

L02

Mají tučnější s delší ploutví také vyšší tělesnou hmotnost?



Vztahy dvou proměnných a vizualizace

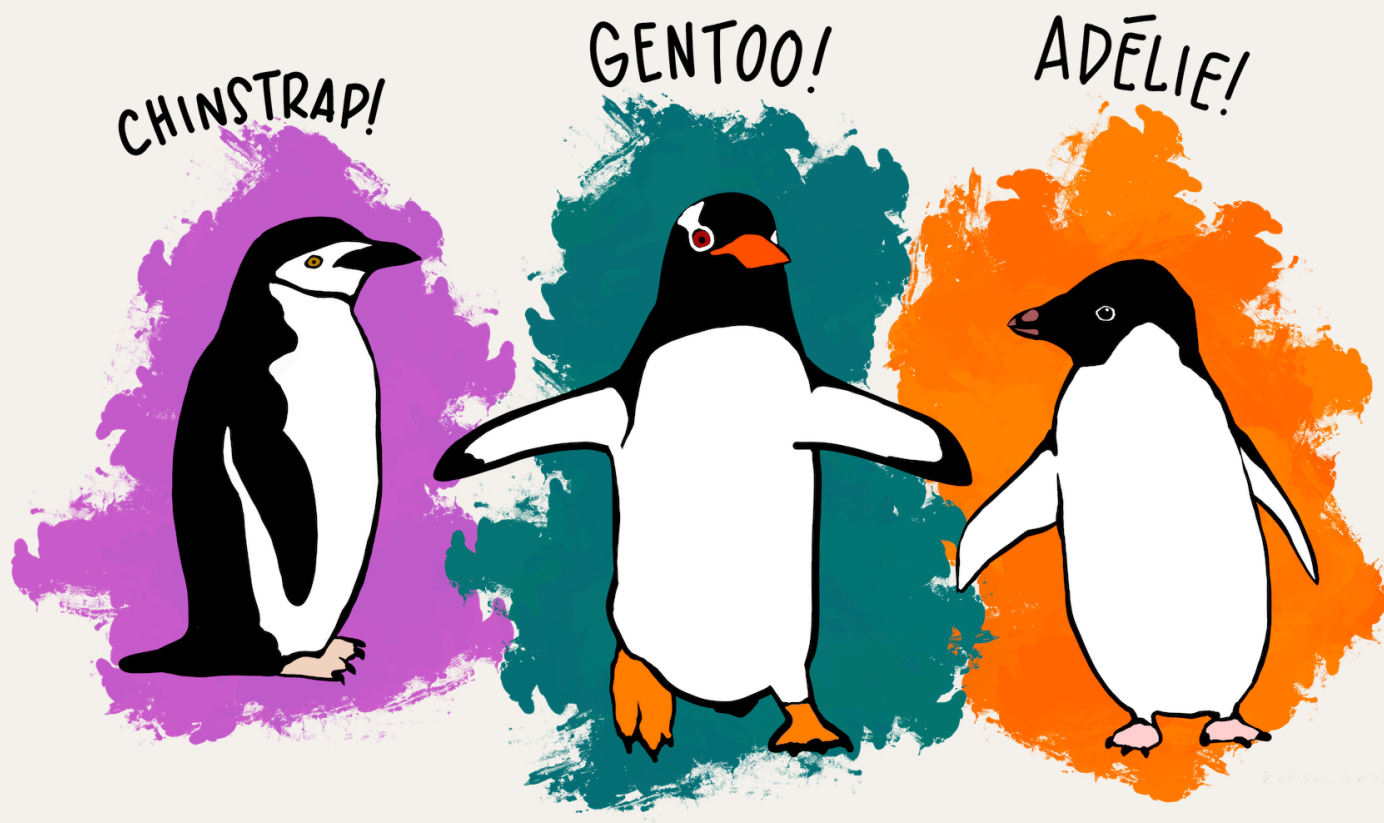
Ilustrace vytvořená pomocí AI.

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

Pomůže nám druhá proměnná?

NAŠE BIOLOGICKÁ OTÁZKA

Mají tučňáci s delší ploutví také vyšší tělesnou hmotnost?



Ilustrace: Allison Horst

Co si pamatujete z minulé lekce?

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

Který z uvedených druhů spí podle tabulky nejdéle?

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

Cheetah

Owl monkey

Mountain beaver

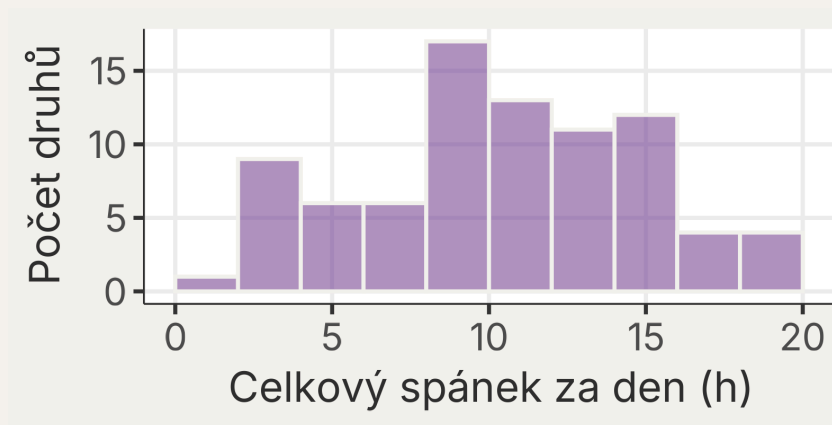
Greater short-tailed shrew

SPRÁVNÁ ODPOVĚĎ

: Owl monkey

Porovnáme hodnoty ve sloupci Spánek za den (h). Nejvyšší hodnota je 17.0 h v řádku Owl monkey.

Ve kterém intervalu délky spánku je podle histogramu nejvíce zastoupených druhů?



10–12 h

12–14 h

8–10 h

16–18 h

SPRÁVNÁ ODPOVĚĎ

: 8–10 h

Nejvyšší sloupec stojí nad intervalem 8–10 h. Poloha sloupce na vodorovné ose určuje interval spánku; jeho výška vyjadřuje počet druhů v tomto intervalu.

Jaká je podle výstupu průměrná délka spánku devíti druhů?

```
vec_spanek_deviti <- c(
  2.9, 3.8, 4.0, 8.0, 8.4,
  9.1, 9.7, 10.1, 12.5
)

mean(x = vec_spanek_deviti)
# [1] 7.611111

median(x = vec_spanek_deviti)
# [1] 8.4
```

SPRÁVNÁ ODPOVĚĎ

: 7.611111 h

Průměr vypočítává funkce `mean()`, která vrácí 7.611111 h. Funkce `median()` vrácí 8.4 h; hodnoty 4.0 h a 9.1 h jsou jednotlivá pozorování ve vektoru.

Výsledky učení

Po absolvování této přednášky budete schopni:

- přečíst **bodový graf** dvou numerických proměnných: směr, tvar a rozptýlení bodů
- vysvětlit, proč jsou biologická data **rozptýlená** i při zřetelném vztahu a co k popisu přidá **korelace**
- zvolit vhodný první graf podle typu dvojice proměnných
- rozlišit popis asociace od tvrzení o její příčině

Dvě měření, jeden bod

Od jedné proměnné ke dvěma

V MINULÉ LEKCI JSME ČETLI JEDNU PROMĚNNOU

Z tabulky, histogramu nebo číselného souhrnu jsme zjišťovali, jak vypadají hodnoty jedné proměnné.

DVOJICE MUSÍ ZŮSTAT SPOLU

Dva oddělené histogramy neukážou, které dvě hodnoty patří stejnému tučňákovi.

Dvě hodnoty patří jednomu tučňákovi



Ilustrační obraz vytvořen AI.

Dvojice měření musí zůstat spojena se stejným pozorováním.

Diskuze se sousedem

NAŠE BIOLOGICKÁ OTÁZKA

Mají tučňáci s delší ploutví také vyšší tělesnou hmotnost?

Jak by podle vás mohl vypadat graf pro zodpovězení této otázky?

30 s přemýšlení — pak sdílejte svůj odhad se sousedem.

Tabulka tučňáků

Druh	Délka ploutve (mm)	Hmotnost (g)
Adelie	181	3750
Adelie	186	3800
Adelie	195	3250

JEDEN ŘÁDEK = JEDEN TUČŇÁK

Které dvě hodnoty z jednoho řádku vytvoří jeden bod grafu?

Úplné dvojice v R

```
1 data_puvodni <-  
2   read.csv(  
3     file = "data/palmer_penguins.csv",  
4     na.strings = ""  
5   )  
6 vec_uplne <-  
7   complete.cases(  
8     data_puvodni$flipper_length_mm,  
9     data_puvodni$body_mass_g  
10  )  
11 data_pary <- data_puvodni[vec_uplne, ]  
12 data_tucnaci <-  
13   data.frame(  
14     druh_tucnaka = data_pary$species,  
15     ostrov = data_pary$island,  
16     delka_ploutve_mm = data_pary$flipper_length_mm,  
17     hmotnost_g = data_pary$body_mass_g  
18   )
```

344 řádků

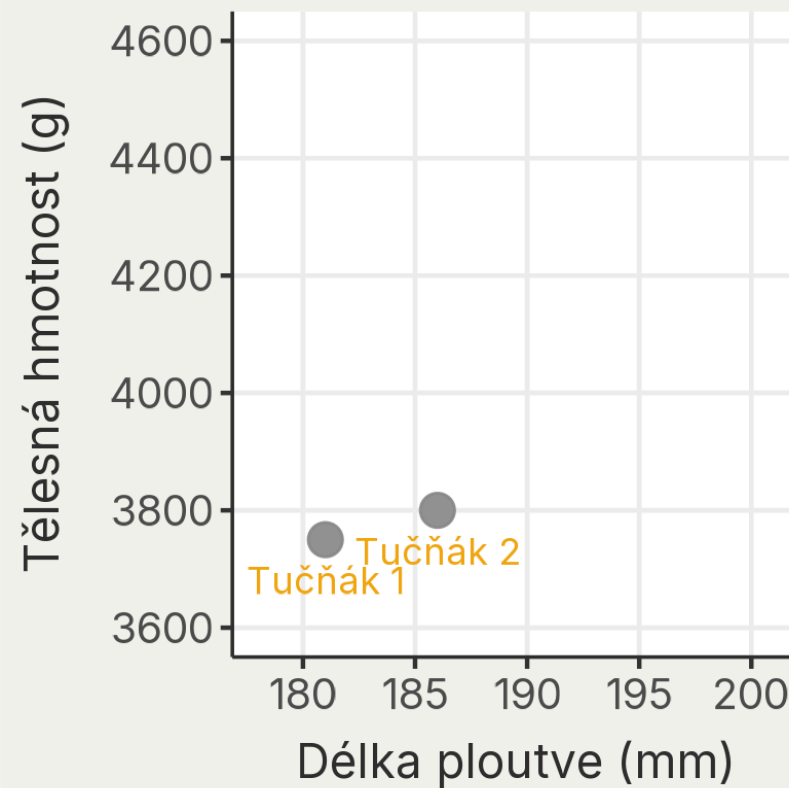
2 řádky bez obou měření

342 úplných dvojic

Neúplný řádek vynecháme celý, aby se dvojice měření nerozpojila.

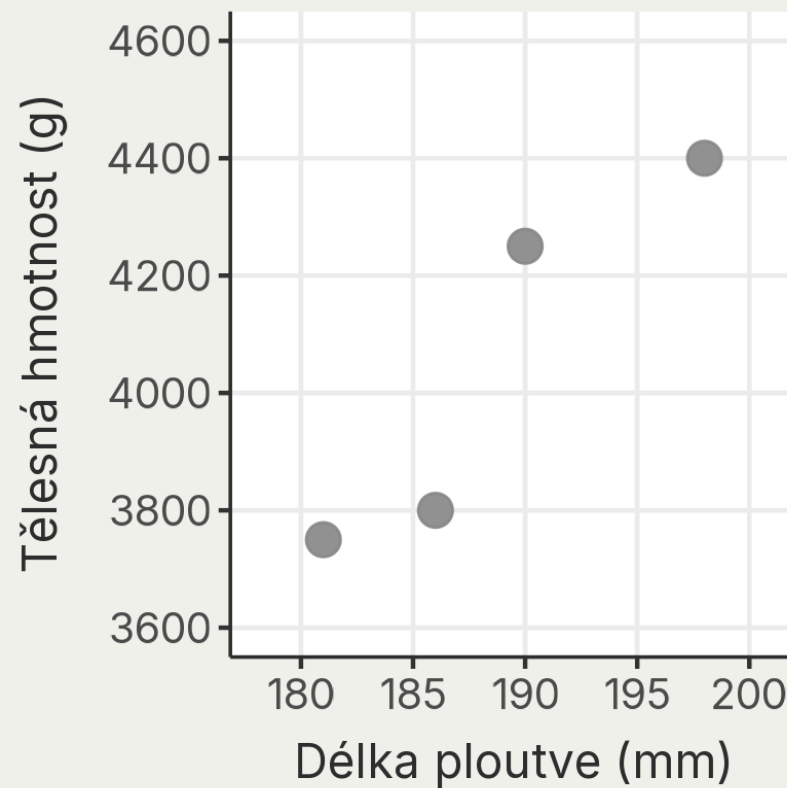
Dva tučňáci

```
1 vec_delka_ploutve_mm <-  
2   c(181, 186)  
3  
4 vec_hmotnost_g <-  
5   c(3750, 3800)  
6  
7 plot(  
8   x = vec_delka_ploutve_mm,  
9   y = vec_hmotnost_g,  
10  xlab = "Délka ploutve (mm)",  
11  ylab = "Tělesná hmotnost (g)",  
12  pch = 16  
13 )
```



