

Kvantitativní analýza výzkumných dat

Kateřina Lukavská, Ph.D.

Co je kvantitativní analýza dat?

Deskriptivní analýza / statistika

- příprava dat pro analýzu (čištění, proměnné, chybějící hodnoty)
- deskriptivní statistika (základní popis získaných dat)

Inferenční analýza / statistika

- použití získaných dat pro vyvozování závěrů o zkoumané populaci a pro testování hypotéz (problematika výběrových souborů, NHST paradigma, statistické testy)

Pro opakování a procvičování lze využít Moodle e-learning

Deskriptivní analýza / statistika

- příprava dat pro analýzu (čištění, proměnné, chybějící hodnoty)
- deskriptivní statistika (základní popis získaných dat)



PADA 1

<https://dl2.cuni.cz/course/view.php?id=3651>

enroll in the course (method: Psy, no key required)

Inferenční analýza / statistika

- použití získaných dat pro vyvozování závěrů o zkoumané populaci a pro testování hypotéz (problematika výběrových souborů, NHST paradigma, statistické testy)



PADA 2

<https://dl2.cuni.cz/course/view.php?id=3828>

enroll in the course (method: Psy, no key required)

Statistika

- odvětví matematiky
- týká se **pravděpodobnosti**
 - Jaká je pravděpodobnost, že padne 6 při jednom hodu šestistěnnou kostkou?
 - Jaká je pravděpodobnost, že padne 6 nebo 1 při jednom hodu šestistěnnou kostkou?
 - Jaká je pravděpodobnost, že padne 6 a poté 1 ve dvou po sobě jdoucích hodech?
- Ve výzkumu používáme statistiku k provádění závěrů o určité populaci (populacích) na základě dat ze vzorků nebo k testování hypotéz v rámci paradigmatu NHST (testování statistické nulové hypotézy).

