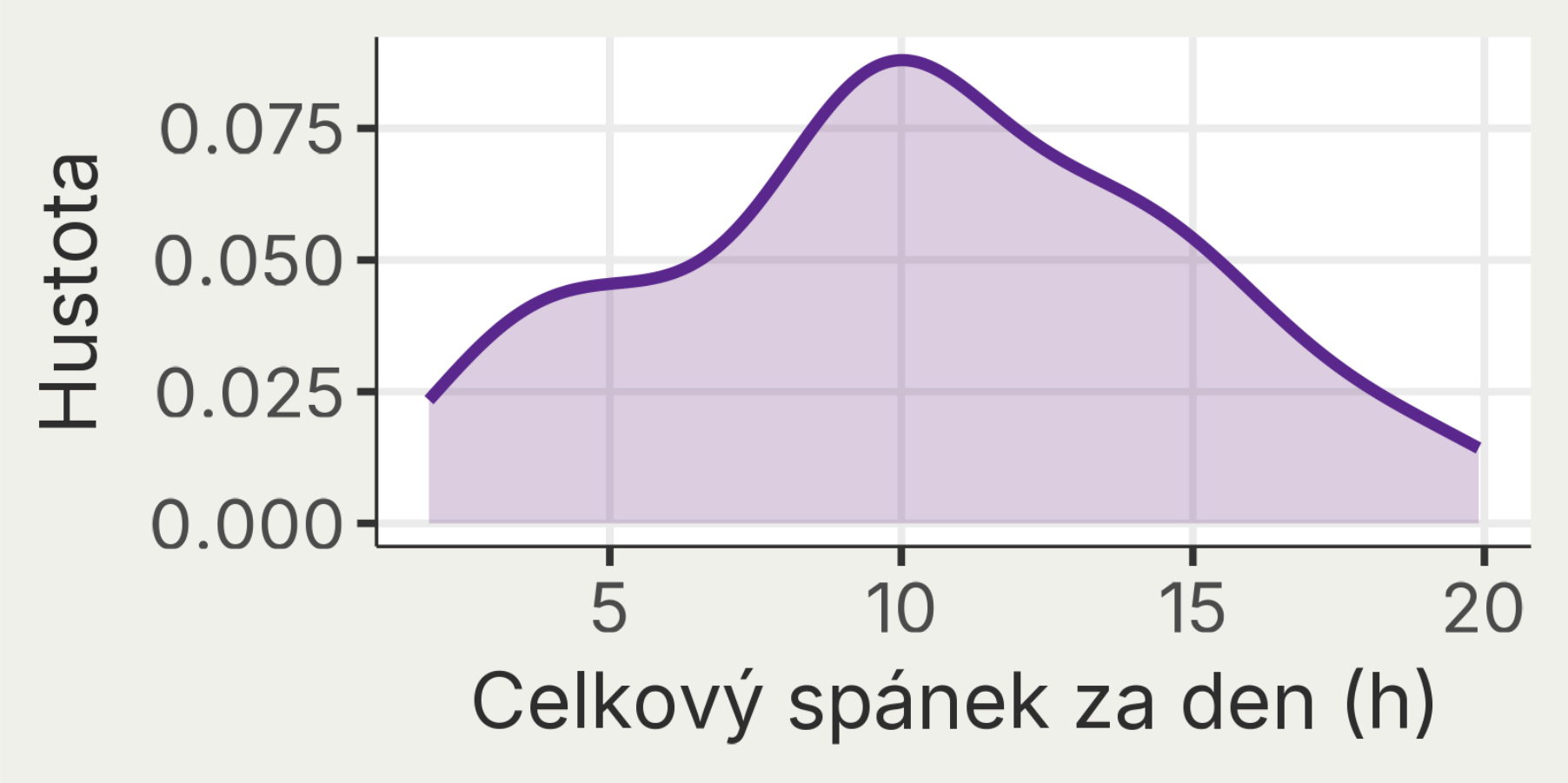
Kde leží nejkratší a nejdelší doba spánku? Kde se body hromadí?

Stejná data, seskupená do intervalů



Histogram ukazuje počty v intervalech, ale skrývá přesné hodnoty jednotlivých druhů.

Stejná data, ale jako hladký tvar



Graf hustoty vyhlazuje naměřené hodnoty do souvislé křivky. Výška ukazuje hustotu, nikoli počet druhů; jednotlivé hodnoty v něm nevidíme.

Který graf byste použili?

Potřebujete zjistit, zda některý druh spí téměř 20 hodin denně.

A	B	C
Bodový graf	Histogram	Graf hustoty

Když nehlasujeme: napište si, který graf zachovává jednotlivé hodnoty a proč je to pro tuto otázku důležité.

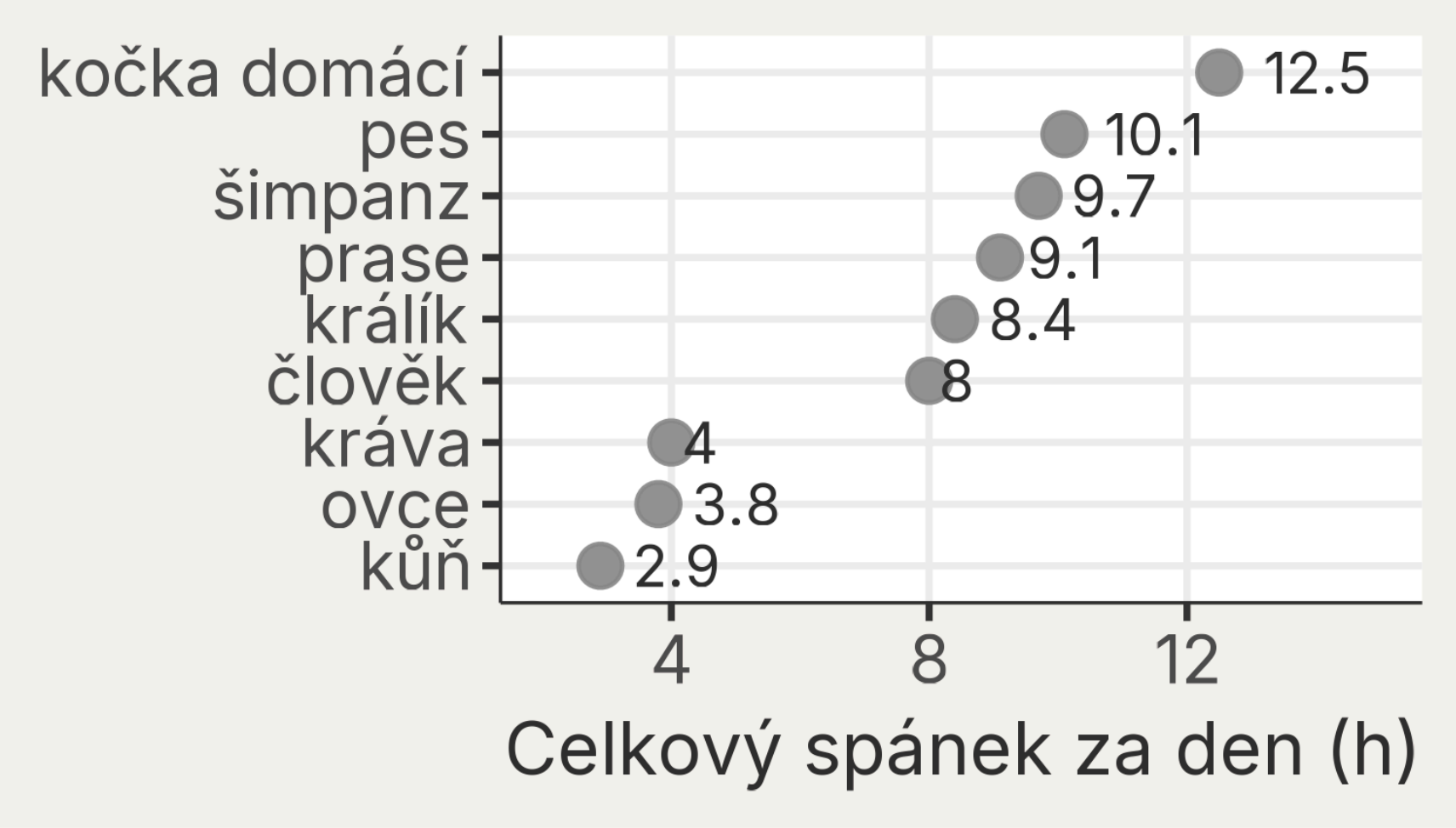
Tři grafy, tři různé otázky

Graf	Nejlépe ukáže	Co skrývá
Bodový graf	jednotlivé hodnoty	(při velkém počtu může být přeplněný)
Histogram	počty v intervalech	přesné hodnoty
Graf hustoty	hladký tvar	počty a konkrétní druhy

> Graf vybíráme podle otázky.

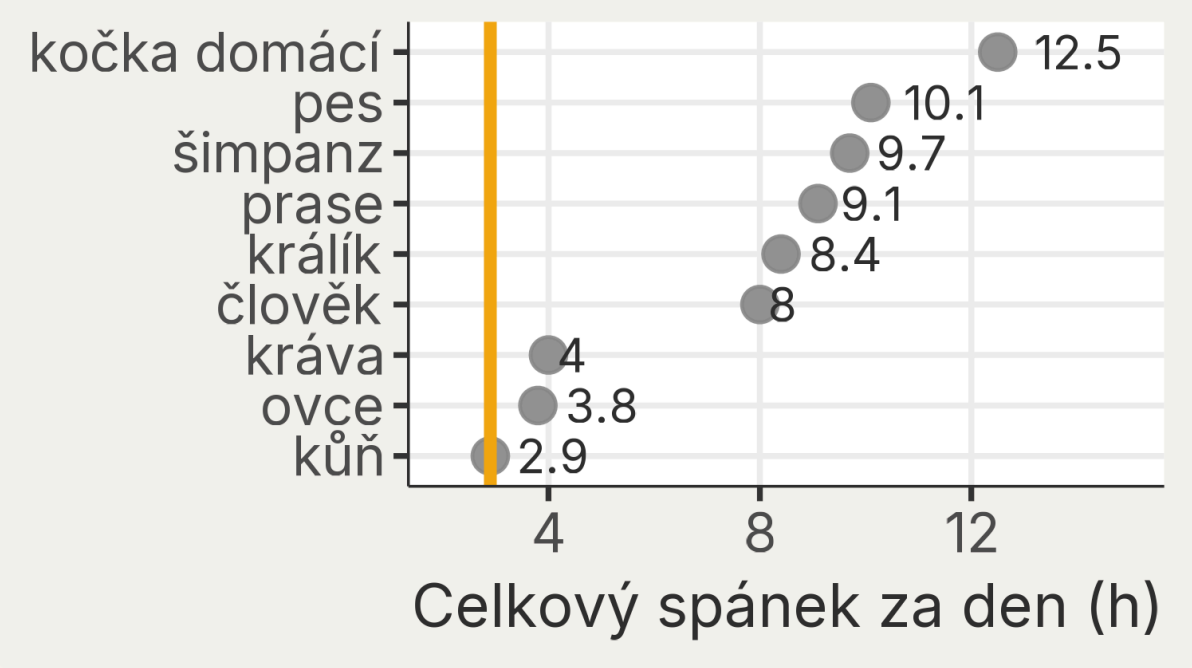
Deskriptivní statistiky

Devět druhů, devět dob spánku



Kam byste umístili jeden bod, který má zastoupit všech devět hodnot?