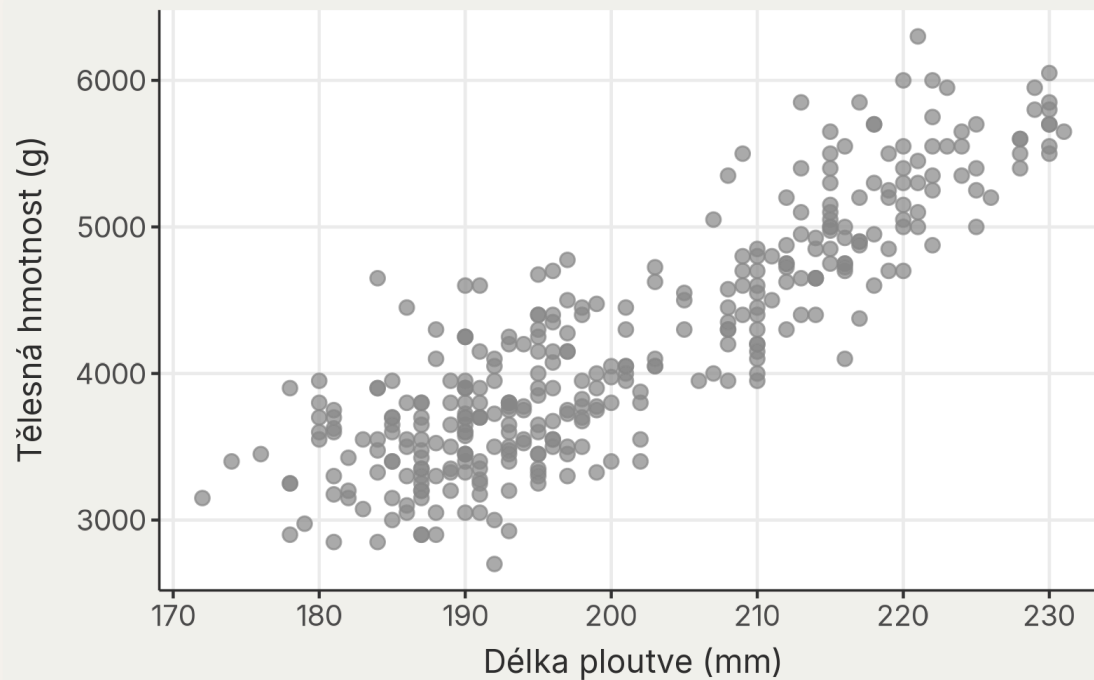
Čtyři tučňáci

```
1 vec_delka_ploutve_mm <-  
2   c(181, 186, 190, 198)  
3  
4 vec_hmotnost_g <-  
5   c(3750, 3800, 4250, 4400)  
6  
7 plot(  
8   x = vec_delka_ploutve_mm,  
9   y = vec_hmotnost_g,  
10  xlab = "Délka ploutve (mm)",  
11  ylab = "Tělesná hmotnost (g)",  
12  pch = 16  
13 )
```

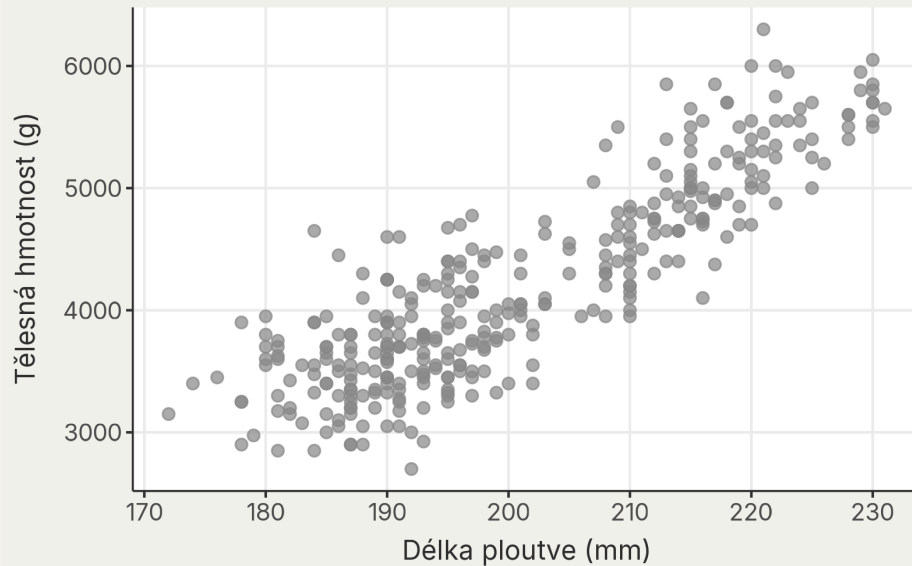


Všichni tučňáci

```
1 plot(  
2   x = data_tucnaci$delka_ploutve_mm,  
3   y = data_tucnaci$hmotnost_g,  
4   xlab = "Délka ploutve (mm)",  
5   ylab = "Tělesná hmotnost (g)",  
6   pch = 16  
7 )
```

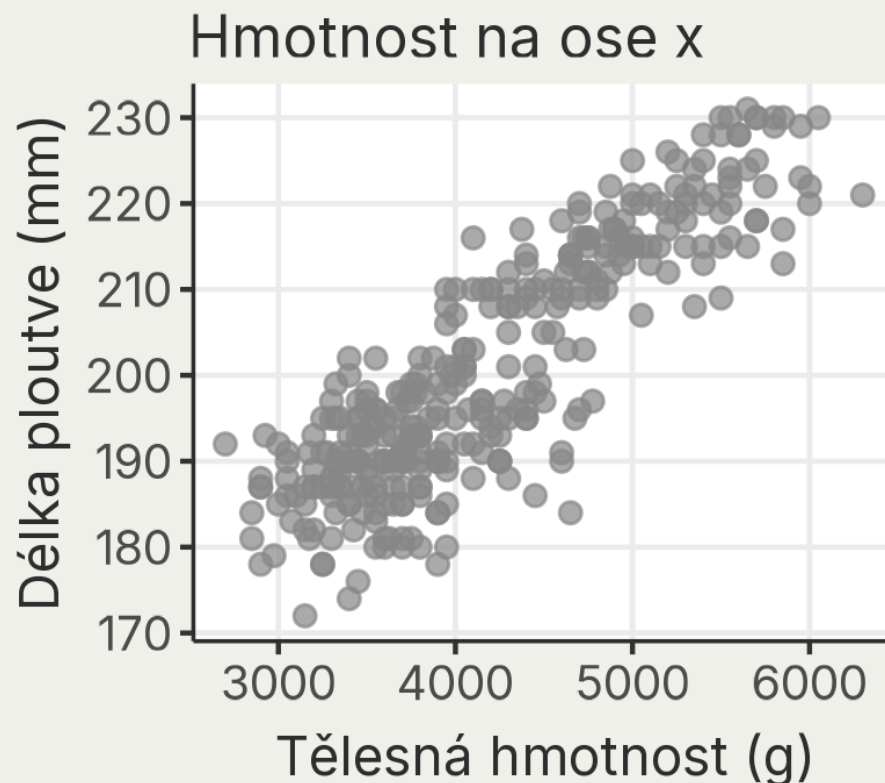
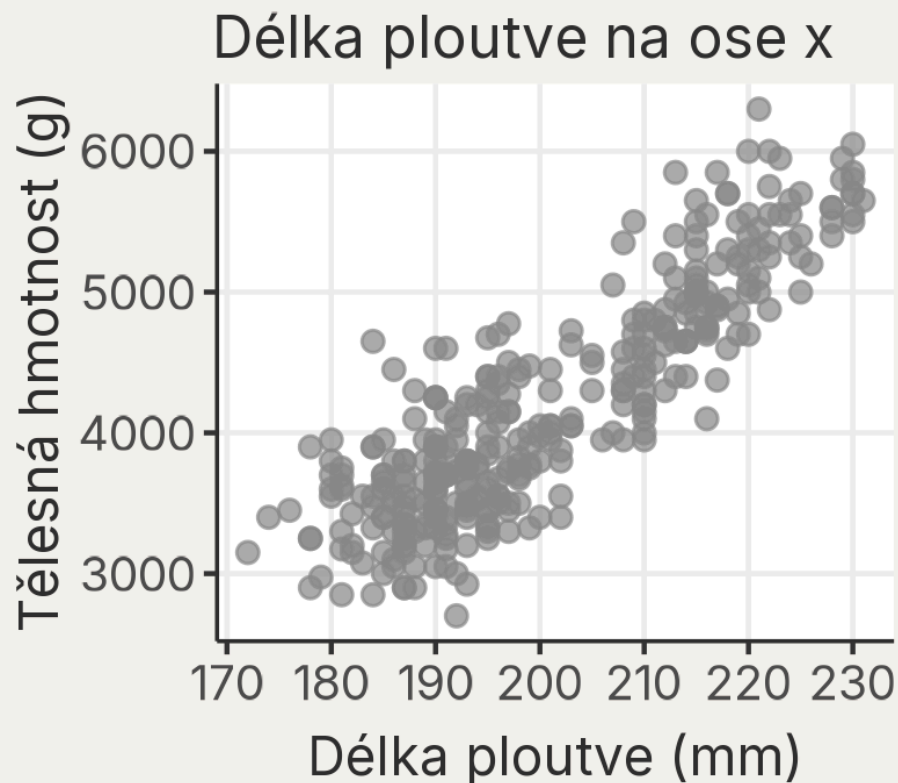


Co v tomto grafu vidíte?



- **Směr**: rostoucí
- **Rozptýlení**: oblak zůstává široký
- **Závěr**: vzor vidíme, příčinu z grafu neurčíme

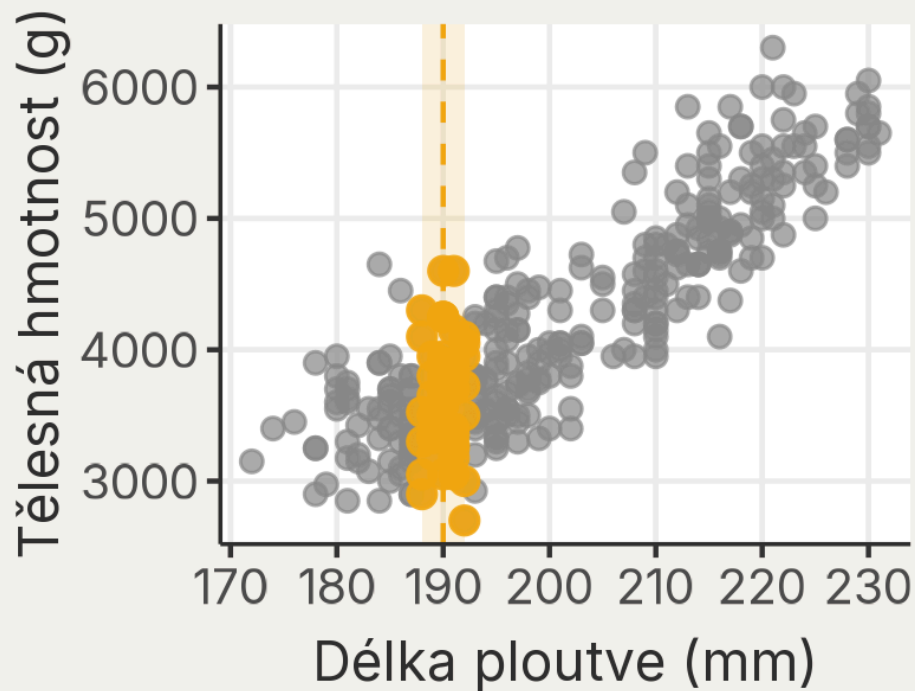
Stejná data, jiné osy



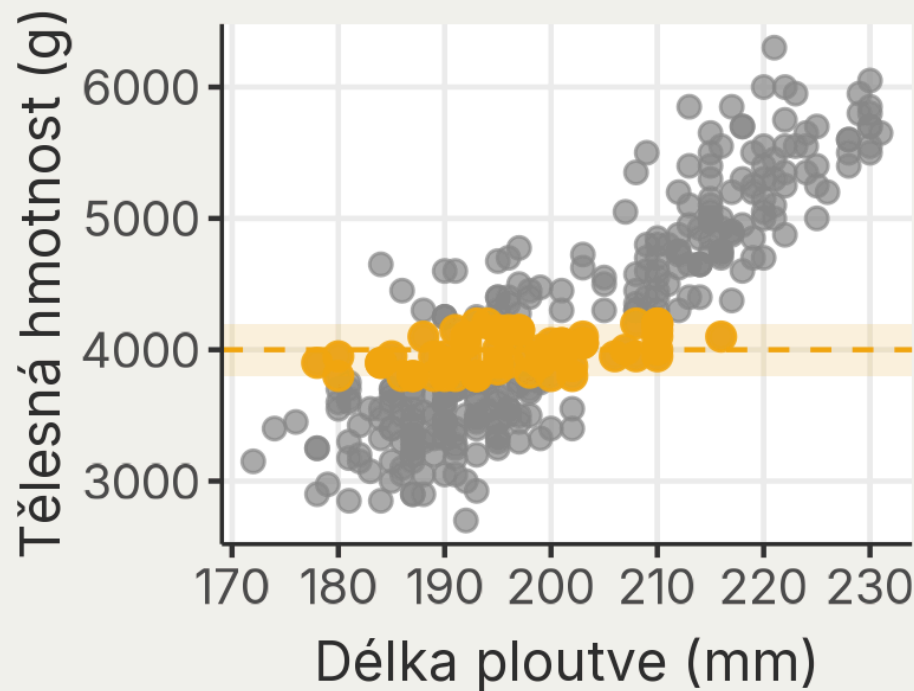
Změnil se vztah mezi délkou ploutve a hmotností, když jsme osy prohodili?

Oblak se jen zrcadlově převrátí podle úhlopříčky. Směr ani těsnost vztahu se nezmění; tytéž dvojice měření zůstávají spojené.

Rozptýlení čteme ve dvou směrech



Podobná délka ploutve, různá
hmotnost



Podobná hmotnost, různá **délka**
ploutve

Když čteme bodový graf, díváme se současně na **směr vztahu** i na to, jak široký oblak zůstává v obou směrech.

Co už můžeme tvrdit? Která formulace je v této chvíli nejlepší?

A

V grafu je vidět vztah, takže jsme prokázali příčinu.

B

V datech je vidět biologicky smysluplný vzor, který chceme popsat přesněji.

C

Protože body neleží na jedné přímce, mezi proměnnými žádný vztah není.

Než hlasujeme: přesvědčte souseda, proč jste vybrali svoji možnost.

Mají tučňáci s delší ploutví také vyšší tělesnou hmotnost?

Mají tučňáci s delší ploutví také vyšší tělesnou hmotnost?

Signál

Delší ploutve **bývají spojeny** s vyšší hmotností.