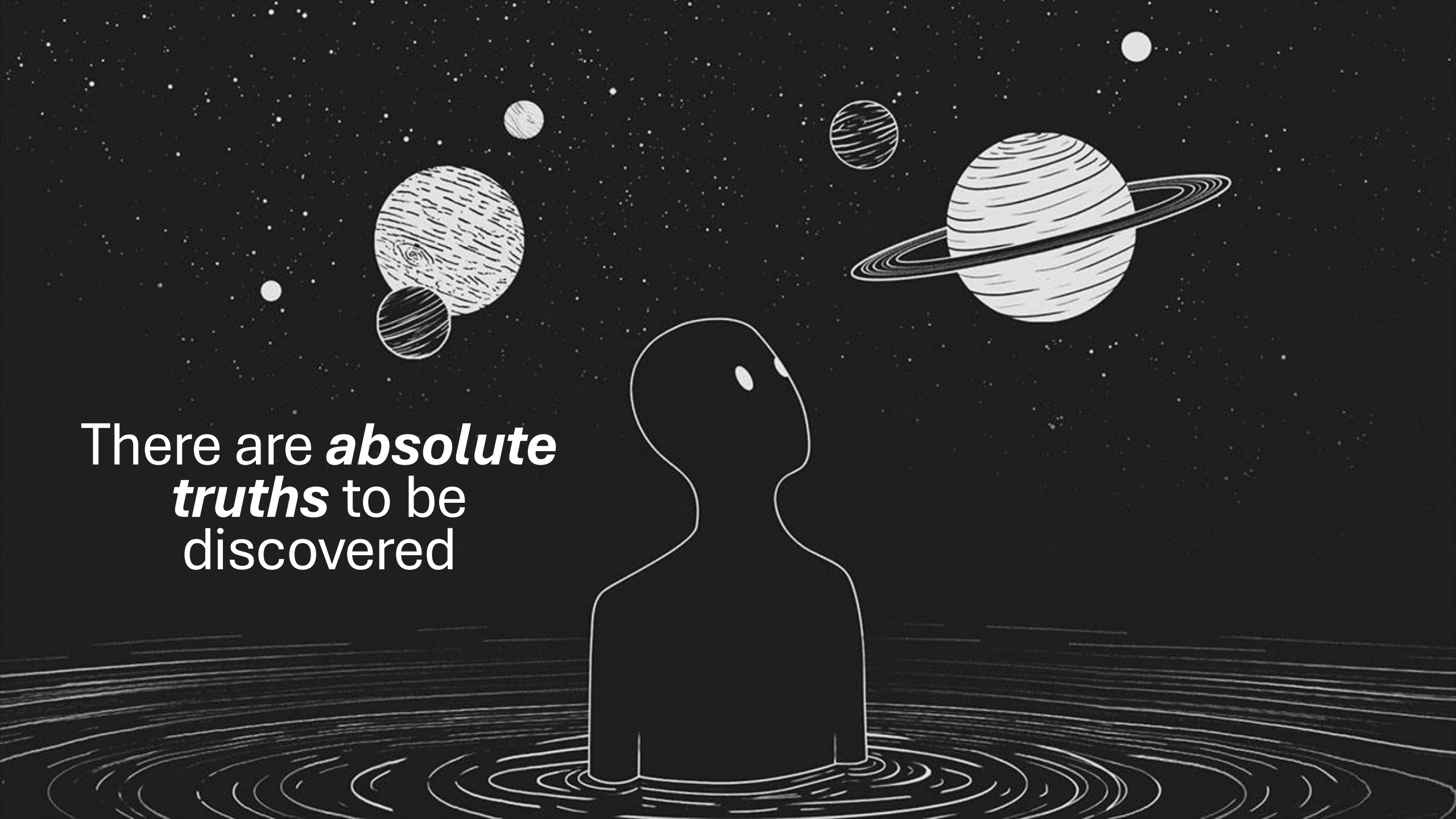
NHST paradigm

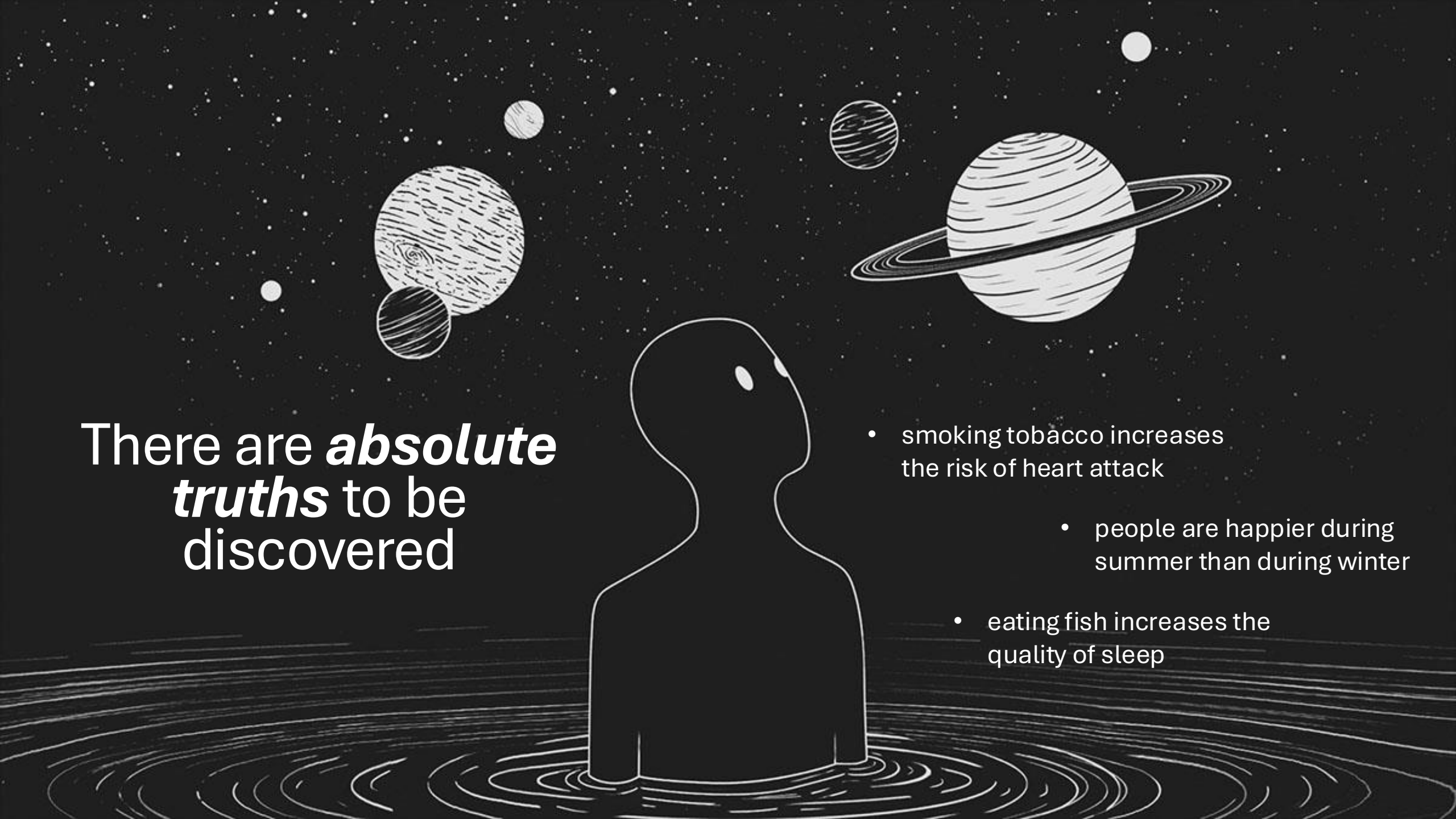
Null Hypothesis Significance Testing

- mainstream... ale alternativy nabývají na popularitě (např. Bayesian statistics)
- Klíčové koncepty:
 - alternativní a nulová hypotéza
 - chyba prvního druhu (a její pravděpodobnost) / alpha error (probability)
 - statistická významnost / significance
 - chyba druhého druhu (a její pravděpodobnost) / beta error (probability)
 - statistická síla / power

Hypothesis / Hypotéza

There are ***absolute
truths*** to be
discovered



A black and white illustration of a person's silhouette looking up at a starry sky. The person is standing in water, with ripples visible around their feet. The sky is filled with stars and several planets, including a large one with rings (Saturn) and others with different textures. The overall mood is contemplative and cosmic.

There are ***absolute truths*** to be discovered

- smoking tobacco increases the risk of heart attack
- people are happier during summer than during winter
- eating fish increases the quality of sleep



...based on experience
(i.e., data)
and theory!

- smoking tobacco increases the risk of heart attack
- people are happier during summer than during winter
- eating fish increases the quality of sleep

Hypotéza

Pojídání ryb zlepšuje kvalitu spánku

[Lidé, kteří jedí ryby (fish-eaters) mají vyšší kvalitu spánku než lidi, kteří nejedí ryby (non-pescatarians)]



Hypotéza

- (musí mít *teoretickou plausibilitu*)
 - = *dává smysl zabývat se její platností (=1)*
- může být **testována**
 - = ***dá se považovat za neplatnou (=0) a dá se získat důkaz, že s určitou pravděpodobností není neplatná (tj. platí, =1)***