Minimum (*minimum*): nejmenší pozorovaná hodnota



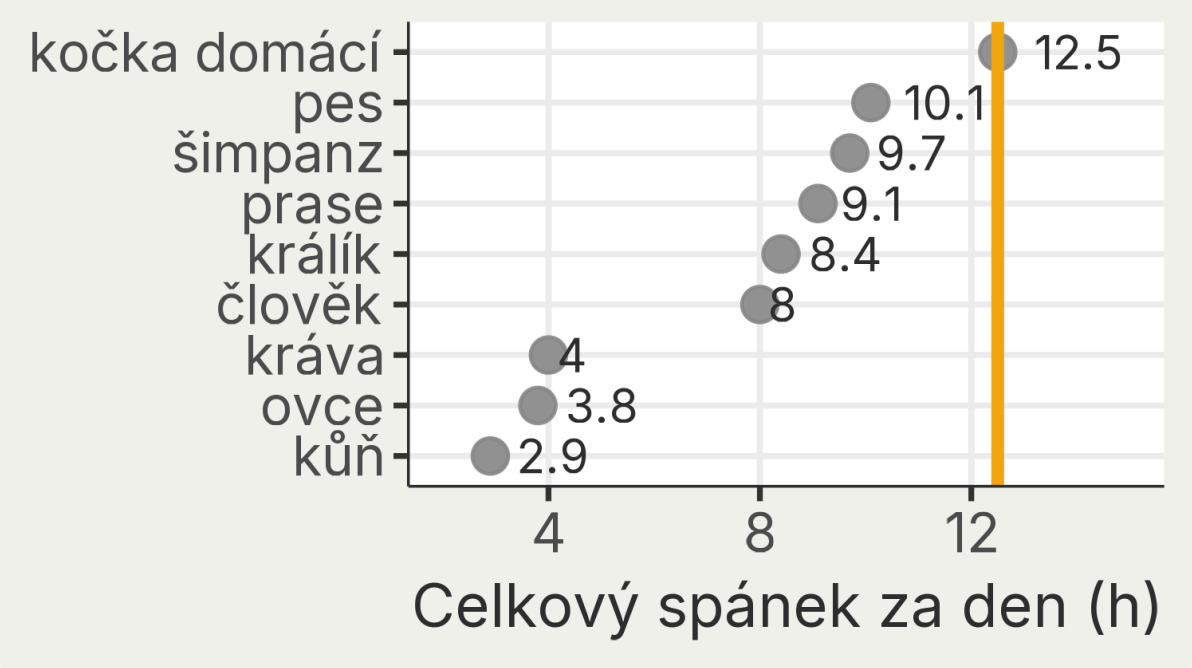

```r
vec_spanek_9 <- c(
  2.9, 3.8, 4, 8, 8.4,
  9.1, 9.7, 10.1, 12.5
)

min(
  x = vec_spanek_9
)
```

```
[1] 2.9
```

Minimum je zde **2.9 h**.

Maximum (*maximum*): největší pozorovaná hodnota



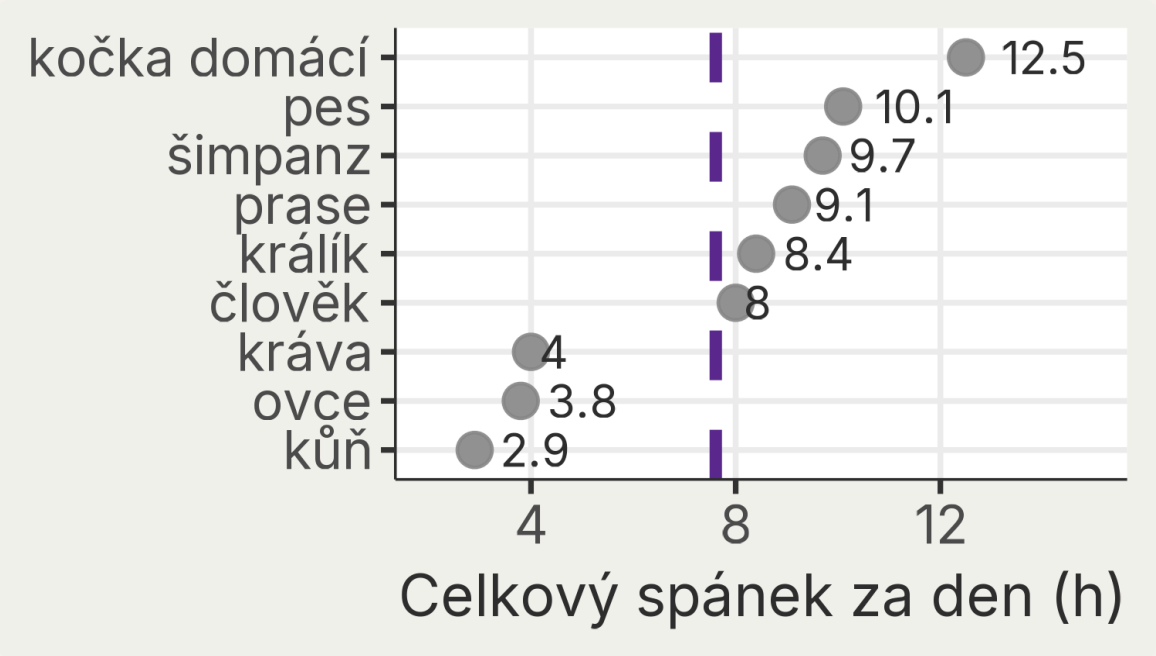

```
1 max(
2   x = vec_spanek_9
3 )
```

```
[1] 12.5
```

Maximum je zde **12.5 h**.

Minimum a maximum ukazují krajní pozorované hodnoty. Když přidáme další druhy, mohou se posunout dál; silně závisejí na velikosti souboru a neříkají, kde leží většina dat.

Průměr: hodnoty sečteme a rozdělíme



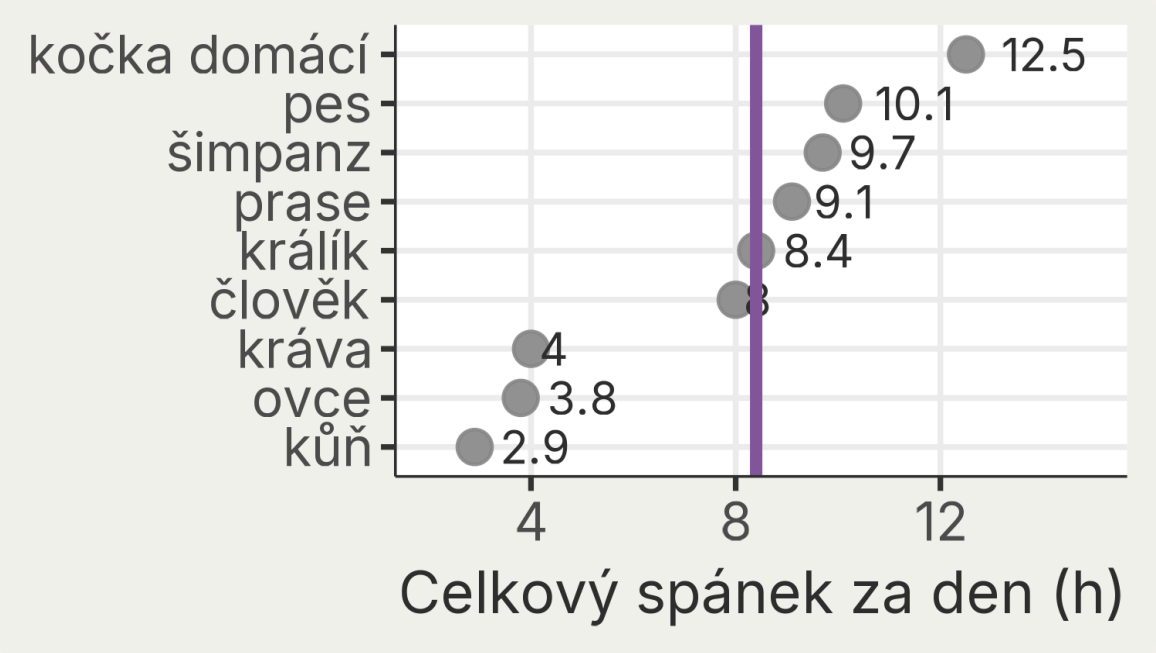

```r
mean(
  x = vec_spanek_9
)
```

```
[1] 7.611111
```

Průměr je zde **7,61 h**.

Aritmetický průměr (*mean*) používá každou hodnotu a může být citlivý na hodnoty daleko od ostatních.

Seřadíme hodnoty a hledáme prostřední



Medián = 8,4 h

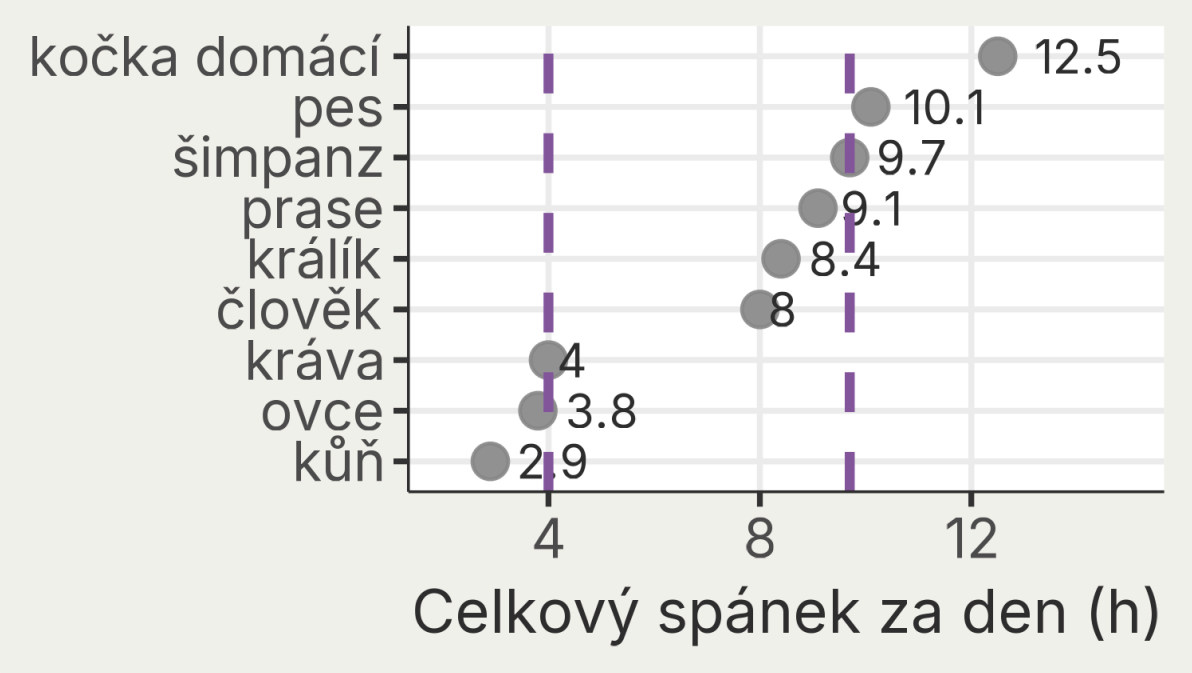
```
1 median(
2   x = vec_spanek_9
3 )
```

```
[1] 8.4
```

Čtyři hodnoty leží níže a čtyři výše.

Medián (*median*) je prostřední hodnota seřazeného souboru.