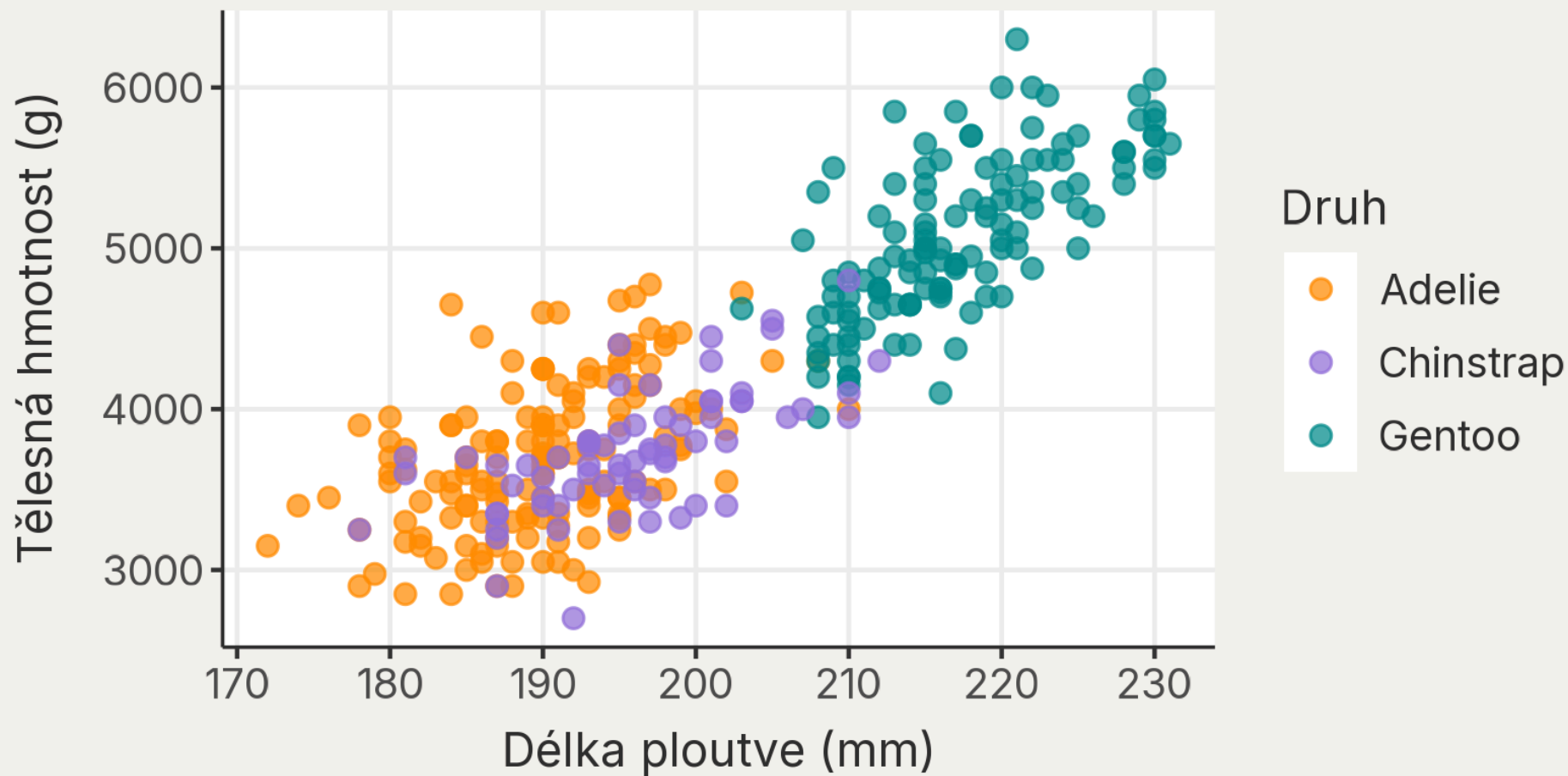
Rozptýlení

Vztah **není** dokonale těsný.

POZOR

Graf popisuje asociaci; její příčinu ani nejistotu z tohoto obrázku neurčíme.

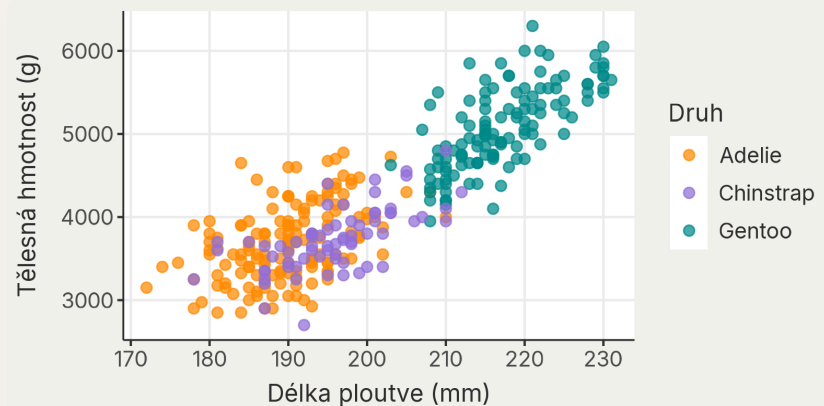
Stejná data, ale s druhem



Co se po obarvení změnilo?

BIOLOGICKÁ LOGIKA PLATÍ I PO OBARVENÍ PODLE DRUHU

- **Směr** oblaku zůstává kladný
 - *tučňáci s delší ploutví bývají těžší*
- **Rozptýlení** má biologickou strukturu
 - část variability pochází z míchání více druhů



Jedno číslo pro společný pohyb

Co by takové číslo muselo umět?

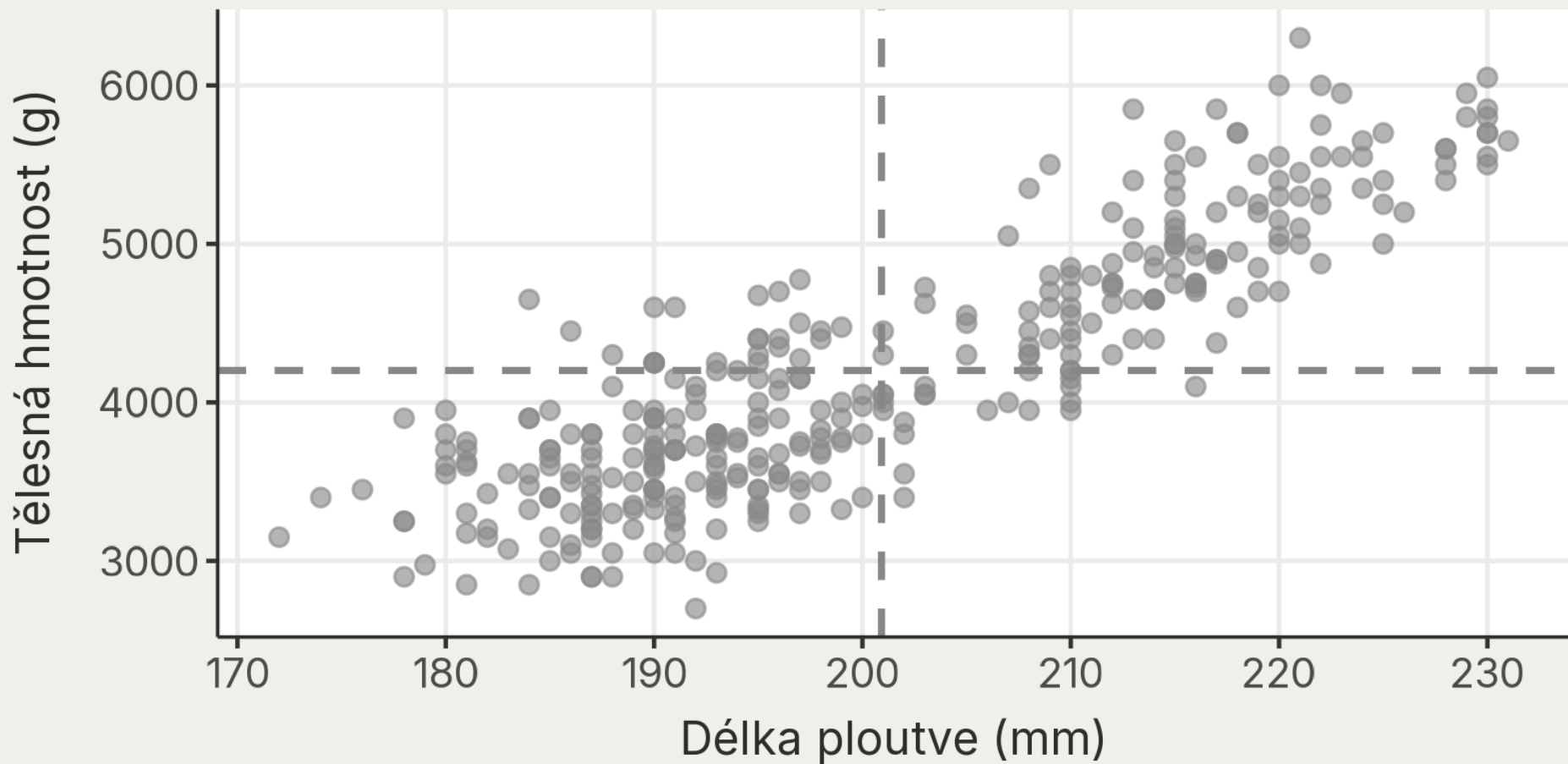
ŘEKNĚTE SOUSEDovi ASPOŇ JEDNU VLASTNOST

30 s každý — pak porovnejte, co jste vymysleli.

Zachytit **směr** a **těsnotu** vztahu a nezáviset na tom, v jakých jednotkách měříme.

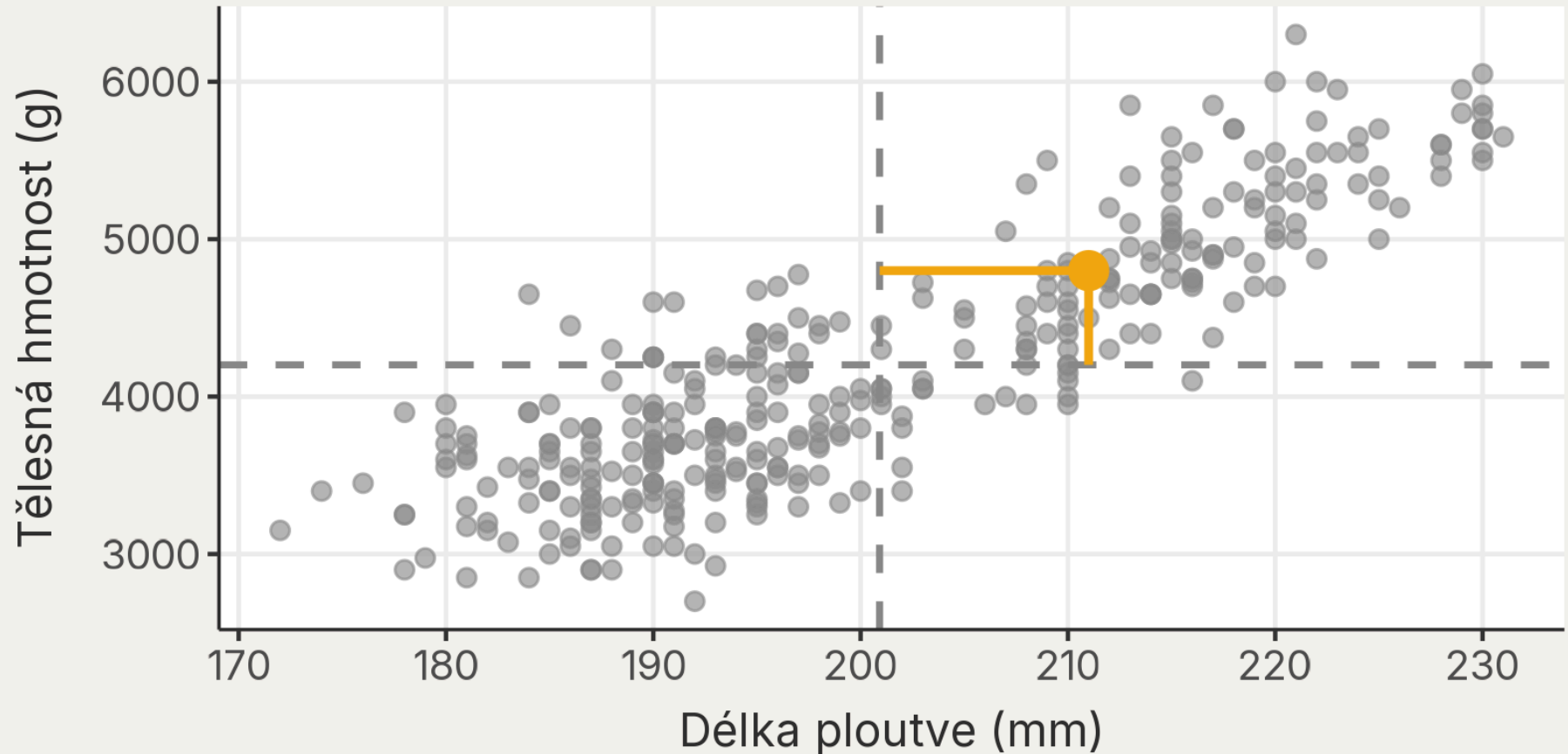
Jak spočítat společný pohyb?

Průměry v grafu



Střed grafu určíme dvěma průměry: délka 200.9 mm, hmotnost 4202 g.

Jeden bod pod lupou

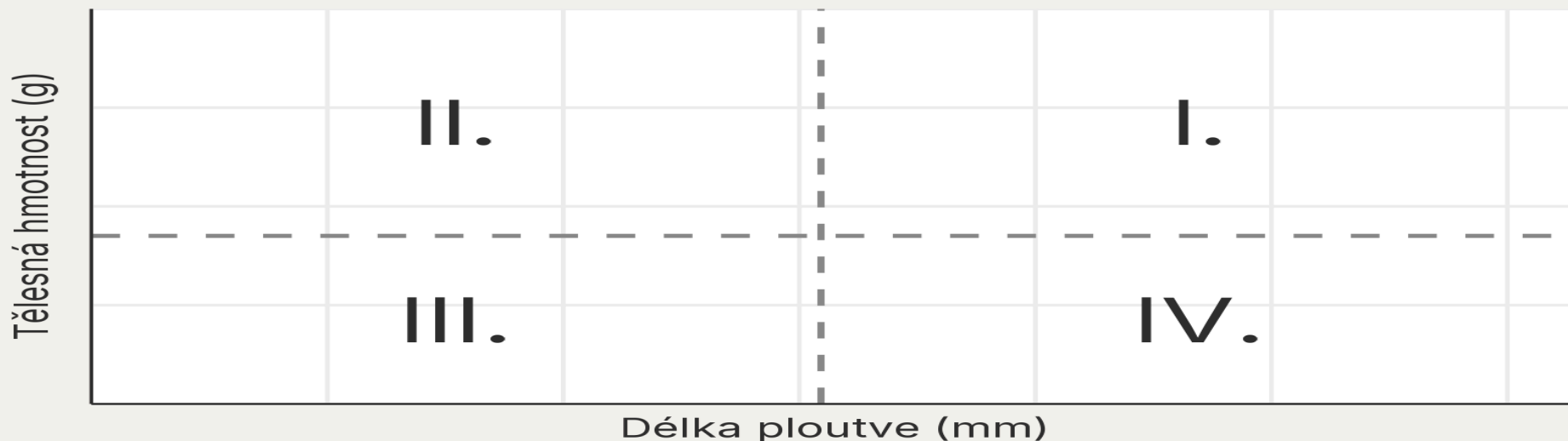


BOD NAD PRŮMĚREM V OBOU PROMĚNNÝCH

Obě odchylky jsou kladné → jejich **součin** je také **kladný**.



Ve které další části grafu budou součiny také kladné?



I. a III. kvadrant

Vpravo nahoře a vlevo dole.

II. a IV. kvadrant

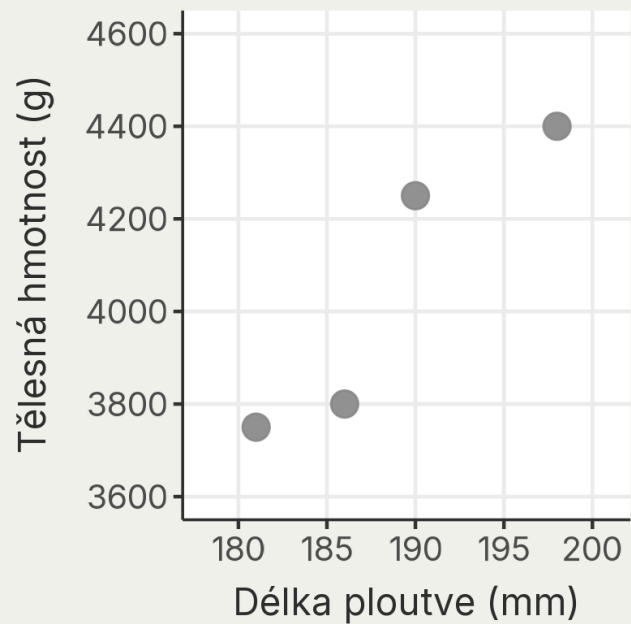
Vlevo nahoře a vpravo dole.

Všechny čtyři stejně

Na poloze vůči středu nezáleží.

Než hlasujeme: vysvětlíte sousedovi znaménka obou odchylek v jedné možnosti.

Čtyři tučňáci ručně



Délka	Hmotnost	Délka odch.	Hmotnost odch.
181	3750	-7,75	-300
186	3800	-2,75	-250
190	4250	1,25	200
198	4400	9,25	350

Průměry:
Délka ploutve: **188,75** mm
Hmotnost: **4050** g

Kovariance je „průměr součinů odchylek“

kovariance = **součet součinů odchylek** / (počet pozorování – 1)

Kovariance sčítá součiny odchylek. Rozhoduje **znaménko i velikost** každého součinu, nikoli samotný počet bodů v kvadrantech.

Délka odch.	Hmotnost odch.	Součin
-7,75	-300	2325
-2,75	-250	687,5
1,25	200	250
9,25	350	3237,5

součet součinů odchylek = 6500

počet pozorování – 1 = 3

kovariance = $6500 / 3 \approx \mathbf{2166,67 \text{ mm} \cdot \text{g}}$

Odchylky a jejich součiny v R

```
1 vec_prumer_delky <-  
2   mean(x = vec_delka_ploutve_mm)  
3 vec_prumer_hmotnosti <-  
4   mean(x = vec_hmotnost_g)  
5 vec_odchylka_delky <-  
6   vec_delka_ploutve_mm - vec_prumer_delky  
7 vec_odchylka_hmotnosti <-  
8   vec_hmotnost_g - vec_prumer_hmotnosti  
9 vec_souciny <-  
10  vec_odchylka_delky * vec_odchylka_hmotnosti
```

Každý součin patří jednomu ze stejných čtyř pozorovaných tučňáků.

Stejný výpočet pomocí `cov()`

```
1 vec_soucin <-  
2   sum(vec_souciny)  
3 vec_pocet_minus_jedna <-  
4   length(x = vec_souciny) - 1  
5 vec_kovariance <-  
6   vec_soucin / vec_pocet_minus_jedna  
7 vec_kovariance
```

```
[1] 2166.667
```

```
1 vec_kovariance <-  
2   cov(  
3     x = vec_delka_ploutve_mm,  
4     y = vec_hmotnost_g  
5   )  
6 vec_kovariance
```

```
[1] 2166.667
```

Záleží na jednotkách?

Co se stane s kovariancí, když délku ploutve přepočteme na centimetry a hmotnost na kilogramy?

Stejní tučňáci: všech 342 úplných dvojic.

V milimetrech a gramech

Kovariance: 9 824 (mm·g)

V centimetrech a kilogramech

Kovariance: 0,982 (cm·kg)

Směr vztahu se nezměnil, ale číslo ano. Proto chceme **bezrozměrný souhrn**: korelaci.

Číslo nezávislé na jednotkách

Kovariance bez jednotek: Pearsonovo r

PEARSONOVA KORELACE

$r = \text{kovariance} / (\text{SD délky ploutve} \times \text{SD hmotnosti})$

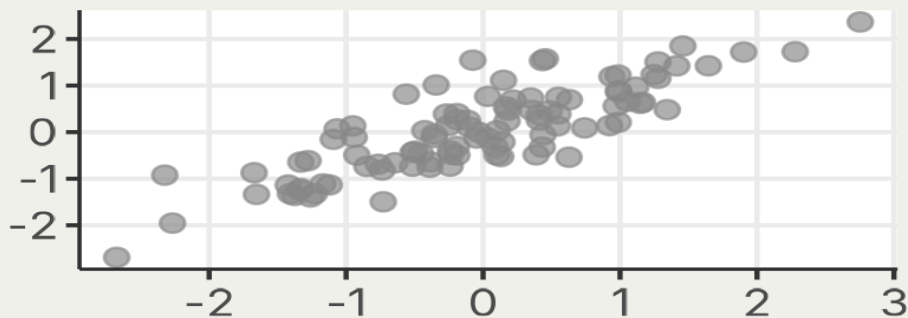
Jednotky v čitateli i jmenovateli se vykrátí.

- r leží vždy mezi -1 a 1
- **znaménko** ukazuje směr vztahu
- čím blíží ± 1 , tím **těsněji** leží body u přímky
- v R: `cor()`

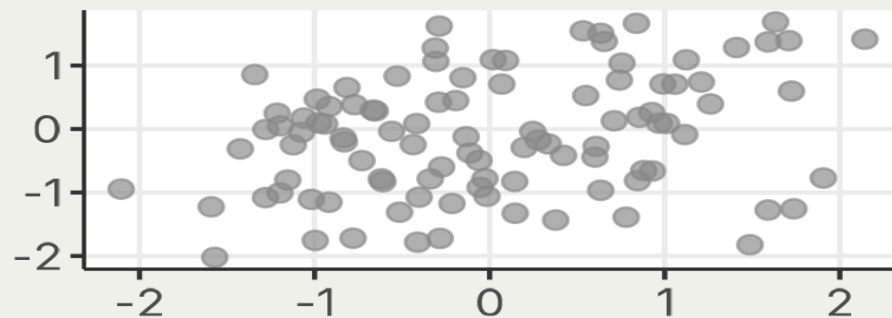
r je **doplňěk ke grafu**, ne jeho náhrada.

Odhadněte korelaci

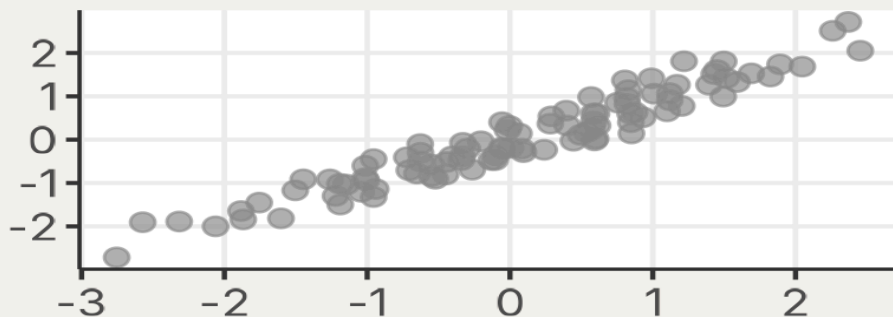
A



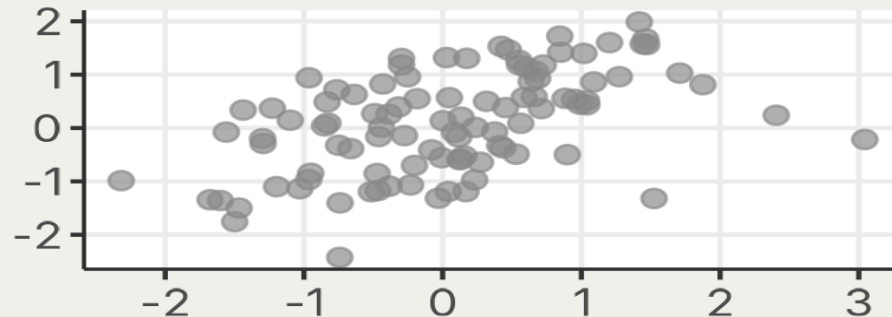
B



C



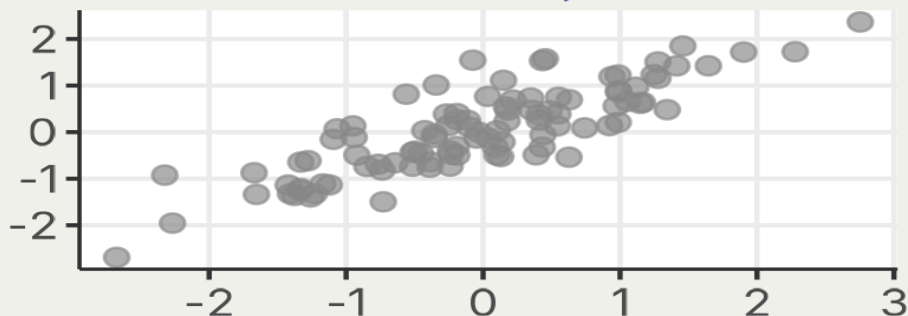
D



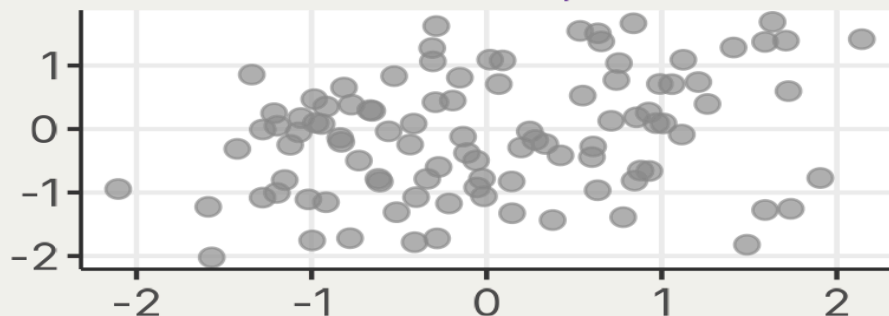
Seřad'te grafy A–D od nejslabšího po nejtěsnější vztah. Jaké r odhadujete pro nejtěsnější z nich?

Stejná data, ale s hodnotou r

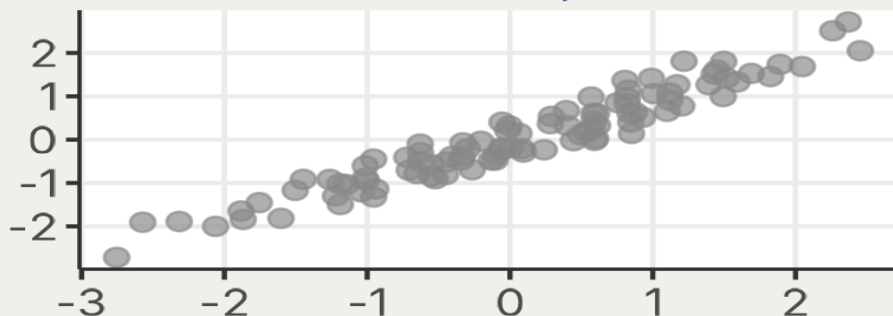
A: $r \approx 0,84$



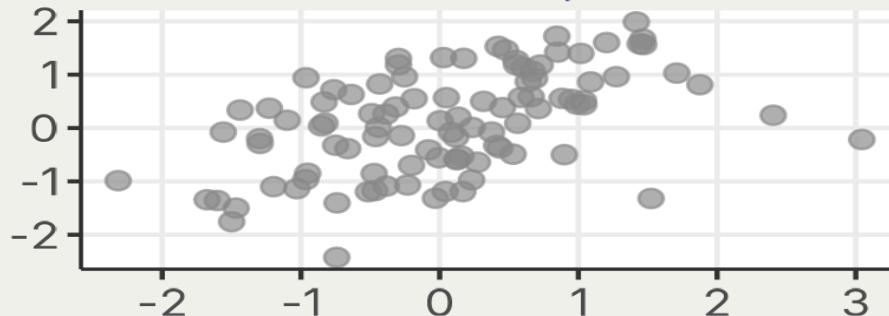
B: $r \approx 0,30$



C: $r \approx 0,96$

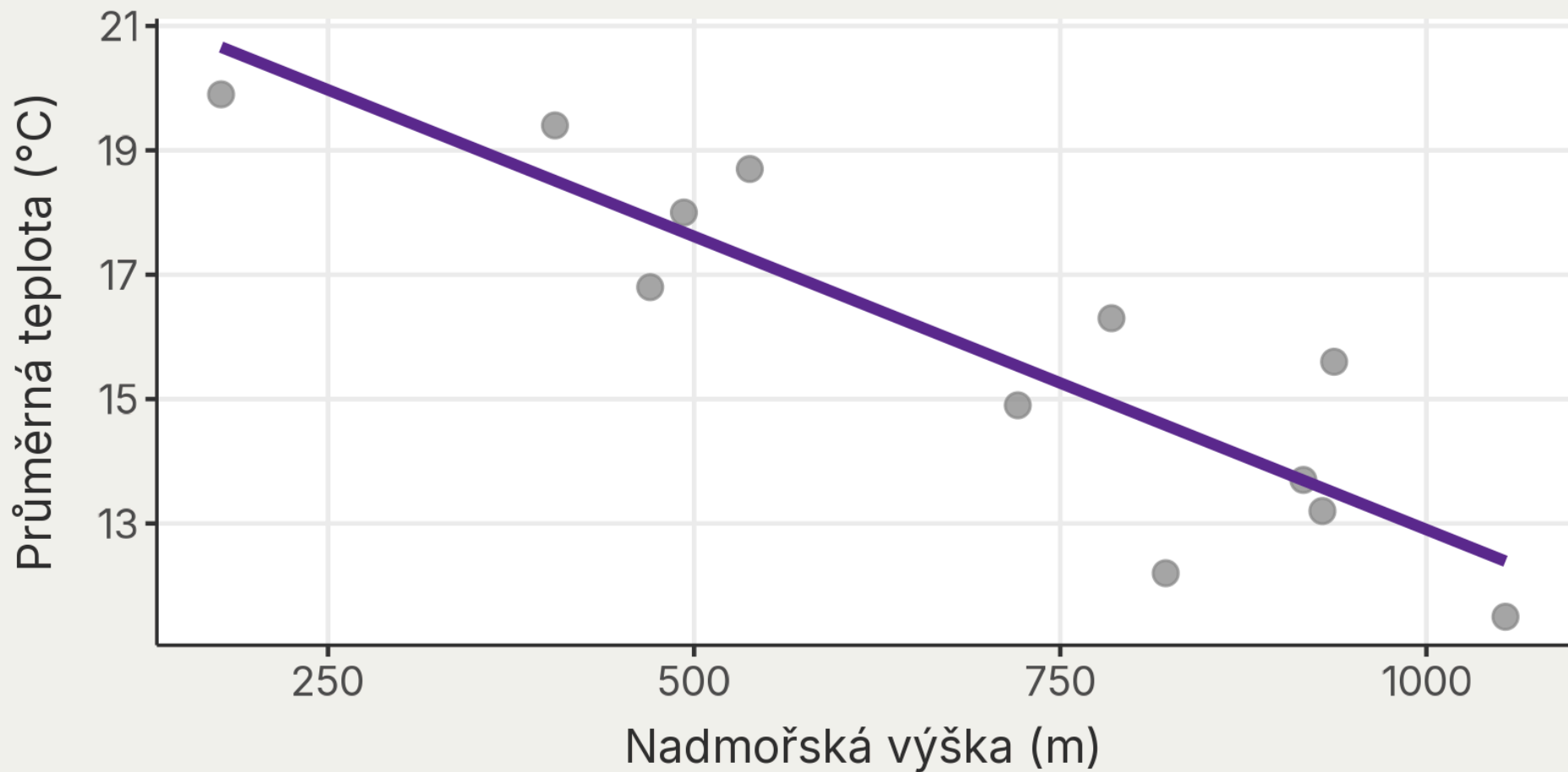


D: $r \approx 0,49$



Čím těsněji body leží podél přímky, tím blíže je r k 1.