Kvartily rozdělí seřazené hodnoty na čtvrtiny



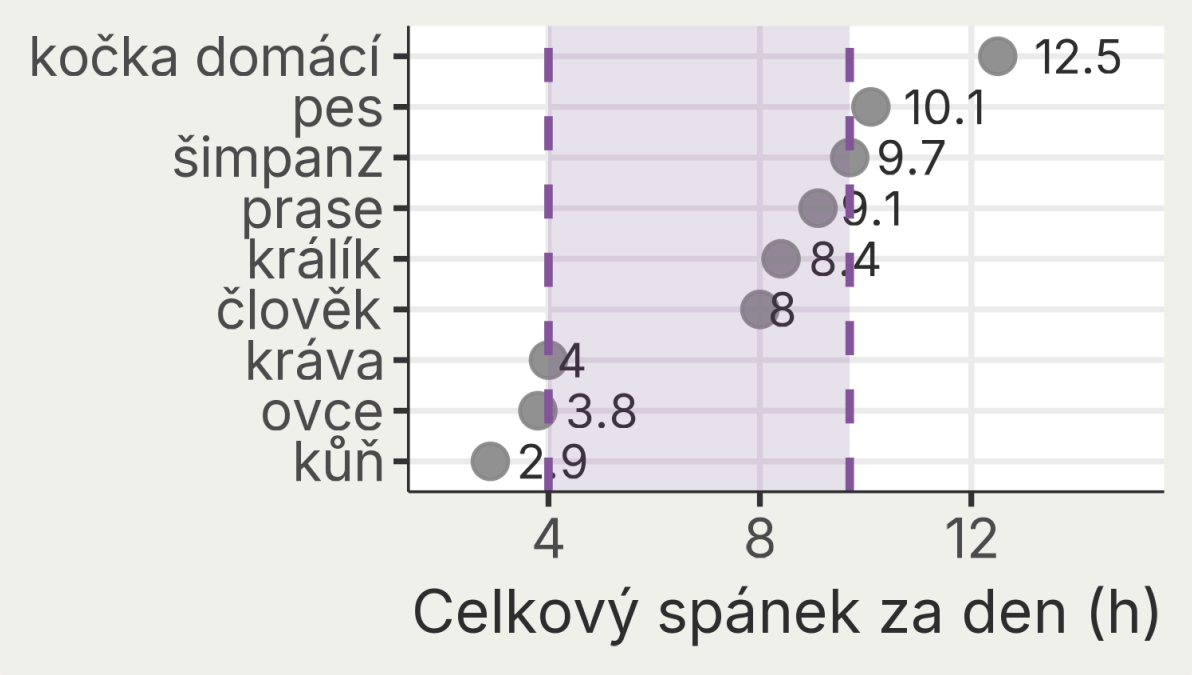

```r
quantile(
  x = vec_spanek_9,
  probs = c(0.25, 0.75)
)
```

```
25% 75%
4.0 9.7
```

Q1 odděluje dolní čtvrtinu, Q3 horní čtvrtinu hodnot.

Kvartily (*quartiles*) Q1 a Q3 určují hranice prostředních 50 % seřazených hodnot.

IQR je šířka prostřední poloviny

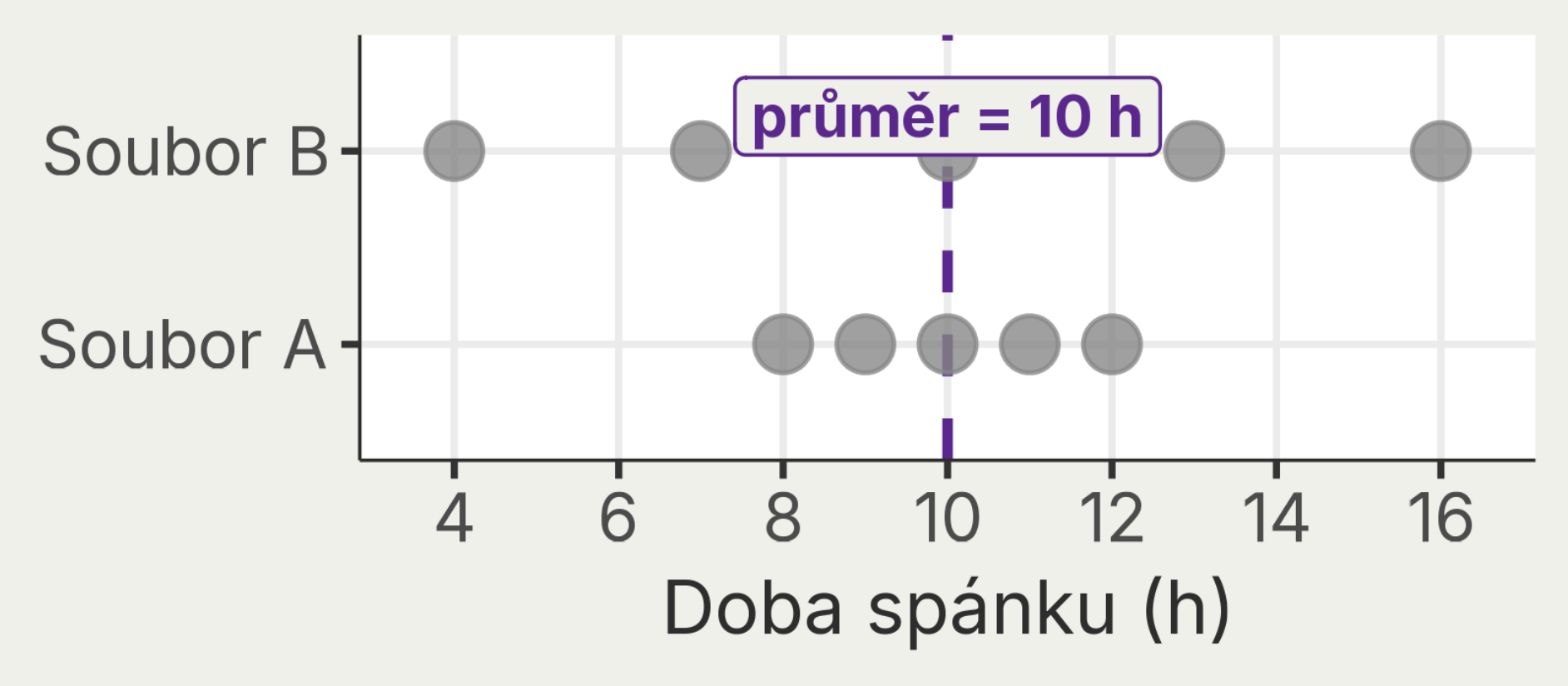



```r
IQR(
  x = vec_spanek_9
)
```

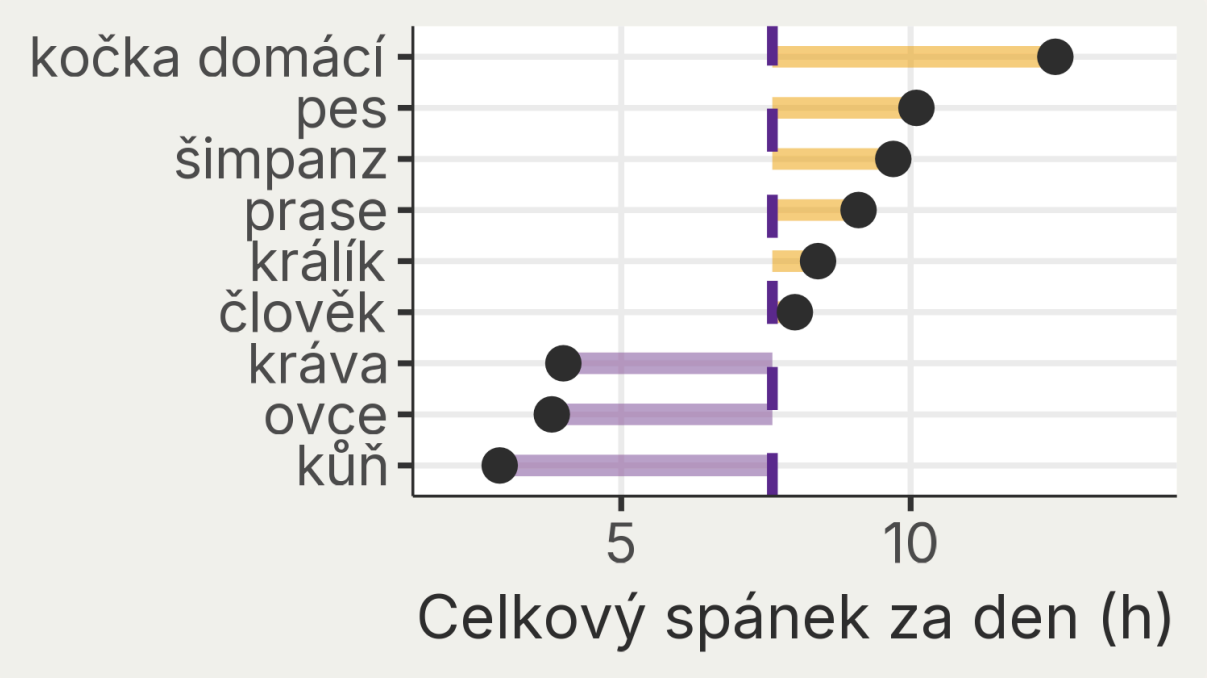
```
[1] 5.7
```

IQR (*interquartile range*) = Q3 − Q1. Popisuje rozptýlení prostřední poloviny hodnot.

Oba soubory mají průměr 10 hodin — jsou stejné?



Odchylka od průměru



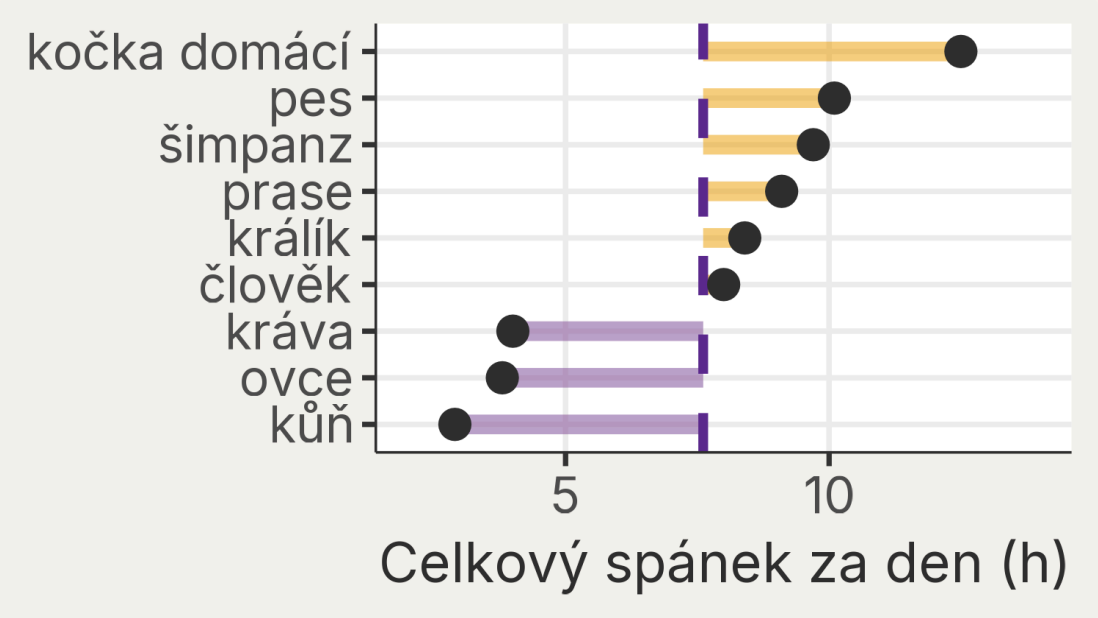
ODCHYLKA (*DEVIATION*)

odchylka doby spánku = doba spánku druhu – průměrná doba spánku

Fialová vede pod průměr, oranžová nad něj.