Výš v horách, chladněji?



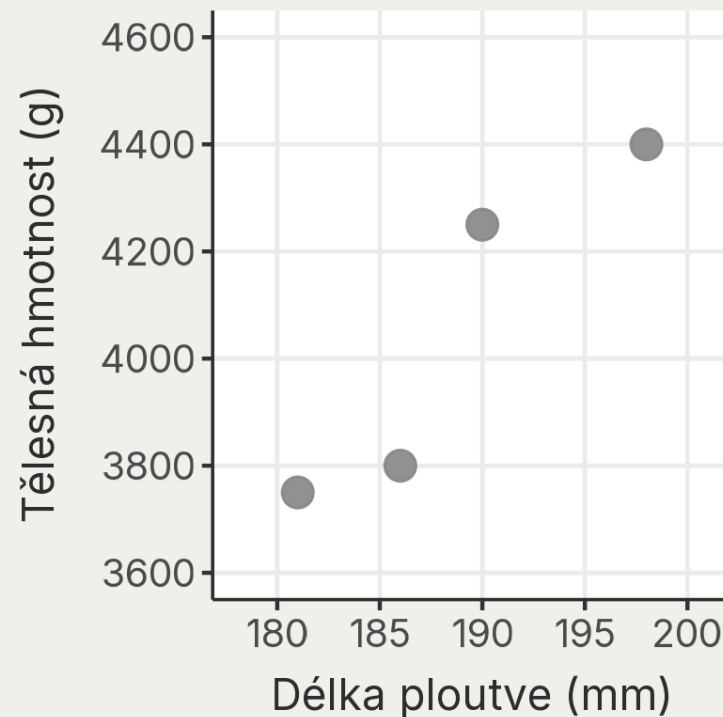
Záporné $r = -0,89$: s vyšší nadmořskou výškou průměrná teplota klesá.

Zkonstruovaný výukový příklad nadmořské výšky a teploty; nejde o terénní měření.

Čtyři tučňáci v kódu

```
1 vec_delka_ploutve_mm <-  
2   c(181, 186, 190, 198)  
3  
4 vec_hmotnost_g <-  
5   c(3750, 3800, 4250, 4400)  
6  
7 cor(  
8   x = vec_delka_ploutve_mm,  
9   y = vec_hmotnost_g,  
10  method = "pearson"  
11 )
```

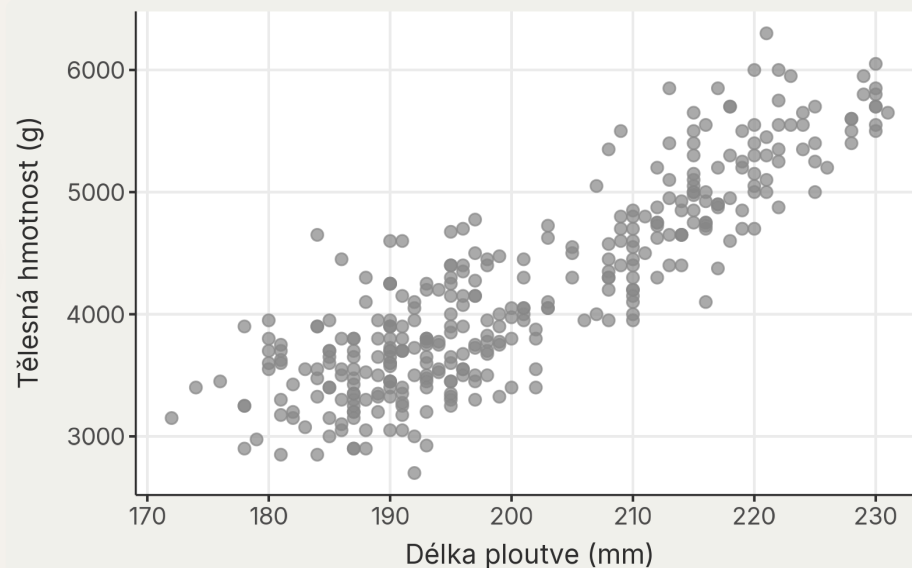
```
[1] 0.9309851
```



Všichni tučňáci v kódu

```
1 cor(  
2   x = data_tucnaci$delka_ploutve_mm,  
3   y = data_tucnaci$hmotnost_g,  
4   method = "pearson"  
5 )
```

```
[1] 0.8712018
```



$r \approx 0,87$: silný rostoucí vztah — a přesto při stejné délce ploutve zůstává rozptýlení hmotnosti.

Pořadí místo hodnot: Spearmanova korelace

SPEARMANOVA KORELACE

Hodnoty nahradí jejich **pořadím** (nejkratší ploutev = 1, druhá nejkratší = 2, ...) a korelaci spočítá z pořadí.

- ptá se, zda vztah **stále roste** (nebo stále klesá), i když není přímkou
- jednotlivá vzdálená pozorování ji ovlivní méně

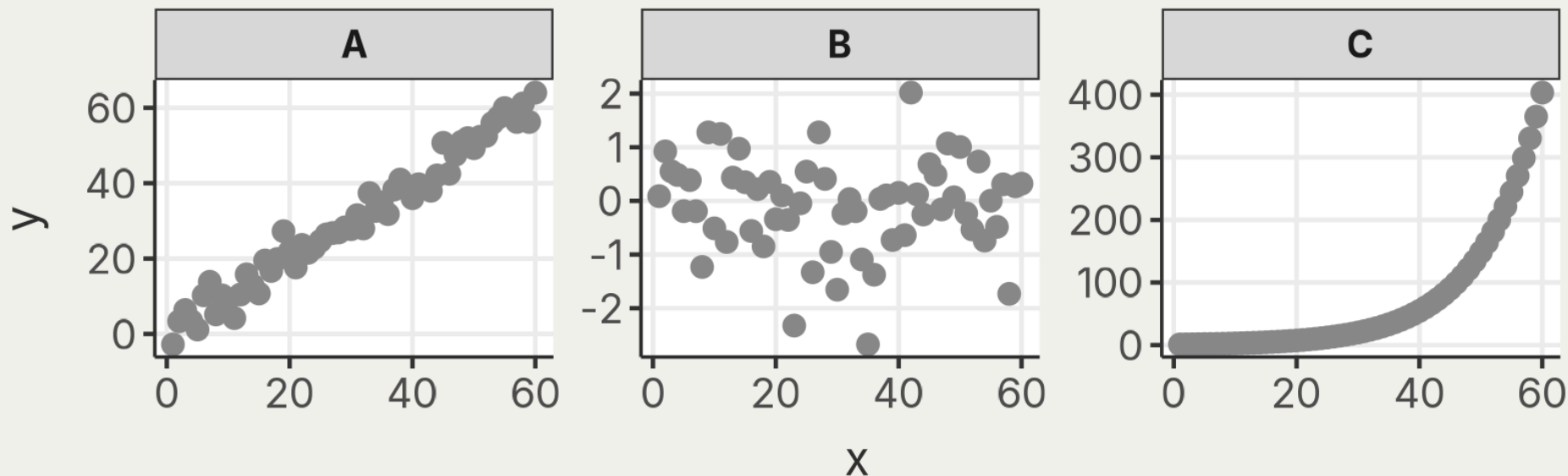
```
1 cor(  
2   x = data_tucnaci$delka_ploutve_mm,  
3   y = data_tucnaci$hmotnost_g,  
4   method = "spearman"  
5 )
```

```
[1] 0.8399741
```

U tučňáků vychází podobně jako Pearson (0,87 a 0,84): oblak roste zhruba po přímkce.

Kde se Pearson a Spearman mohou lišit?

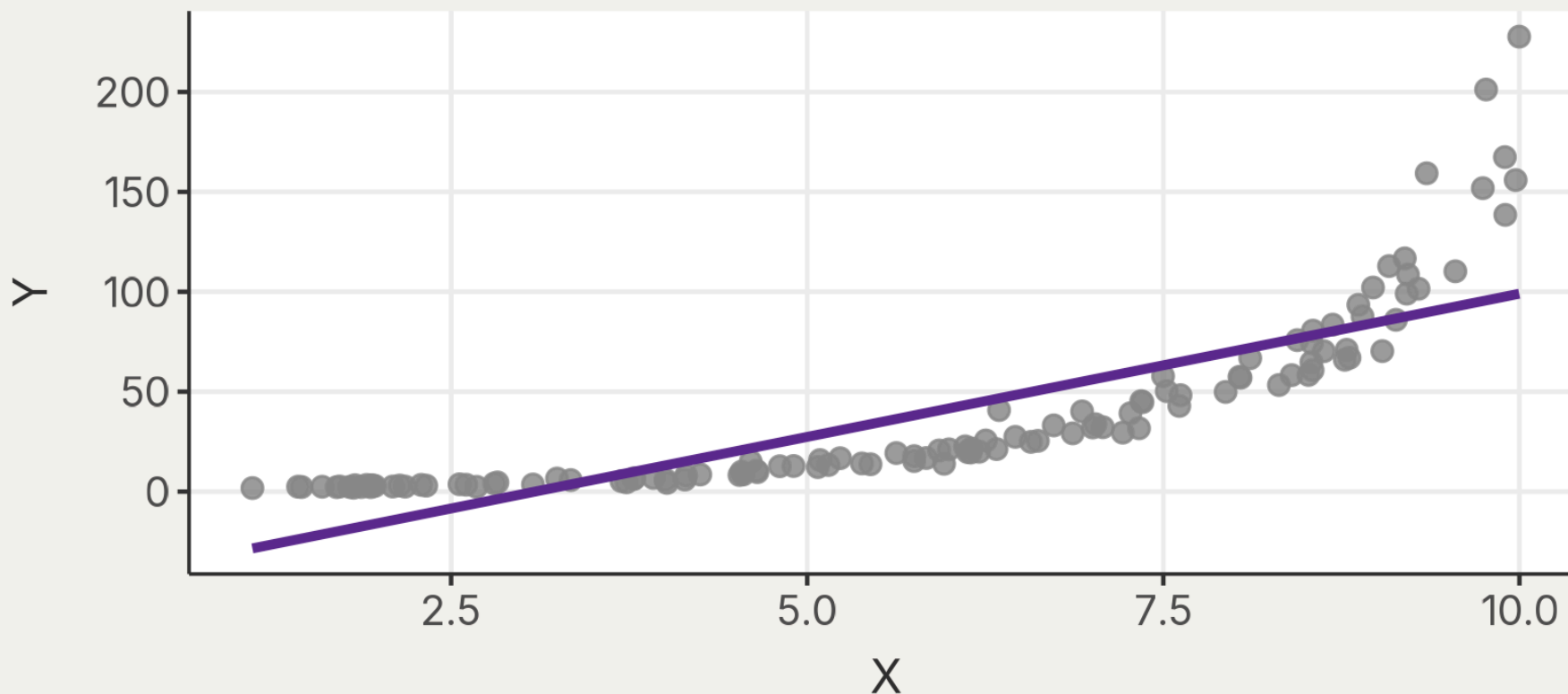
Který z těchto tří grafů podle vás povede k největšímu rozdílu mezi oběma koeficienty?



Největší rozdíl: **C**. Pořadí roste, ale vztah není přímka: Pearson **0,82**, Spearman **1,00**.