Délka barevného úseku ukazuje, jak daleko je každá hodnota od aritmetického průměru.

Odchylka u kočky



HODNOTY Z HORNÍHO ŘÁDKU

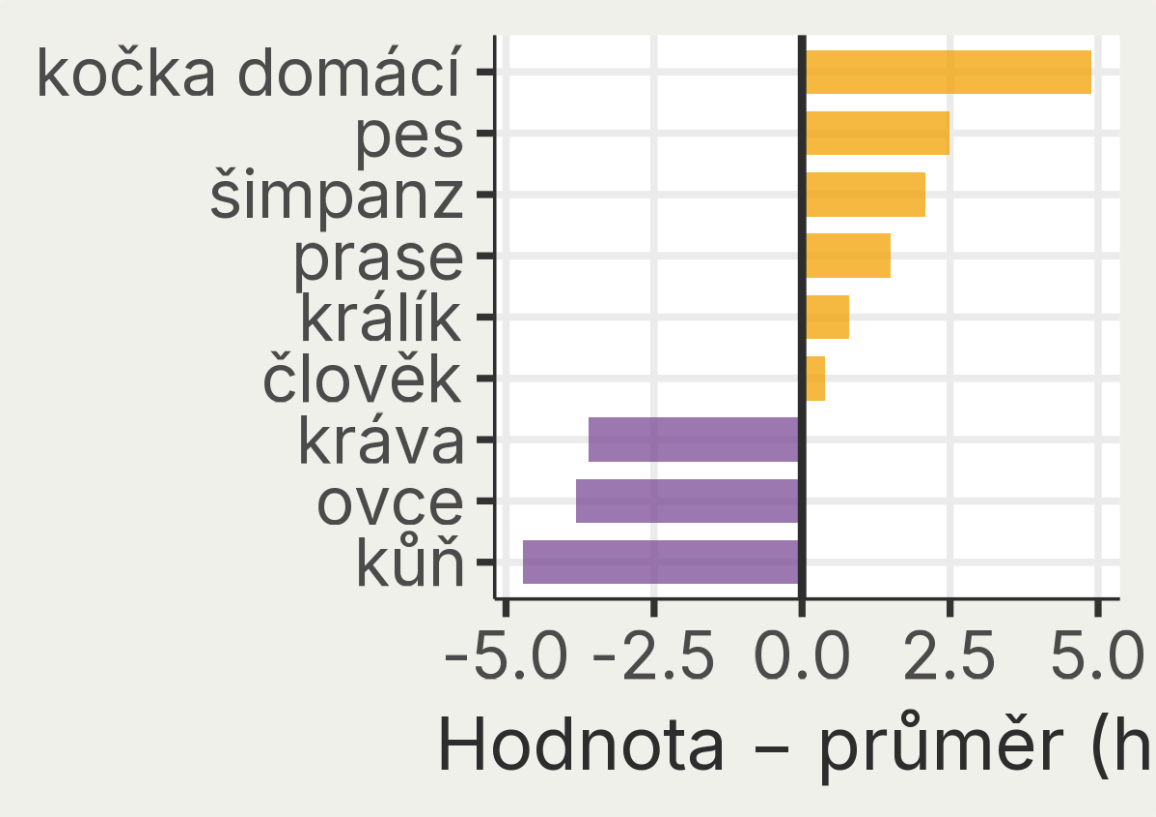
Kočka domácí: **12,5 h**
Aritmetický průměr: **7,61 h**

DOSAZENÍ

12,5 h – 7,61 h = 4,89 h

Kladný výsledek říká, že kočka domácí spí déle než průměr devíti druhů.

Odchylky se navzájem vyruší



SOUČET ODCHYLEK

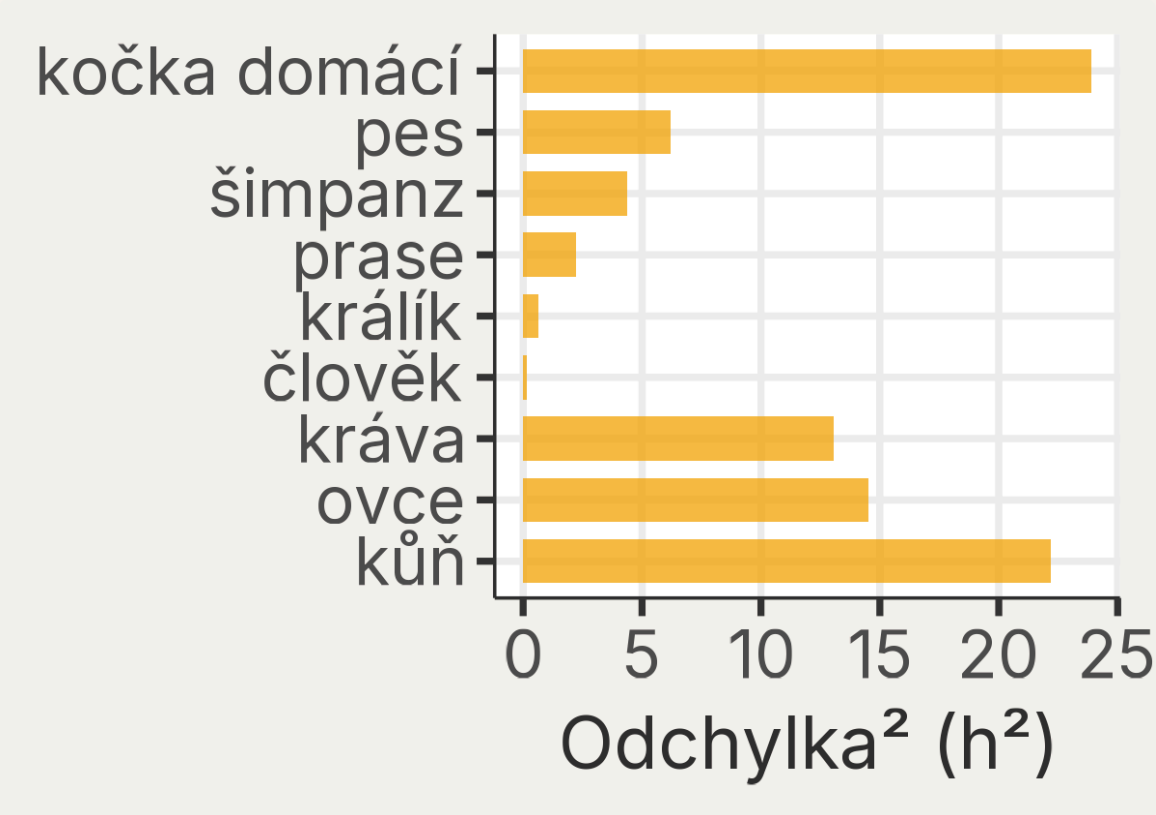
součet odchylek =
kladné odchylky +
záporné odchylky

= 0,00 h

Kolem průměru se navzájem vyváží.

Když prostý součet vždy vyjde přibližně nula, jak zachováme velikost vzdáleností?

Odchylky umocníme na druhou



ČTVEREC ODCHYLKY

čtverec odchylky = odchylka²

Všechny příspěvky jsou nezáporné.

$(12{,}5 - 7{,}61)^2 = 4{,}89^2 = 23{,}90\ h^2$

Nejvíce zde přispívá kočka domácí, protože leží nejdále od průměru.

Stejný krok rozepíšeme pro všech devět druhů

VŠECHNY HODNOTY JSOU PŘEVZATÉ Z GRAFU

$$(2{,}9 - 7{,}61)^2 + (3{,}8 - 7{,}61)^2 + (4{,}0 - 7{,}61)^2 +$$
$$(8{,}0 - 7{,}61)^2 + (8{,}4 - 7{,}61)^2 + (9{,}1 - 7{,}61)^2 +$$
$$(9{,}7 - 7{,}61)^2 + (10{,}1 - 7{,}61)^2 + (12{,}5 - 7{,}61)^2$$
$$= 87{,}21\ h^2$$

Součet devíti čtverců odchylek shrnuje jejich celkovou velikost do jednoho čísla.

Výběrový rozptyl

VÝPOČET SLOVY

výběrový rozptyl (*sample variance*) =
součet čtvercových odchylek dělený
počtem hodnot zmenšeným o jednu

DOSAZENÍ HODNOT

$$87{,}21\ \text{h}^2 / (9 - 1) =$$
$$87{,}21\ \text{h}^2 / 8 = 10{,}90\ \text{h}^2$$

Jednotkou rozptylu jsou **hodiny na druhou**. Proto se rozptyl hůře interpretuje přímo.

Vzorec shrnuje stejný výpočet

OBECNÝ VZOREC

$$s^2 = \Sigma(x_i - \bar{x})^2 / (n - 1)$$

```
var(
  x = vec_spanek_9
)
```

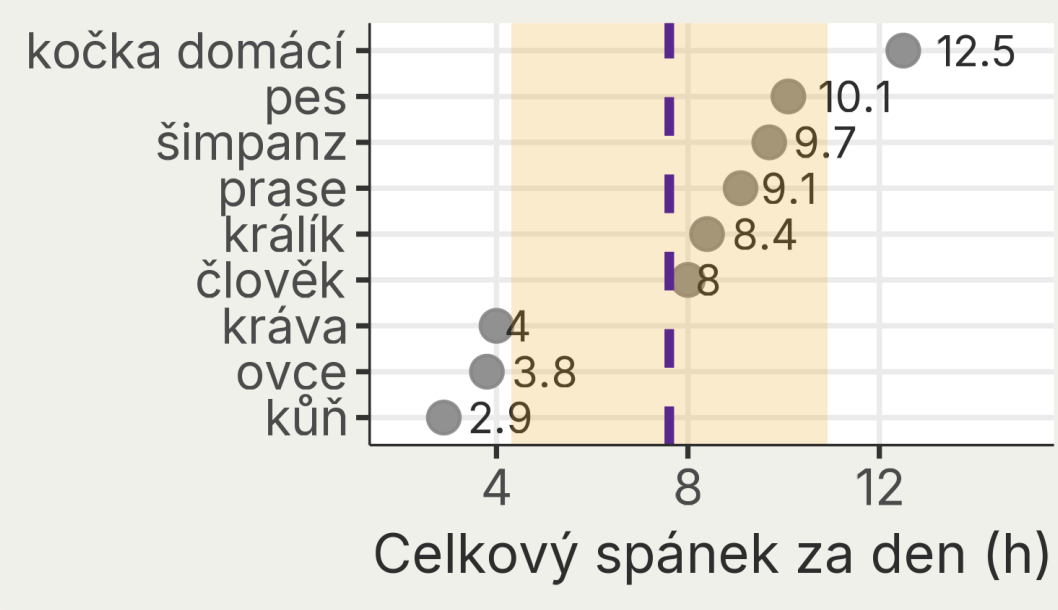
```
[1] 10.90111
```

CO ZNAMENAJÍ SYMBOLY?

- s^2 — výběrový rozptyl;
- x_i — doba spánku jednoho druhu;
- $\bar{x}$ — aritmetický průměr;
- n — počet hodnot;
- Σ — sečteme příspěvky všech druhů.

Vzorec zapisuje stejný výpočet pro libovolný počet hodnot.

Směrodatná odchylka



VZTAH K ROZPTYLU

směrodatná odchylka = druhá odmocnina z rozptylu

DOSAZENÍ HODNOT

$$\sqrt{10{,}90\ \text{h}^2} = 3{,}30\ \text{h}$$

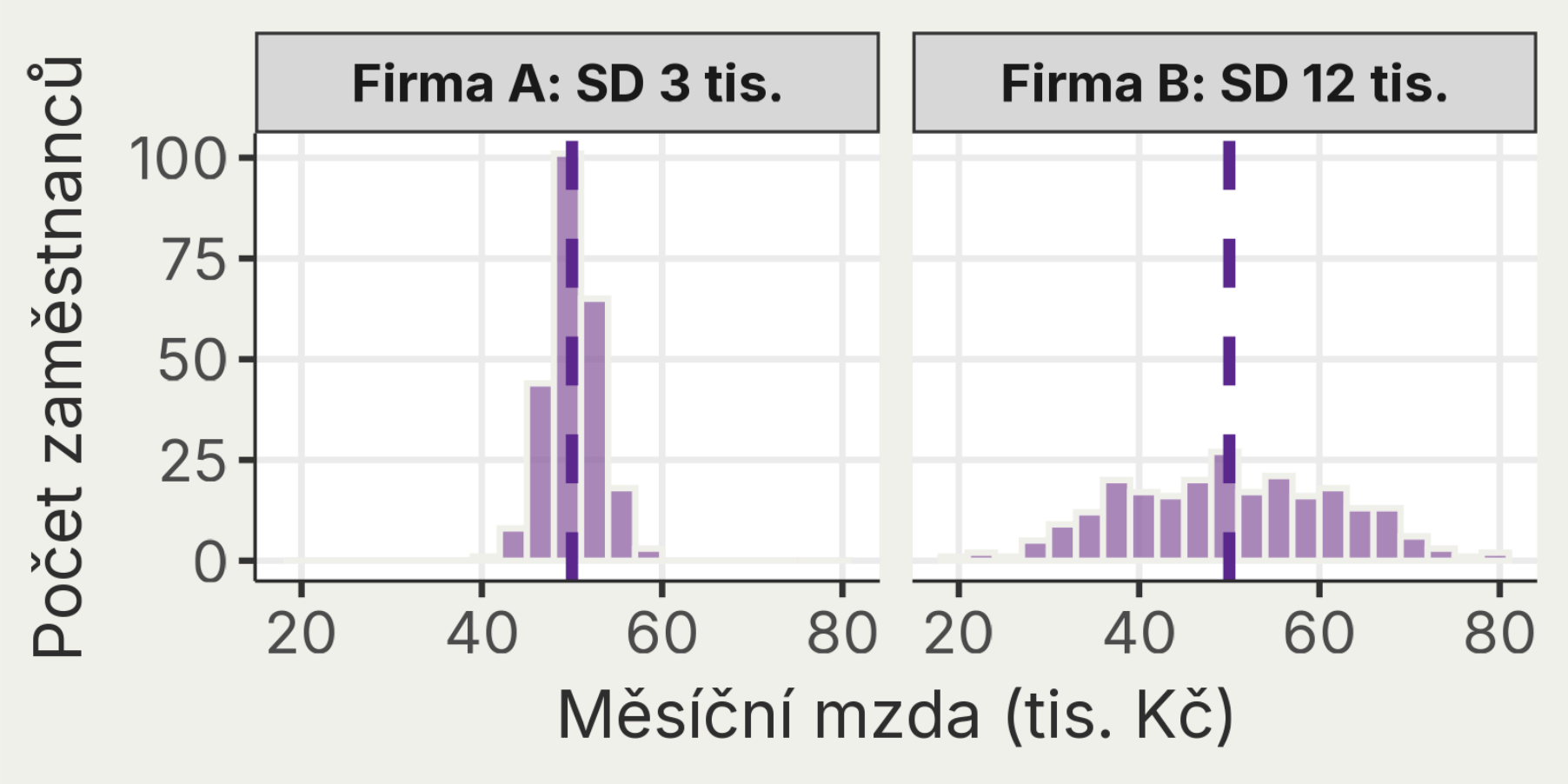
VZOREC

$$s = \sqrt{s^2}$$

s je směrodatná odchylka a **s²** výběrový rozptyl.

V R stejný krok provede `sd(x = vec_spanek_9)`. **SD (*standard deviation*)** je zpět v původních jednotkách.

Stejná průměrná mzda, jiné SD



V obou firmách je průměr 50 tisíc Kč. Vyšší SD ve firmě B (12 tisíc Kč oproti 3 tisícům Kč) znamená širší histogram.

Které souhrny patří k sobě?

Vyberte dvojici, která popisuje střed a rozptýlení podobným způsobem.

A	B	C
průměr + IQR	průměr + SD	medián + SD

Dvě užitečné dvojice

PRŮMĚR + SD

Používají všechny hodnoty a jsou citlivější k hodnotám daleko od středu.

MEDIÁN + IQR

Vycházejí z pořadí hodnot a jsou odolnější vůči vzdáleným hodnotám.

Volbu neurčuje zvyk. Rozhoduje tvar dat a otázka, kterou souhrnem zodpovídáme.

Zpět k titulku

Malý netopýr a velký slon jsou v jedné tabulce

Co znamená „běžná“ hmotnost?

BIOLOGICKÁ OTÁZKA

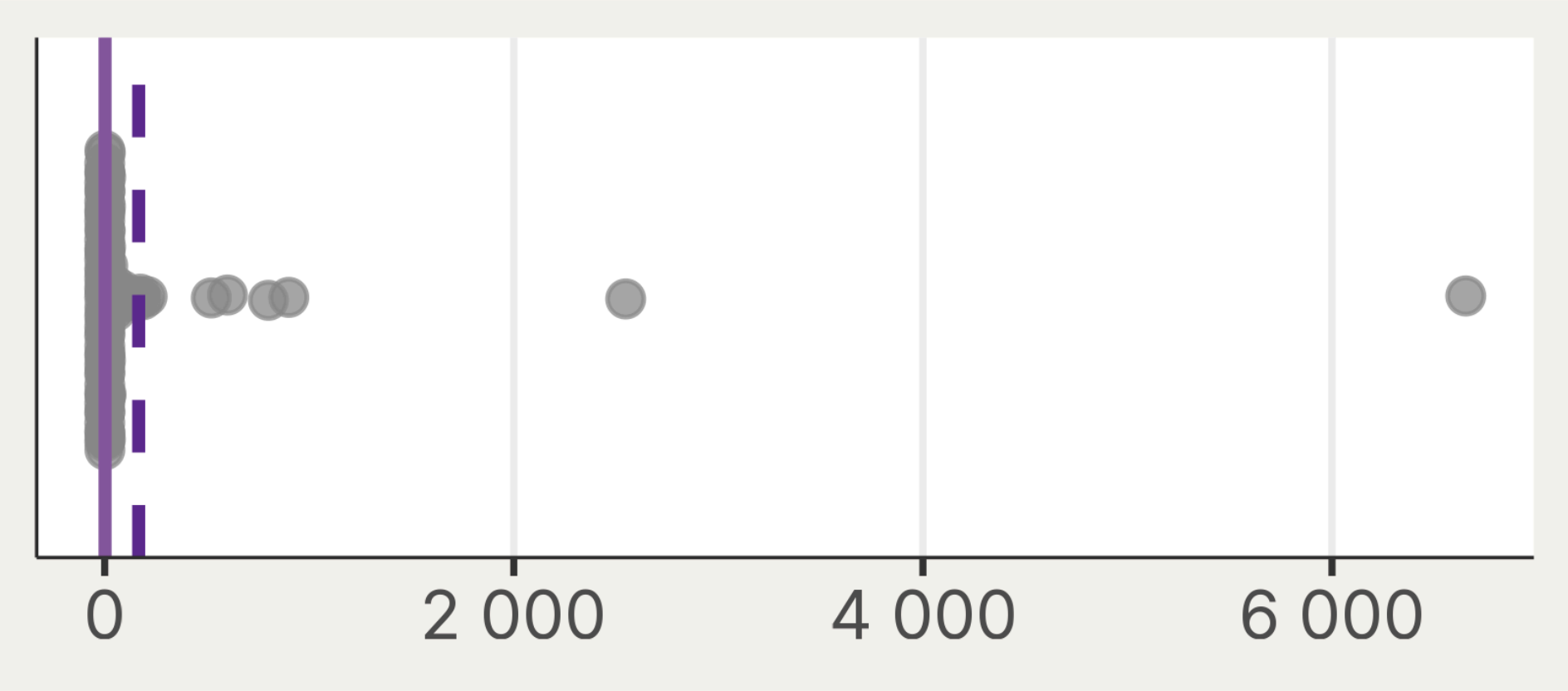
Má jedno číslo zastoupit druhy od velmi malých savců až po slony?

Volbu souhrnu opřeme o tvar dat v grafu.



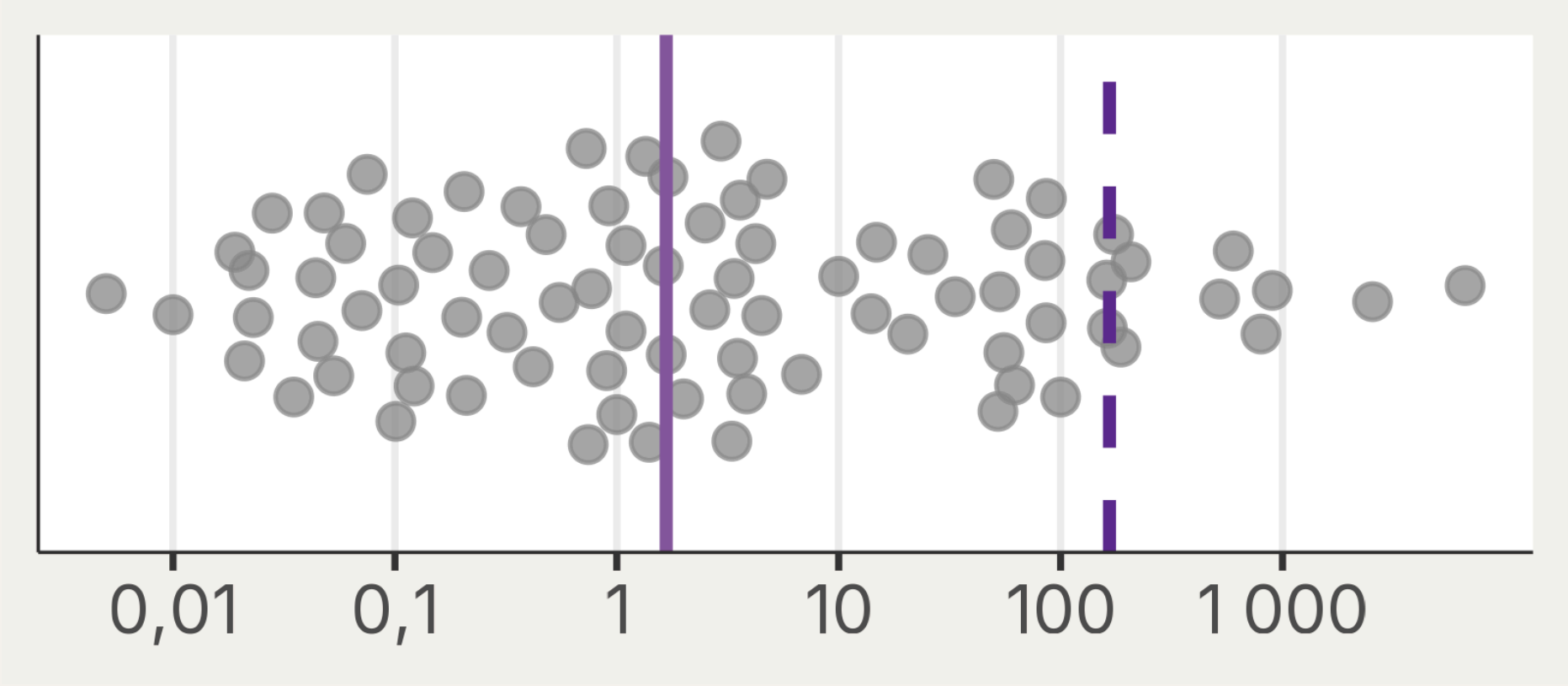
Ilustrace vytvořená pomocí AI.