Rostoucí vztah nemusí být přímka

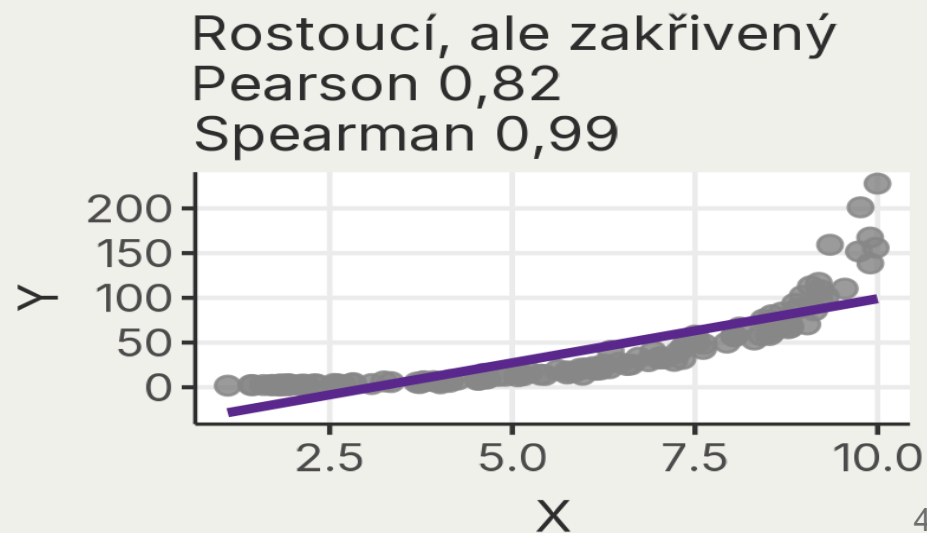
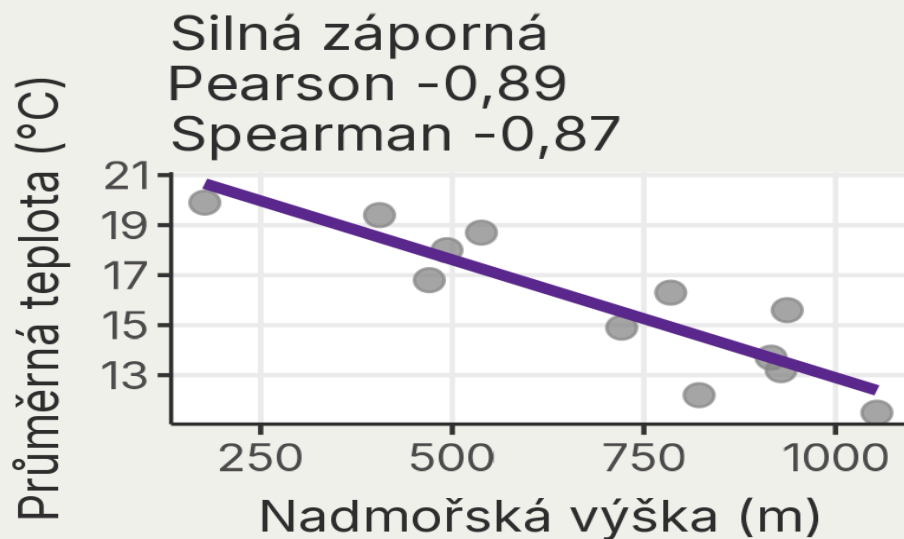
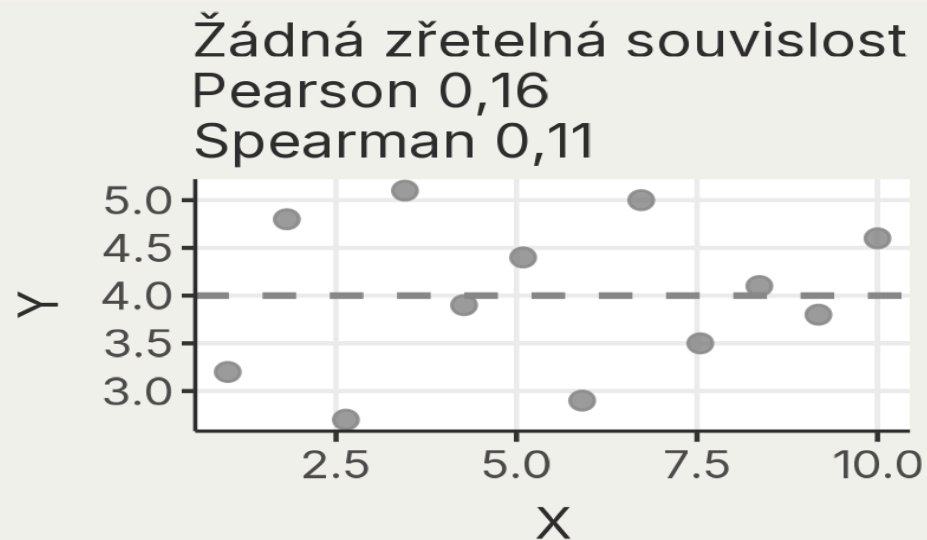
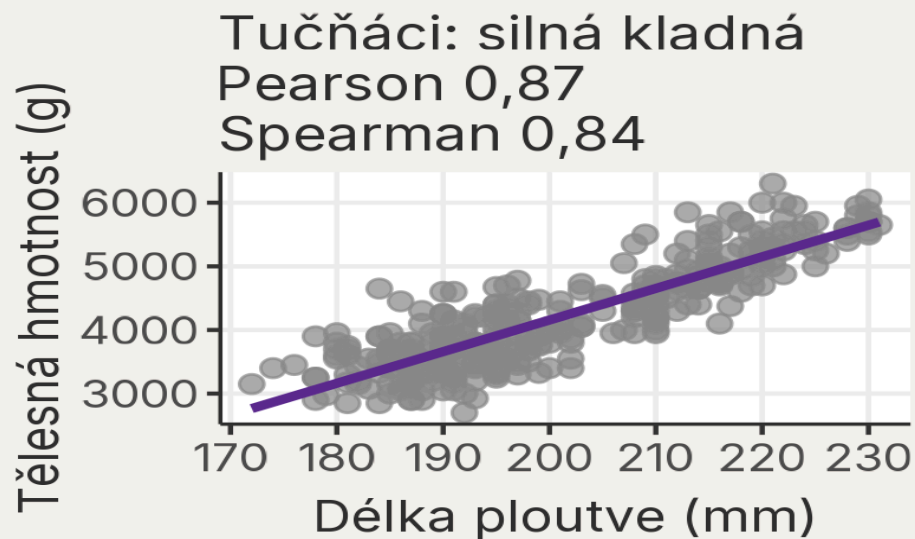
Stále rostoucí, ale zakřivený vztah



Stejný typ vztahu jako v grafu C, tentokrát s rozptýlenými body. Pearson 0,82, Spearman 0,99: pořadí hodnot roste téměř dokonale, ale body neleží na přímce.

Zkonstruovaný výukový příklad; nejde o měření tučňáků.

Čtyři situace korelace



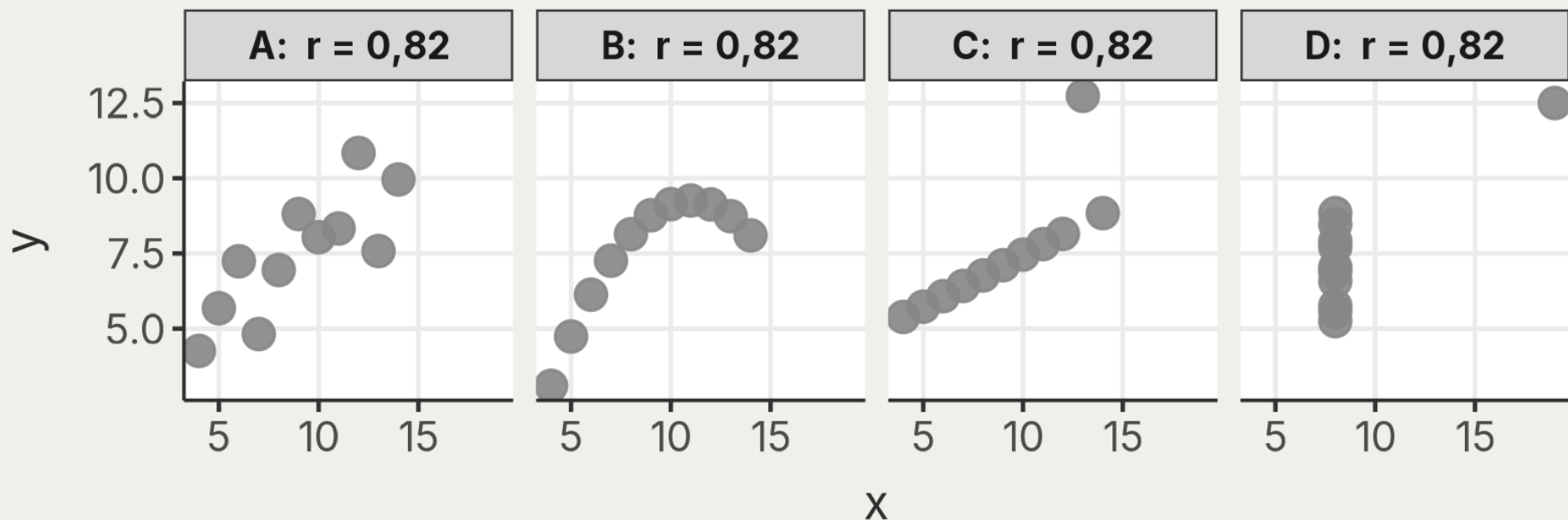
Čtyři soubory, stejná čísla

Soubor	Průměr x	Průměr y	Rozptyl x	Rozptyl y	r
A	9,00	7,50	11,00	4,13	0,82
B	9,00	7,50	11,00	4,13	0,82
C	9,00	7,50	11,00	4,12	0,82
D	9,00	7,50	11,00	4,12	0,82

Mají všechny čtyři soubory stejný vztah mezi x a y?

Načrtněte sousedovi, jak by mohl vypadat bodový graf jednoho z nich.

Stejná čísla, jiné grafy



A rozptýlená
přímka

B křivka

C přímka a jeden
vzdálený bod

D jediný bod
vytváří „vztah“

Stejně **průměry, rozptyly i r** mohou popisovat **úplně jiné vztahy**. Proto se nejdřív díváme na graf.

Co nám to říká právě o tučňácích?

Mají tučňáci s delší ploutví také vyšší tělesnou hmotnost?

Pearson i Spearman vyšly podobně

Pearson: 0,87, Spearman: 0,84.

Graf ukázal rostoucí oblak. Oba koeficienty shrnují celkovou asociaci; vztahy uvnitř druhů mohou být jiné.

Korelace neřekne, odkud vztah pochází

Část vztahu může souviset s druhem nebo celkovou velikostí těla.

Korelace ukazuje asociaci, ne příčinu



SPOLEČNÁ ZMĚNA DVOU PROMĚNNÝCH

Korelace popisuje, zda se dvě proměnné mění společně.

SAMA NEODDĚLÍ PŘÍČINU

Korelace sama neodliší přímý efekt od sdílené příčiny.

Jak vybrat první graf?

Tři dvojice, tři různé otázky

Délka ploutve × hmotnost

Druh tučňáka × hmotnost

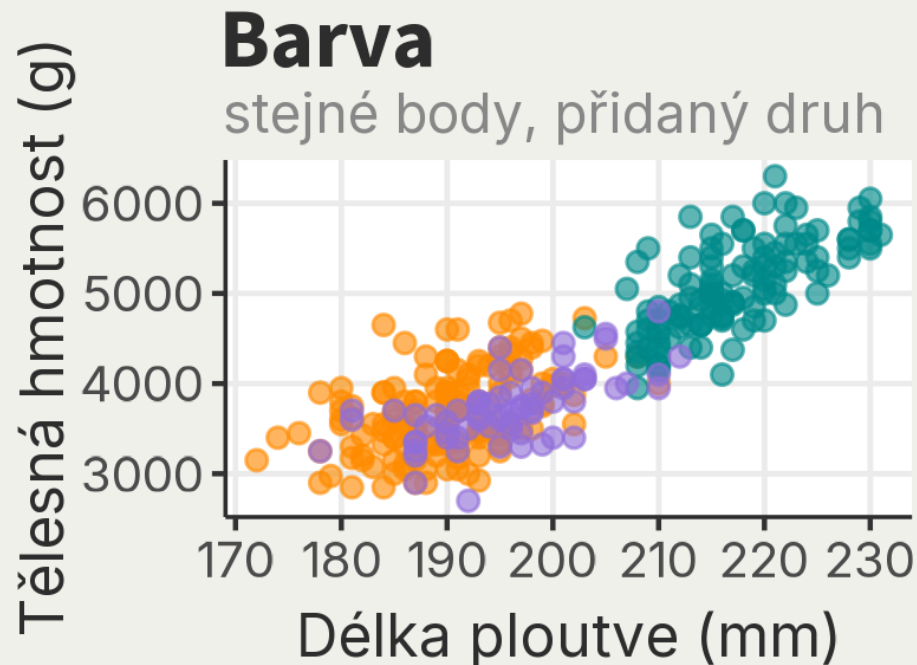
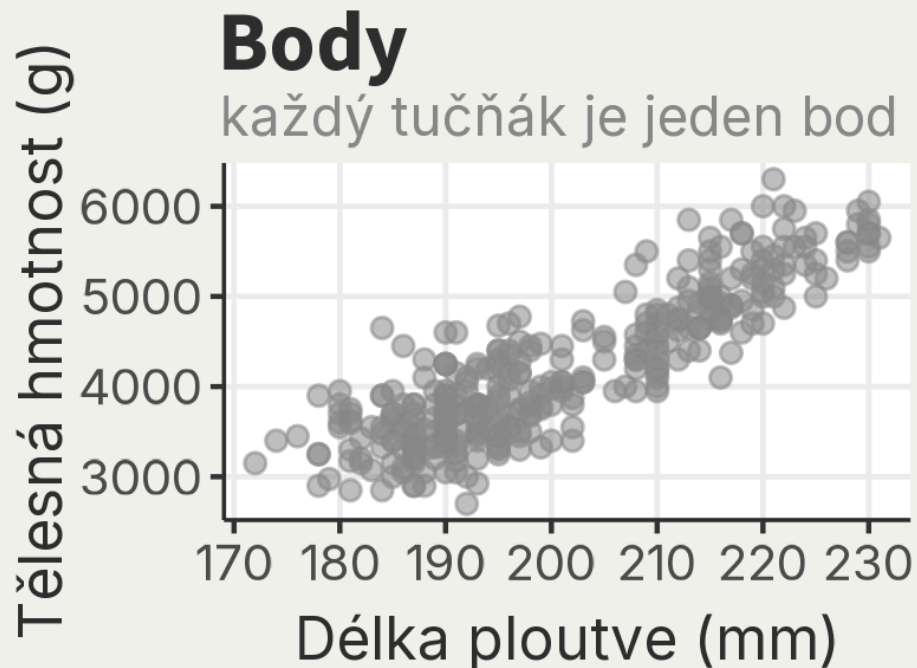
Druh tučňáka × ostrov

SKUPINKA 3–4 · 2 MIN

Ke každé dvojici navrhnete první graf a jednou větou zdůvodníte, co z něj chcete přechít.

Ploutev × hmotnost: body, nebo barva?