Vracíme se k tělesné hmotnosti



Několik velmi těžkých druhů roztáhne osu tak silně, že většina bodů splývá u nuly.

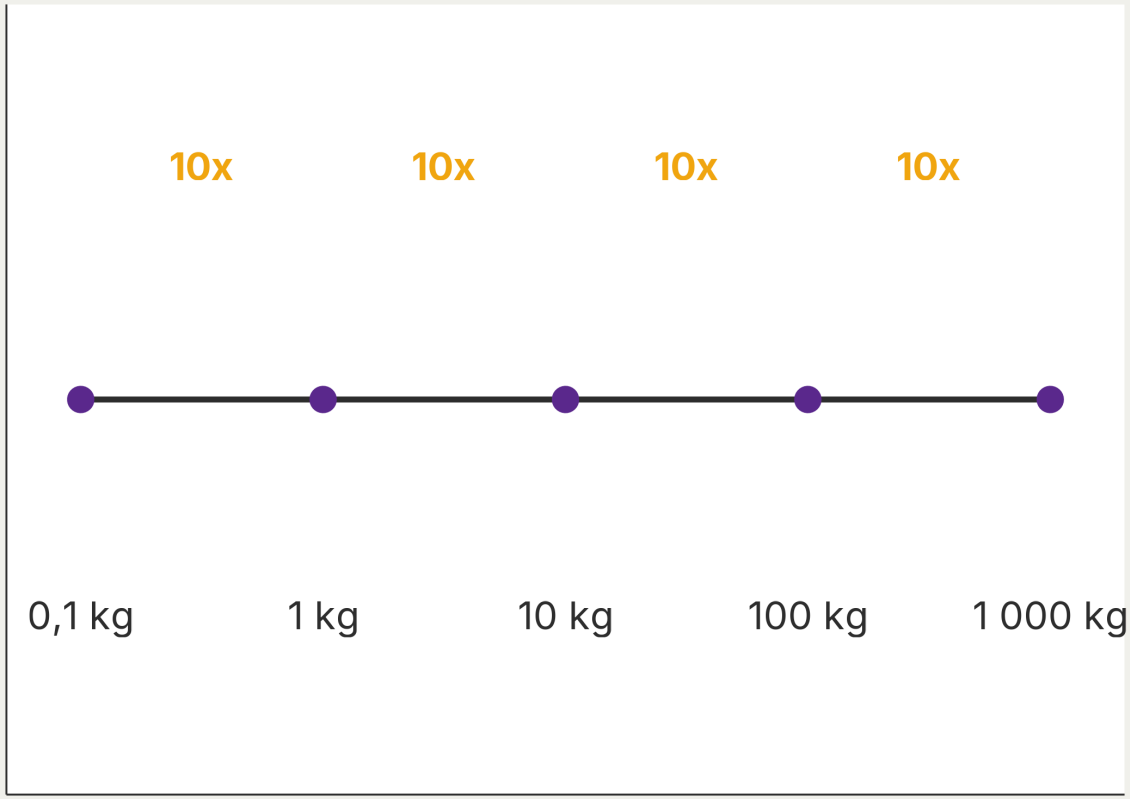
Stejná data na logaritmické ose



Graf nezměnil data. Jen nám dovolil vidět lehké i těžké druhy současně. Plná fialová čára je medián, přerušovaná tmavá čára průměr.

Co znamená logaritmická osa?

Stejná vzdálenost na ose znamená stejný násobek.



Úsek od 1 do 10 kg je stejně dlouhý jako úsek od 10 do 100 kg. Nulu ani zápornou hodnotu na logaritmické ose zobrazit nelze.

Jak dvě redakce počítaly?

TITULEK A · PRŮMĚR

„Průměrný savec váží **166 kg**.“

TITULEK B · MEDIÁN

„Běžný savec váží jen **1,67 kg**.“

A) Pouze A.

B) Pouze B.

C) Oba mohou být správně.

D) Nemáme dost informací.

Obě čísla jsou správně vypočtená. Ani jedno však není univerzální „typický savec“.

Statistická redakce: opravte titulek

DŮKAZY

- řádek představuje zastoupený **druh**
- rozdělení hmotnosti je ovlivněno několika velmi těžkými druhy;
- průměr = 166 kg, medián = 1,67 kg.

Přepište jeden titulek tak, aby byl pravdivý i po přečtení grafu.

Co z tabulky smíme tvrdit?

POCTIVÝ ZÁVĚR

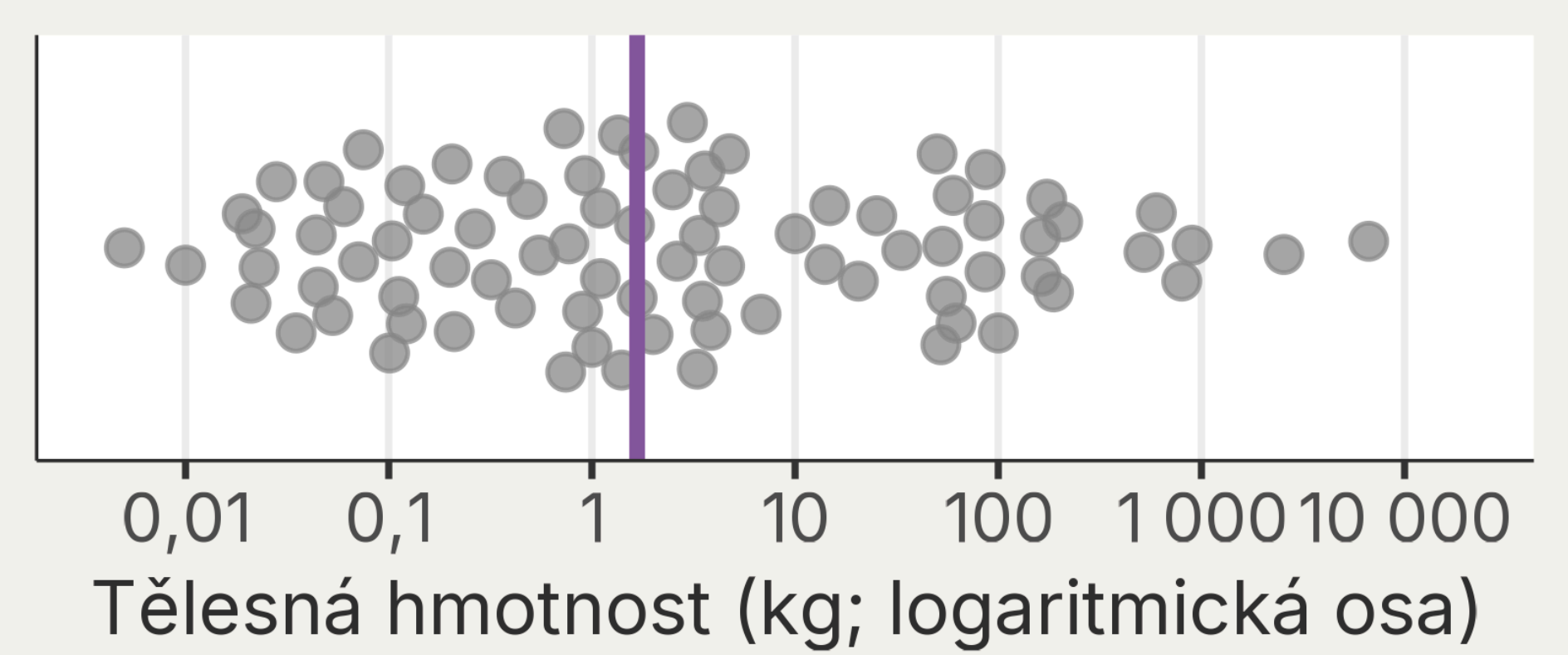
V této tabulce je medián hmotnosti zastoupeného druhu **1,67 kg**, ale několik velmi těžkých druhů posouvá průměr na **166 kg**.

POZOR!

Tabulka není náhodným výběrem všech druhů ani všech jednotlivých savců. Čísla proto nepopisují „náhodného savce na Zemi“.

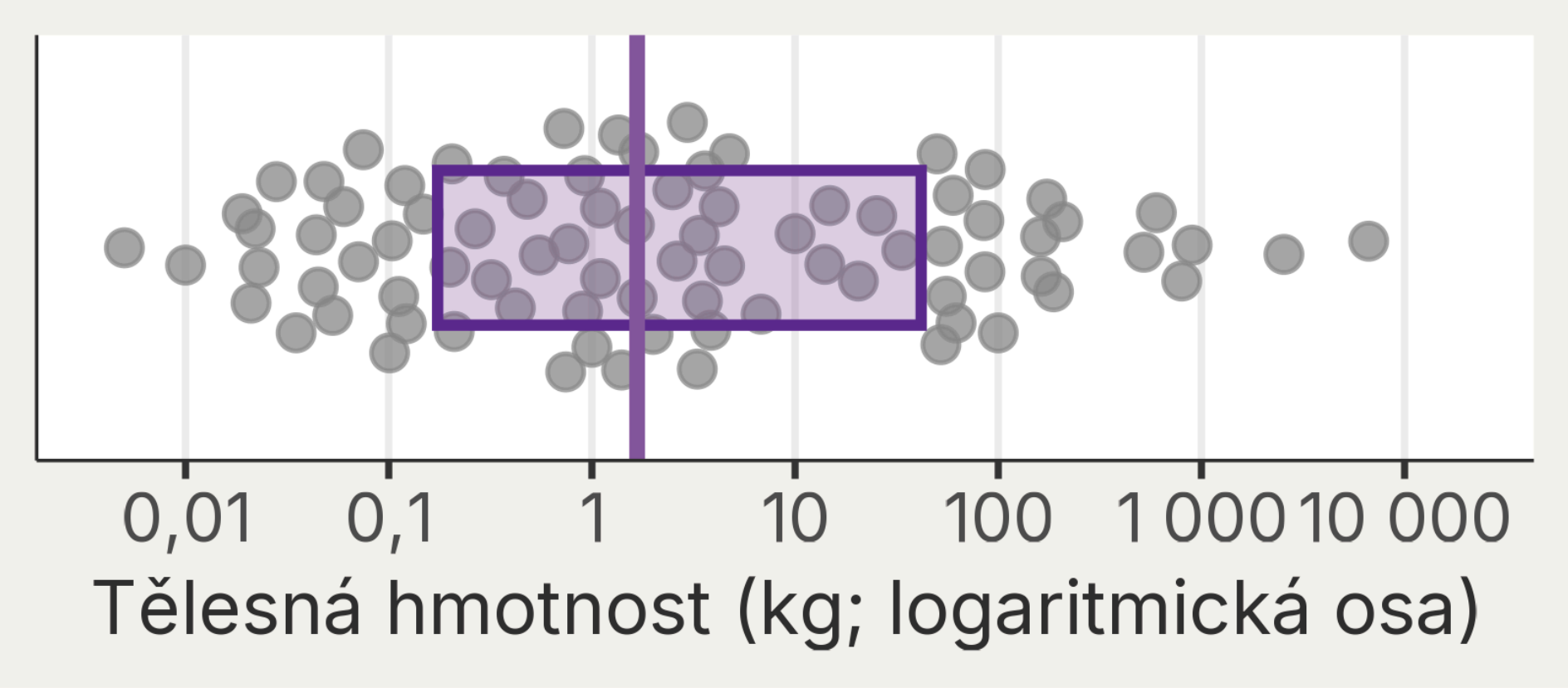
Statistiky jako vizualizace

Správný titulek můžeme doplnit grafem



„Běžný zastoupený druh váží **1,67 kg**.“ Graf ukazuje data i medián za tímto titulkem.

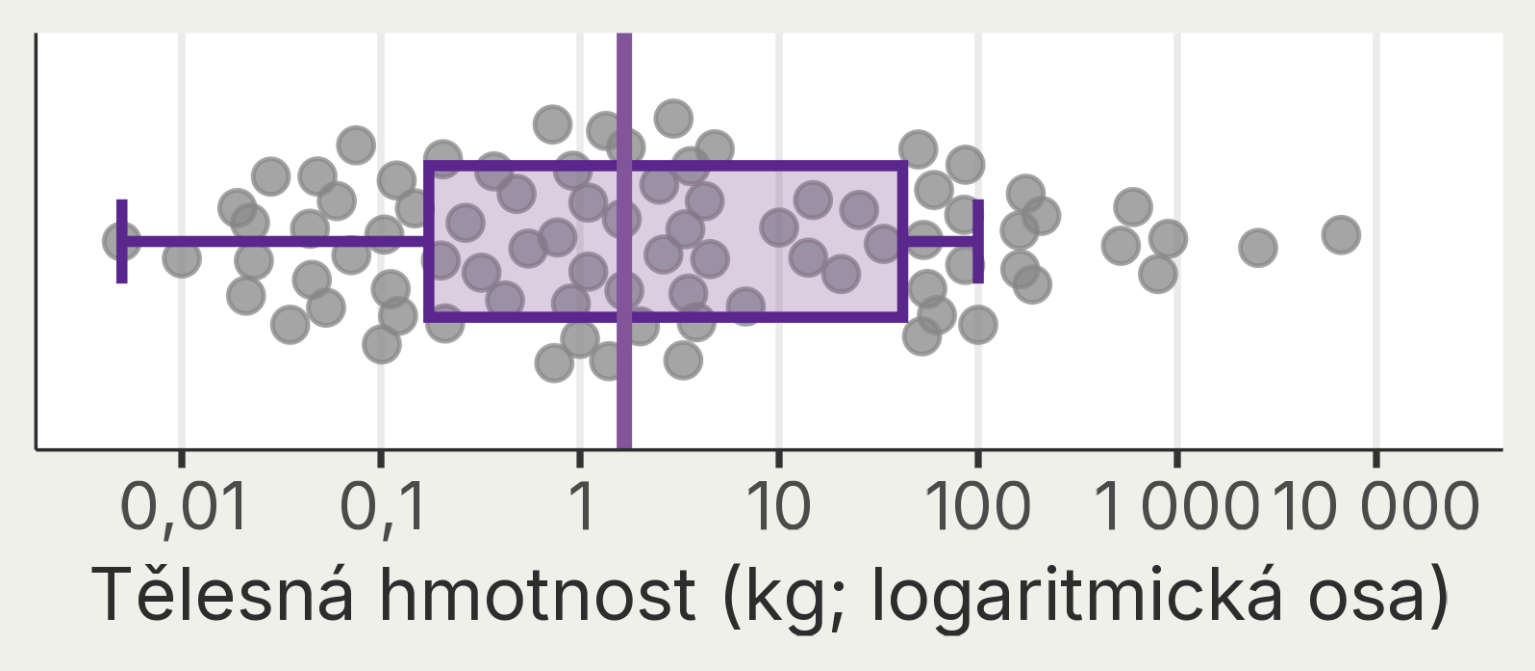
Přidáme prostřední polovinu



Šířka krabice je **IQR**: prostředních 50 % hodnot.

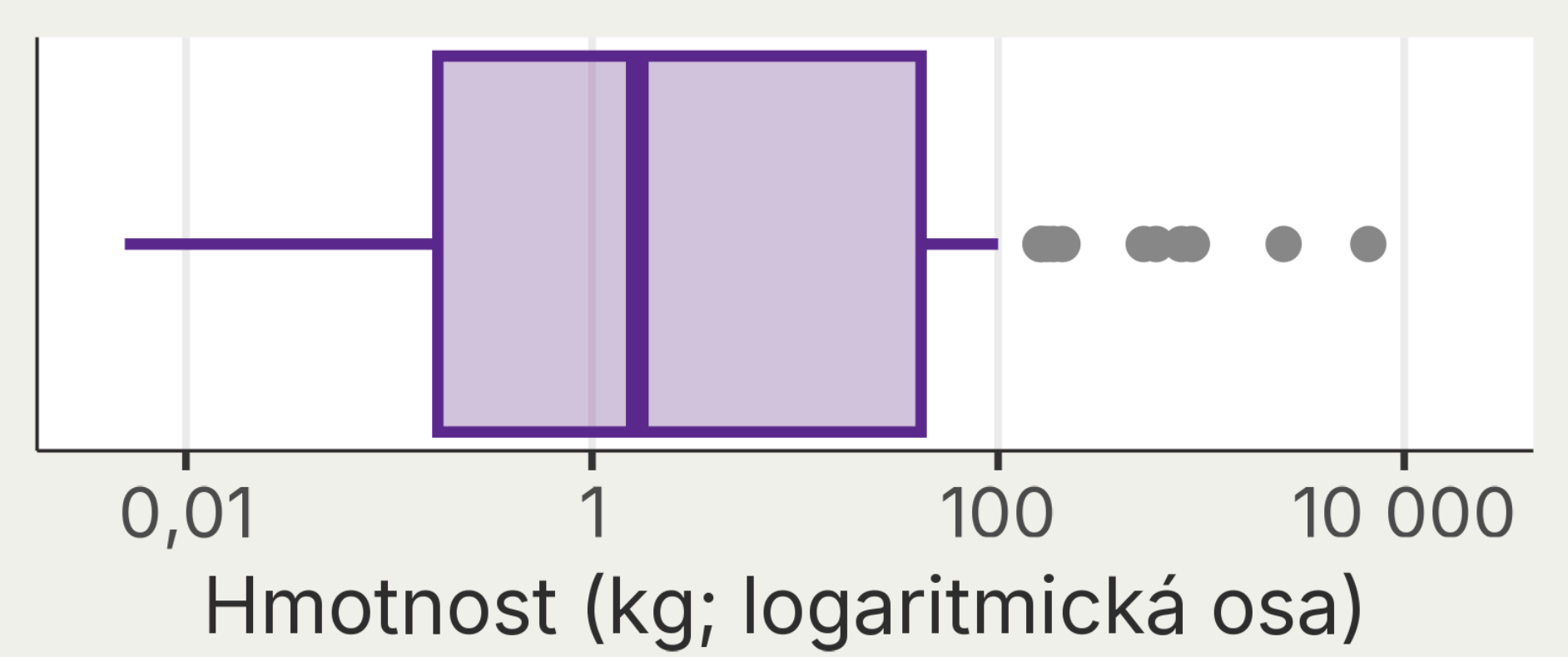
Přidáme vousy

Vous končí u nejkrajnější pozorované hodnoty nejvýše 1,5 × IQR od okraje krabice. Body dál od krabice vyznačíme zvlášť; nejsou automaticky chybou. Pravidlo počítáme v původních kilogramech.



Body za vousy mohou být zajímavé nebo neobvyklé, ale nejsou automaticky chybné.

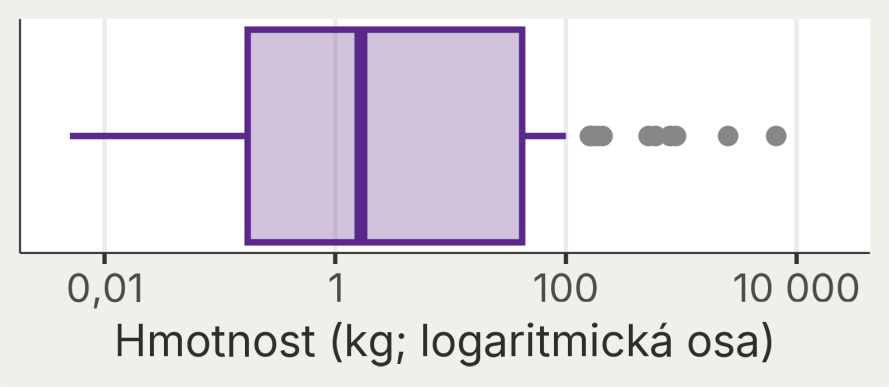
Z bodů vznikl krabicový graf



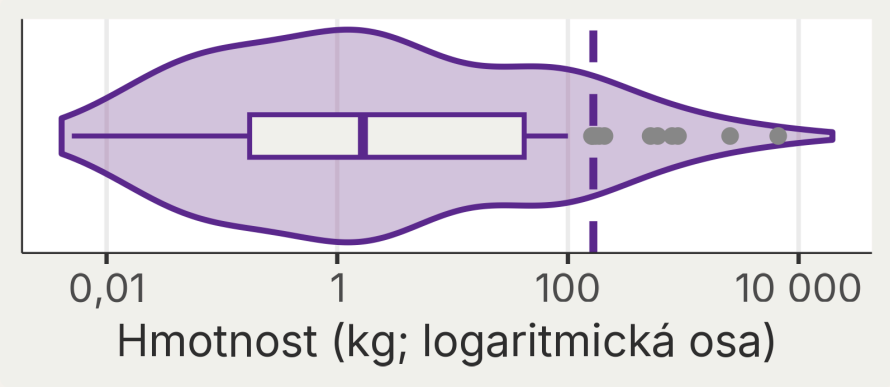
Krabicový graf (*boxplot*) je mapa několika souhrnů, ne fotografie všech hodnot.

Co krabice schovala?

Krabicový graf



Krabicový + houslový graf



přerušovaná tmavá čára = průměr

plná čára v krabici = medián

Houslový graf vrací tvar

KRABICOVÝ GRAF

Rychle porovná medián, IQR, vousy a možné odlehlé hodnoty.

HOUSLOVÝ GRAF

Ukáže, kde se hodnoty hromadí, ale přesná četnost se z jeho šířky čte obtížně.

Housle mohou být informativnější, když je důležitý `tvar`; `body` zůstávají nejtransparentnější první kontrolou.

Typy proměnných

Jeden netopýr, různé typy údajů

U netopýra můžeme zaznamenat:

- dobu spánku v hodinách;
- váhu v kilogramech;
- počet pozorovaných jedinců;
- druh;
- typ potravy;
- velikostní kategorii.



Ilustrace vytvořená pomocí AI.

> Typ neurčuje vzhled čísla. Určuje jej význam proměnné.

Číslo 37 může znamenat čtyři různé věci

Zápis	Co znamená	Jak vznikl
37,0	tělesná teplota	měření
37	počet mláďat	počítání
37	číselný kód lokality	název kategorie
3	stupeň poškození	uspořádaná kategorie

Rozhoduje vzhled hodnoty, nebo její význam?