Numerická × numerická: obě proměnné mají číselnou osu

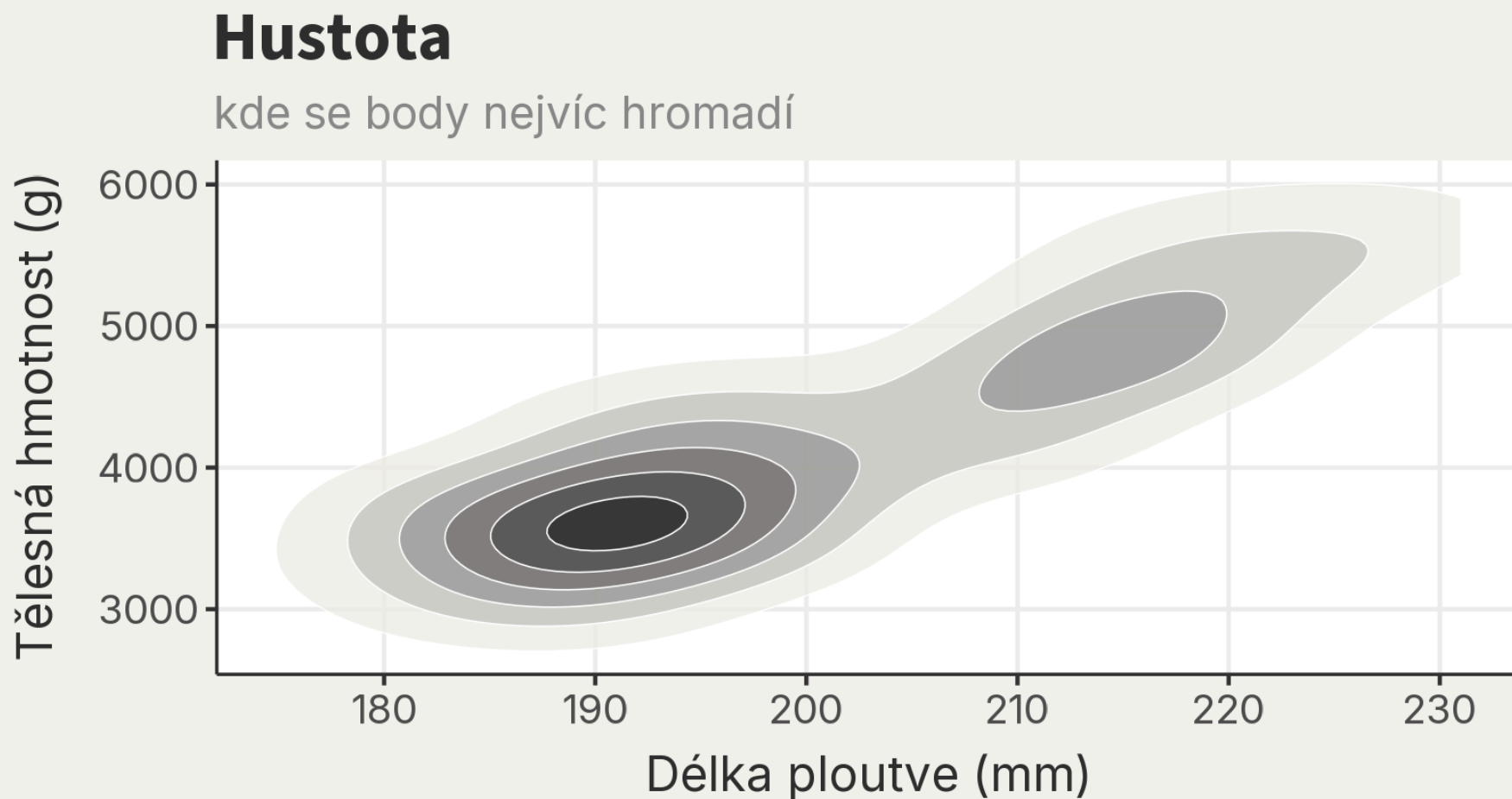


● Adelie ● Chinstrap ● Gentoo

Každý tučňák je jeden bod: první volba pro dvě numerické proměnné.

Barva nepřidá osu, ale novou otázku: je vztah stejný ve všech

Ploutev × hmotnost: kde je bodů nejvíc?



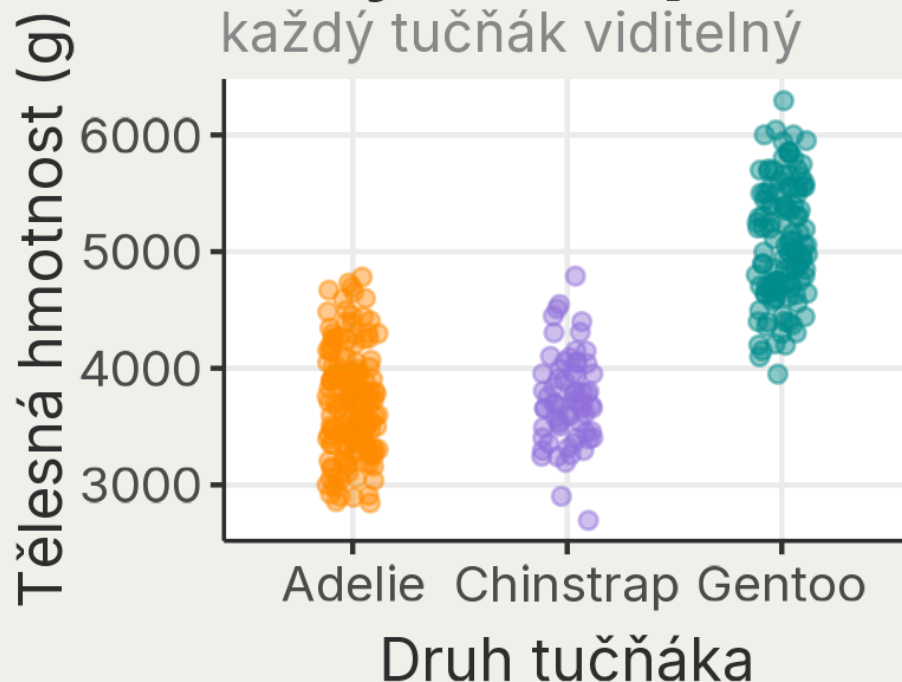
Hustota odpovídá na jinou otázku: kde je dat nejvíc? Jednotlivé body ale zmizí.

Druh × hmotnost: body, nebo krabice?

Kategoriální × numerická: jedna proměnná vytváří skupiny, druhá je číslo

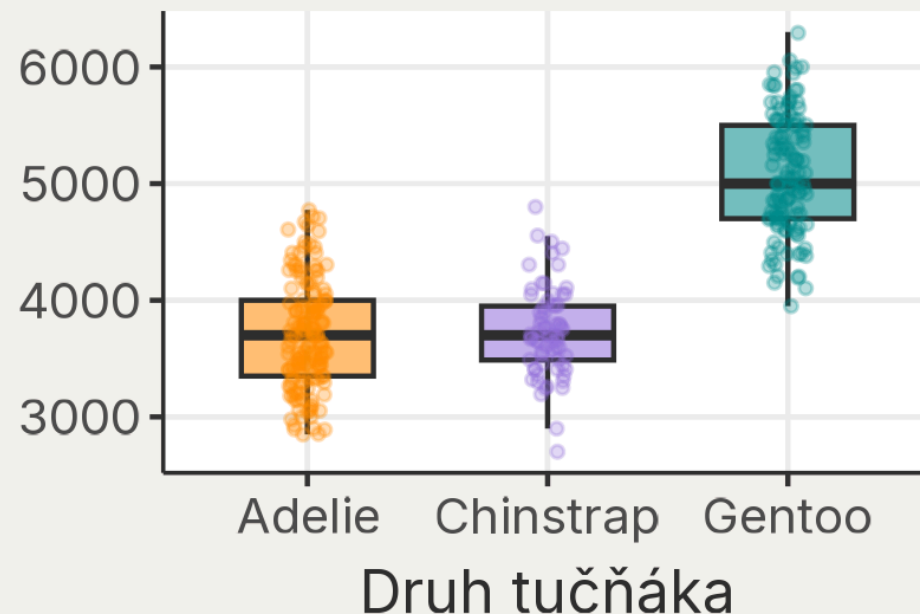
Body ve skupinách

každý tučňák viditelný



Boxplot + body

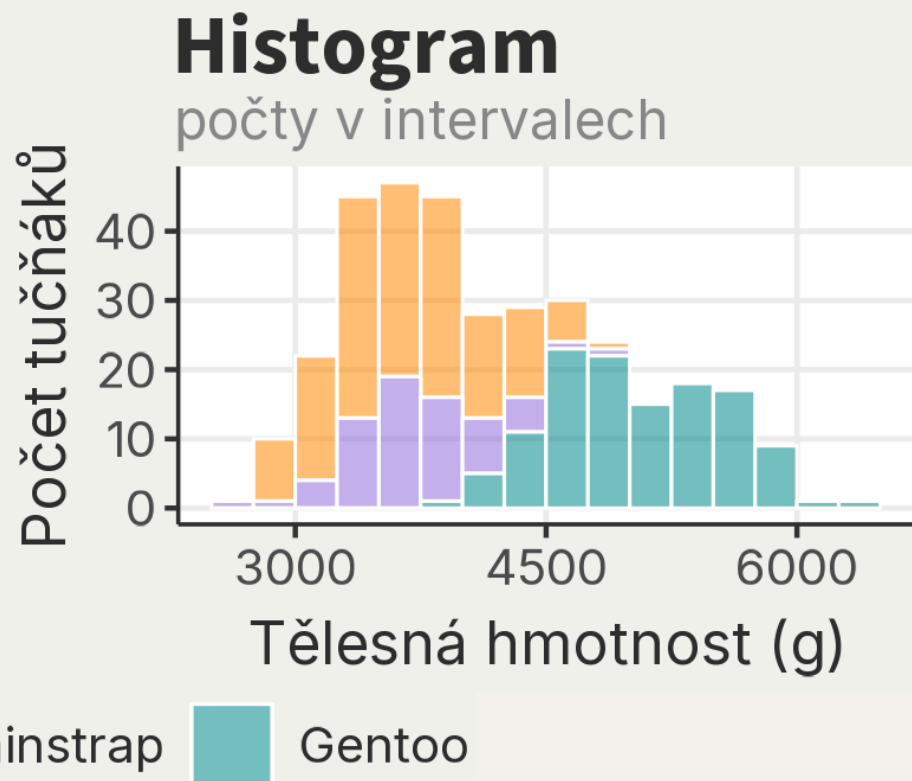
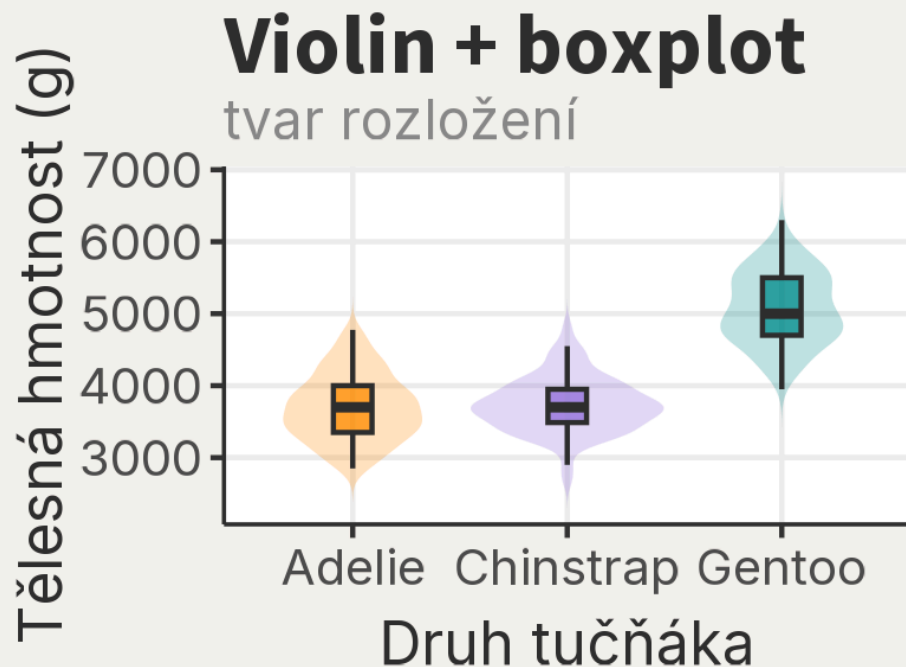
souhrn i jednotlivé body



Body ve skupinách nic neskrývají:
vidíme každého tučňáka.

Boxplot dá rychlý souhrn, ale
samotná krabice by skryla tvar dat.

Druh × hmotnost: jaký tvar má rozložení?



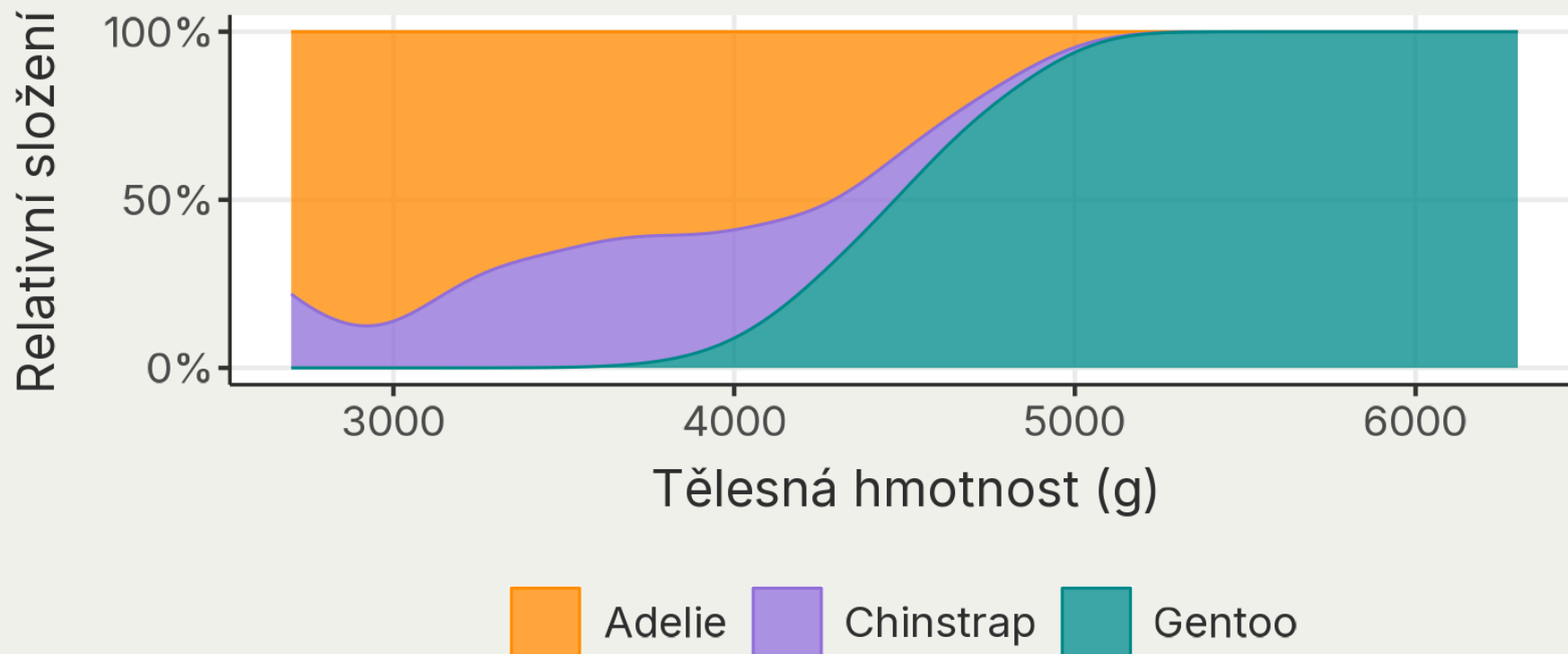
Violin ukáže tvar rozložení; krabice uvnitř dá rychlý souhrn.

Histogram ukáže skutečné počty tučňáků v intervalech hmotnosti.

Druh × hmotnost: který druh převažuje?

Skládaná hustota

který druh převažuje při dané hmotnosti



Skládaná hustota neukazuje počty. Ukazuje relativní složení druhů podél hmotnosti.

Který graf byste udělali jako první?

Otázka: liší se tělesná hmotnost mezi druhy a jak moc se druhy překrývají?

Body ve skupinách

Vidíme každého tučňáka.