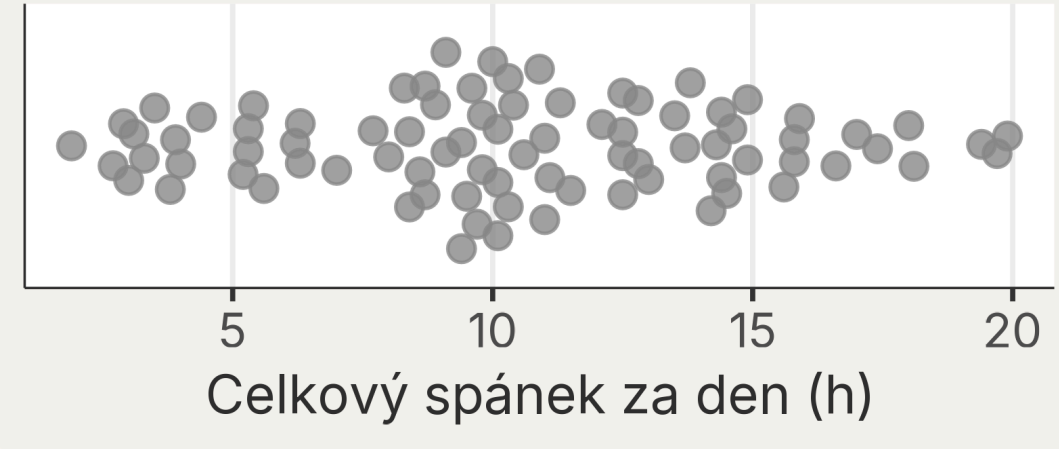
Typ určuje význam a možné operace

- **spojitá numerická**: měření na číselné škále;
- **diskrétní numerická**: výsledek počítání;
- **nominální kategoriální**: názvy bez přirozeného pořadí;
- **ordinální kategoriální**: kategorie s přirozeným pořadím.

> Technická třída objektu v R pomáhá s kontrolou, ale statistický typ určuje význam hodnot a způsob jejich vzniku.

Doba spánku vznikla měřením

Spojitá numerická proměnná: mezi dvěma hodnotami mohou být další **měření**.



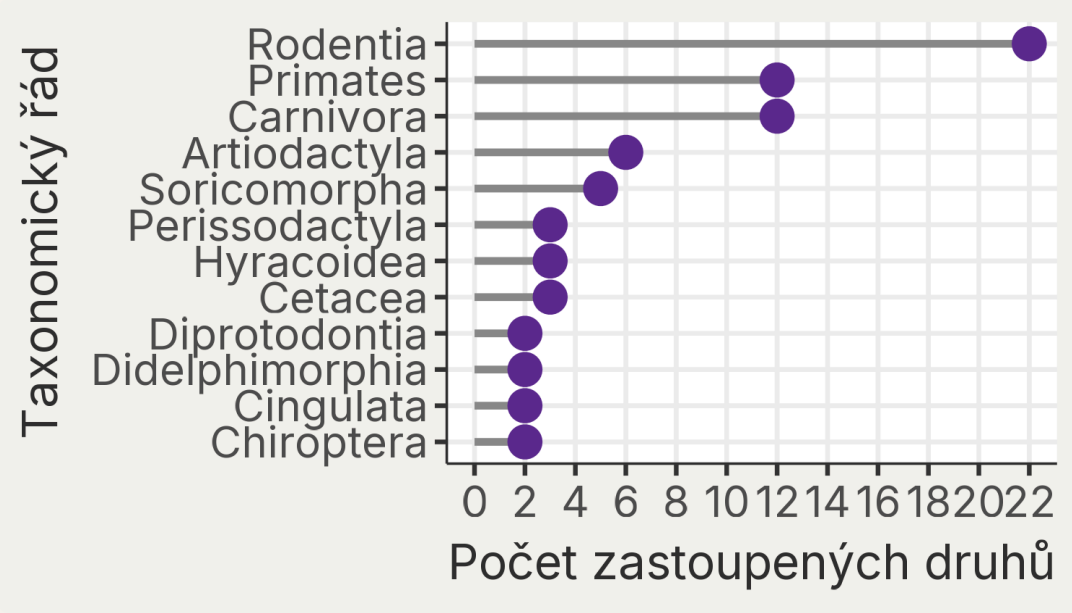

```
1  vec_spanek_h <-
2    ggplot2::msleep[["sleep_total"]]
3
4  class(x = vec_spanek_h)
```

```
[1] "numeric"
```

```
1  typeof(x = vec_spanek_h)
```

```
[1] "double"
```

Počet druhů v řádu jsme spočítali



Diskrétní numerická proměnná: vznikla počítáním; může nabývat pouze celočíselných hodnot (např. 1, 5, 10).

```
vec_pocet_druhu <-
  as.integer(
    x = table(
      ggplot2::msleep[["order"]]
    )
  )

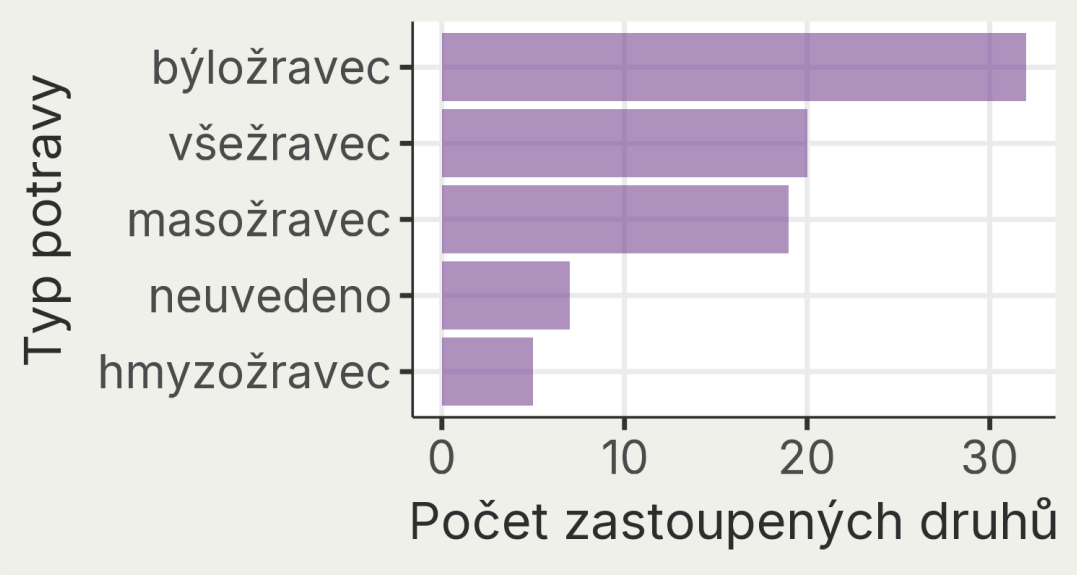
class(x = vec_pocet_druhu)
```

```
[1] "integer"
```

```
str(object = vec_pocet_druhu)
```

```
int [1:19] 1 6 12 3 2 2 2 2 2 3 ...
```

Typ potravy je název bez pořadí



Nominální kategoriální proměnná: smysl dávají četnosti a podíly, ne průměr.

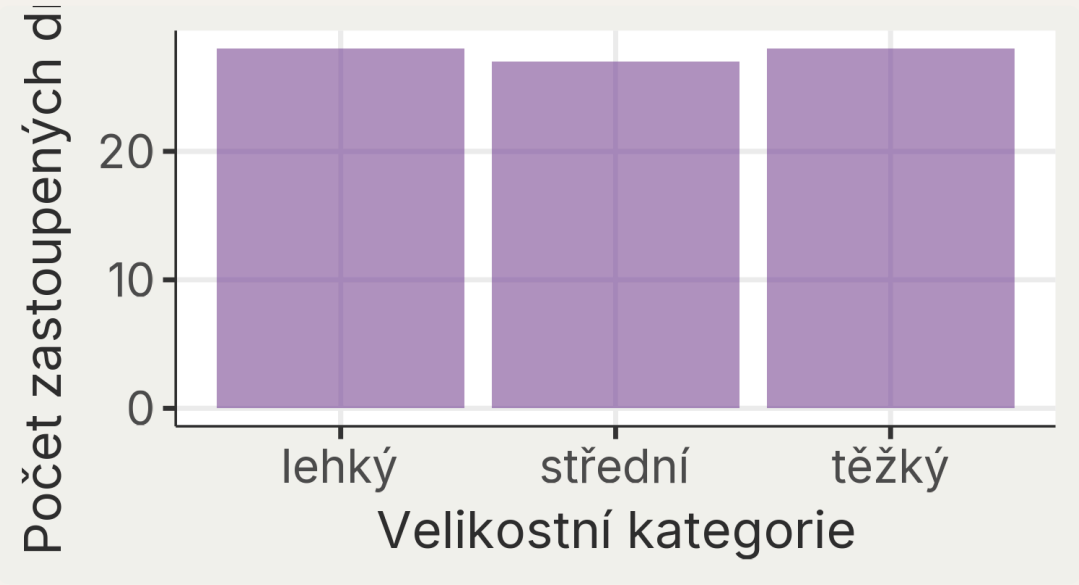
```
 1  vec_typ_potravy <-
 2    ggplot2::msleep[["vore"]]
 3
 4  vec_chybi <-
 5    is.na(vec_typ_potravy)
 6
 7  vec_typ_potravy[vec_chybi] <-
 8    "neuvedeno"
 9
10  vec_typ_potravy <-
11    factor(x = vec_typ_potravy)
12
13  class(x = vec_typ_potravy)
```

```
[1] "factor"
```

```
 1  str(object = vec_typ_potravy)
```

```
 Factor w/ 5 levels "carni","herbi",..: 1
5 2 5 2 2 1 4 1 2 ...
```

Velikostní kategorie: lehký–střední–těžký



Ordinální kategoriální proměnná: pořadí je součástí objektu: lehký < střední < těžký.

```
poradi <-
  c("lehký", "střední", "těžký")

vec_velikost <-
  ordered(
    x = poradi,
    levels = poradi
  )

class(x = vec_velikost)
```

```
[1] "ordered" "factor"
```

```
str(object = vec_velikost)
```

```
 Ord.factor w/ 3 levels "lehký"<"střední"
<..: 1 2 3
```

Určete typ proměnné

MOŽNOSTI

spojitá numerická · diskrétní numerická · nominální kategoriální · ordinální kategoriální

Proměnná	Jak vznikla?
pH vody v rybníce	měřením
počet mláďat ve vrhu	počítáním
druh stanoviště	názvem kategorie
stupeň poškození listu	uspořádanou kategorií

Přiřaďte ke každé proměnné jeden ze čtyř typů.

Od typu proměnné k prvnímu popisu

Typ proměnné	Vhodný první souhrn	Vhodný první graf
numerická spojitá	medián + IQR nebo průměr + SD	body, histogram, hustota, boxplot
numerická diskrétní	četnosti; někdy medián/průměr	body nebo sloupce počtů
kategoriální nominální	četnosti a podíly	sloupcový graf
kategoriální ordinální	četnosti; někdy medián pořadí	uspořádaný sloupcový graf

> Typ proměnné zužuje, které grafy, souhrny a později i modely dávají smysl.

Závěr

Shrnutí

- Každý řádek spojíme s jeho skutečnou jednotkou pozorování.
- Data zhodnotíme graficky před výběrem souhrnného čísla.
- Průměr + SD; medián + IQR.
- Význam proměnné určuje její typ. Podle toho vybereme vhodný graf a vhodný souhrn.

Co potřebujete vědět, než řeknete „typický“?

Jaké tři otázky položíte člověku, který vám ukáže jedno souhrnné číslo bez grafu?

> V následující lekci přidáme druhou proměnnou a zeptáme se, zda spolu dvě měření souvisejí.