Boxplot + body

Rychleji čteme typickou hodnotu a rozptýlení.

Violin / histogram

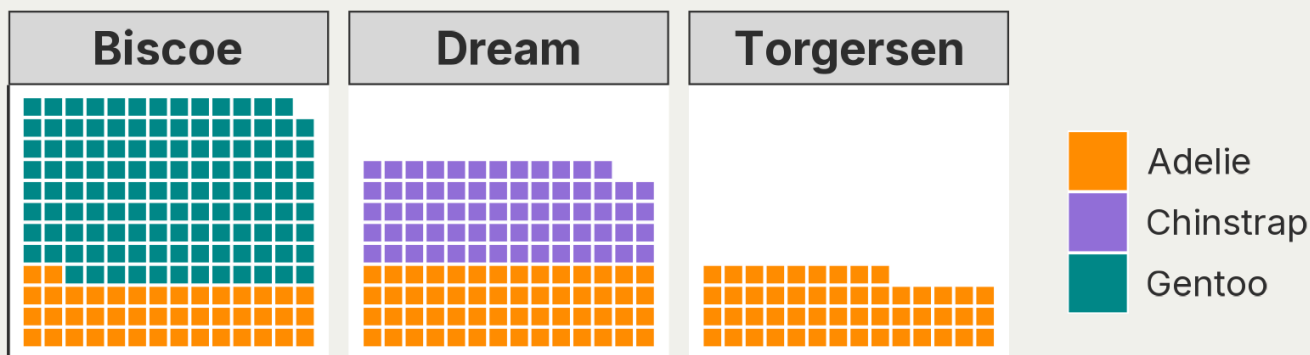
Lépe čteme tvar nebo počty v intervalech.

Druh × ostrov: každý tučňák jedním čtverečkem

Kategoriální × kategoriální: čteme počty nebo podíly v kombinacích kategorií

Waffle plot

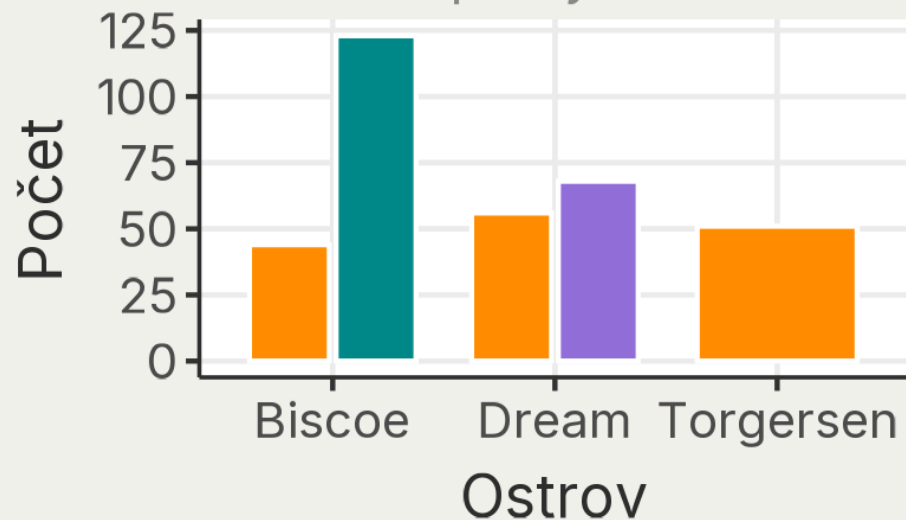
každý čtvereček je jeden tučňák



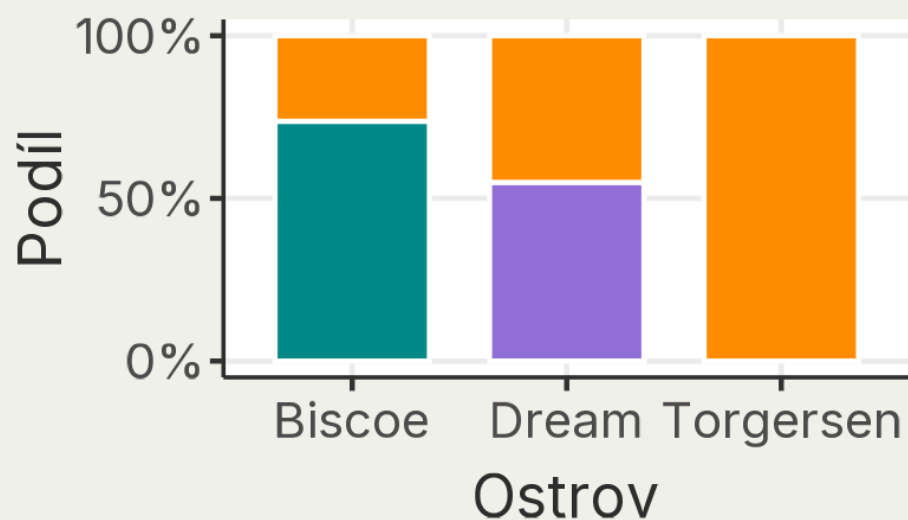
Waffle plot ukáže počty jako jednotky: vidíme složení i velikost skupin najednou.

Druh × ostrov: sloupce počtů a podílů

Skupinové sloupce
absolutní počty



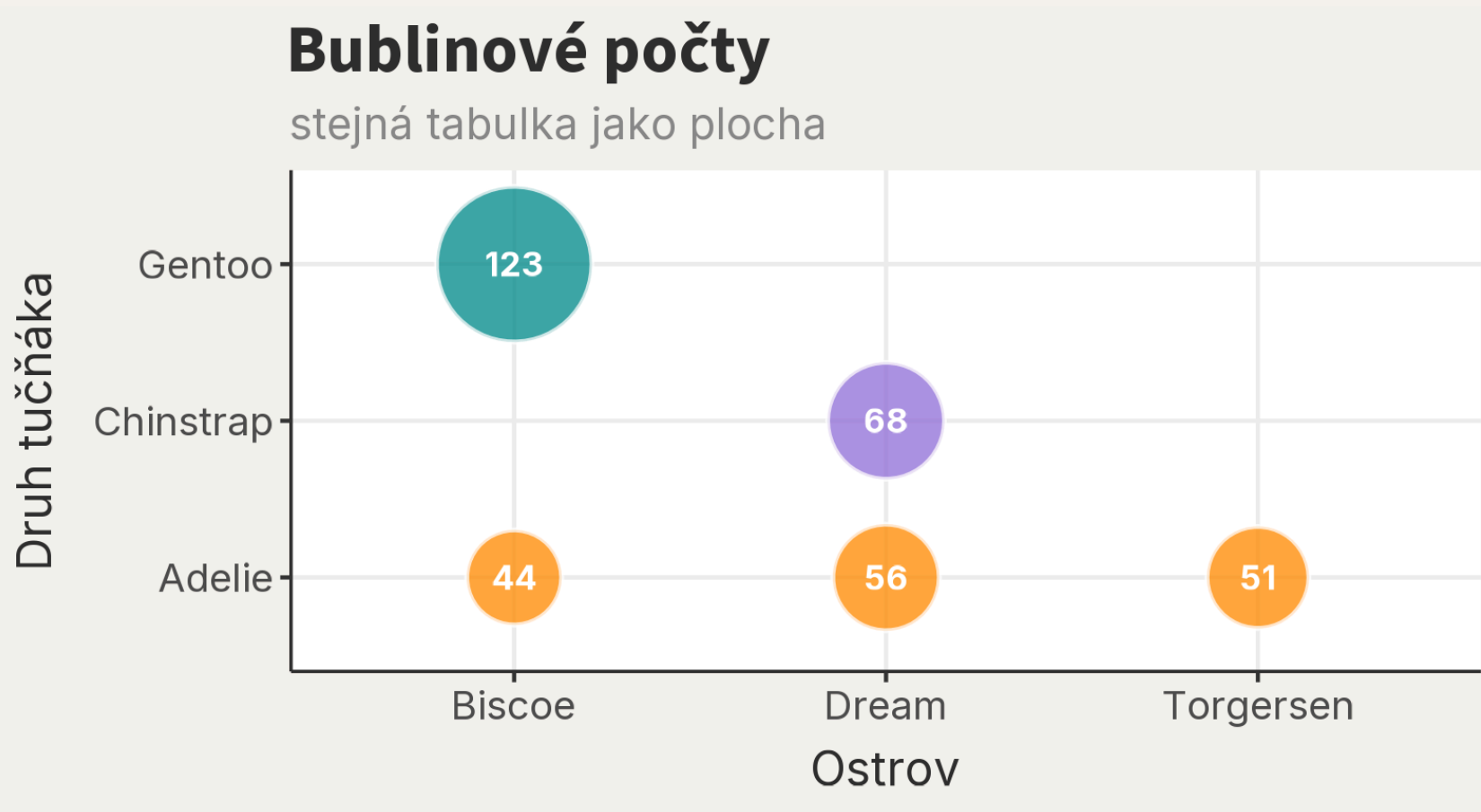
Skládané podíly
složení každého ostrova



Skupinové sloupce porovnají absolutní počty mezi ostrovy.

Skládané podíly ukážou složení ostrova, ale skryjí, kolik tučňáků tam bylo.

Druh × ostrov: hezčí, nebo čitelnější?



Bublíny jsou efektní, ale porovnávání ploch je pro oko těžší než porovnávání délek sloupců.

Kdy potřebujeme počty a kdy podíly?

POČTY

Kdy potřebujeme vědět, kolik pozorování opravdu máme?

PODÍLY

Kdy nás zajímá složení skupiny bez ohledu na její velikost?

Který graf zvolit jako první?

Typy proměnných	Dobrý první graf	Co v něm čteme
numerická × numerická	bodový graf	směr, tvar, rozptýlení, odlehlá pozorování
kategoriální × numerická	body ve skupinách; boxplot + body	rozdíly mezi skupinami, překryv, rozptýlení uvnitř skupin
kategoriální × kategoriální	tabulka četností; sloupce počtů nebo podílů	složení kombinací, počty vs. podíly

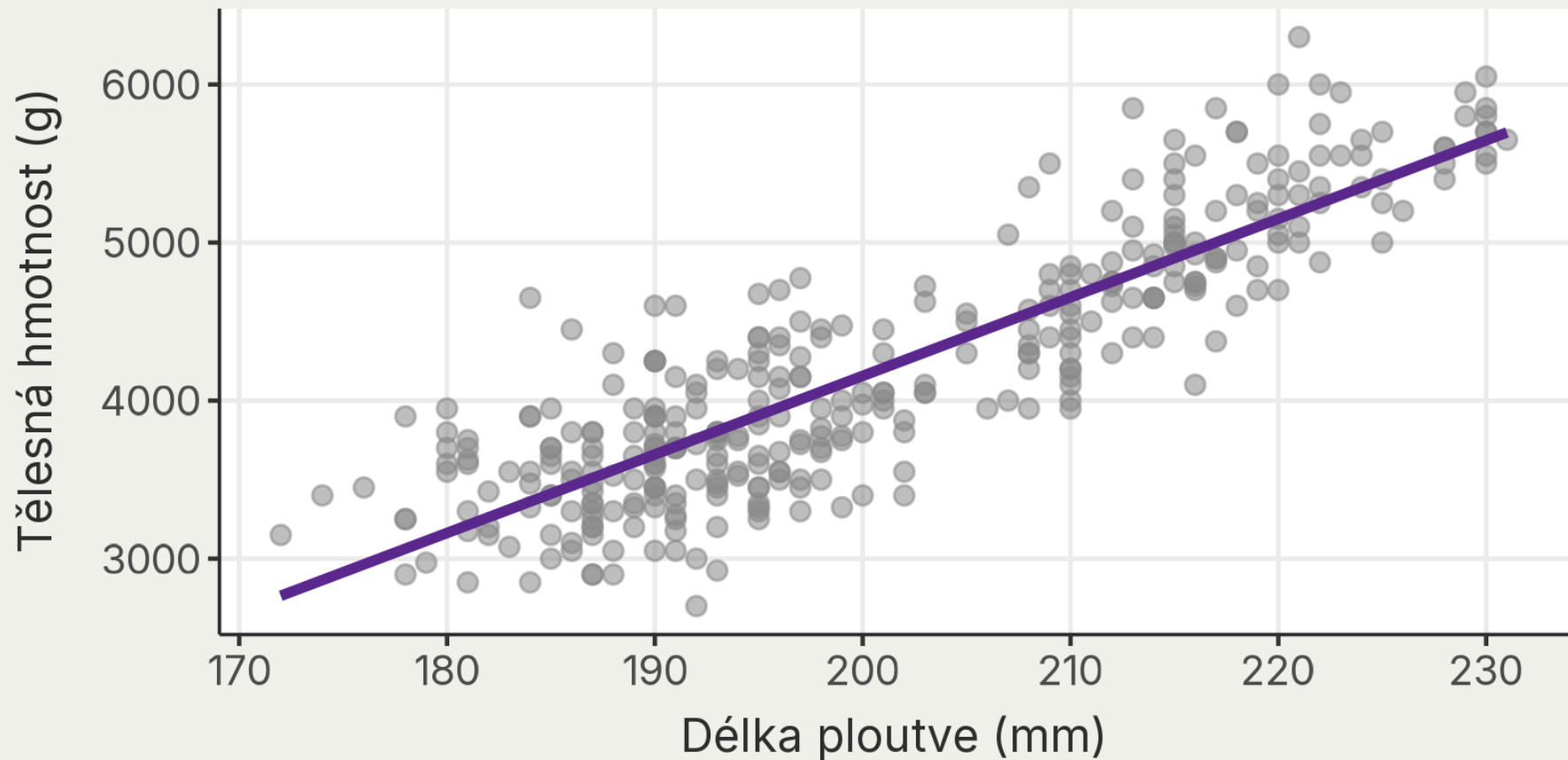
Typy proměnných zúží možnosti.
Biologická otázka rozhodne, co má
graf zvýraznit.

CO NESMÍME SKRÝT?

Boxplot skryje tvar dat, podíly skryjí
počty.

Závěr

Co přidá **model**?



Přímka popisuje **vztah v původních jednotkách**: o kolik gramů se typicky mění hmotnost na každý 1 mm délky ploutve? To bude téma následující lekce.

Shrnutí

- Pro dvě numerické proměnné je první volbou **bodový graf**.
- Z grafu čteme současně **směr vztahu** i **rozptýlení bodů**.
- **Korelace** vztah shrne číslem, ale nenahradí samotný graf.
- Typ dvojice proměnných zúží výběr grafu; otázka rozhodne, co má graf ukázat.
- Viditelný vzor ještě není důkaz **příčiny**.

Co potřebujete vidět, než napíšete „delší ploutev, těžší tučňák“?

Jaké tři otázky položíte autorovi, který vám ukáže jen $r = 0,87$ bez grafu?

ŘEKNĚTE SOUSEDŮVI JEDNOU VĚTOU

Kterou z otázek byste položili jako první, a proč?