Null hypothesis / Nulová hypotéza

Pojídání ryb zlepšuje kvalitu spánku

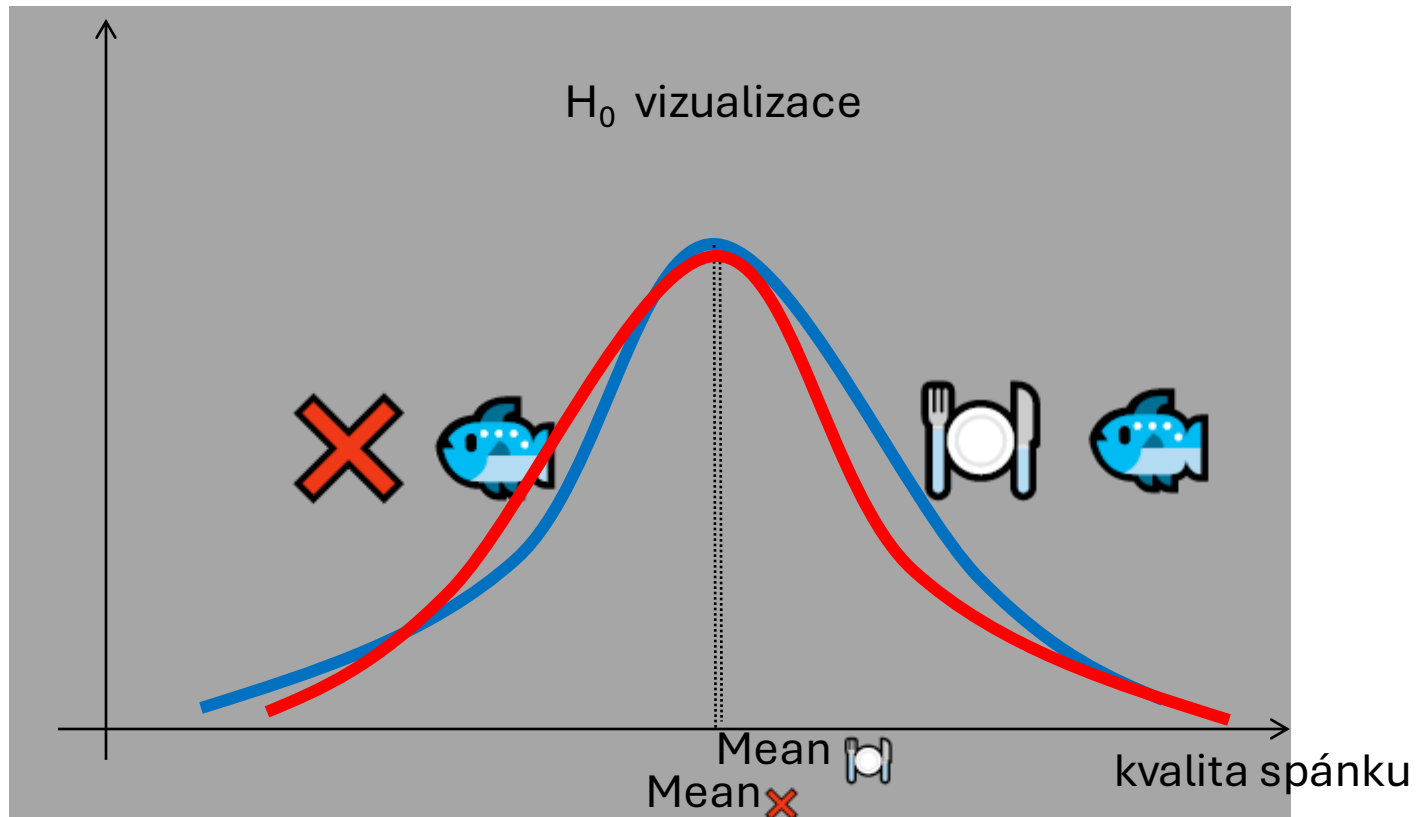
[Lidé, kteří jedí ryby (fish-eaters) mají vyšší kvalitu spánku než lidi, kteří nejedí ryby (non-pescatarians)]

Pojídání ryb nezlepšuje kvalitu spánku

[Lidé, kteří jedí ryby (fish-eaters) nemají vyšší kvalitu spánku než lidi, kteří nejedí ryby (non-pescatarians)]

Nulová hypotéza

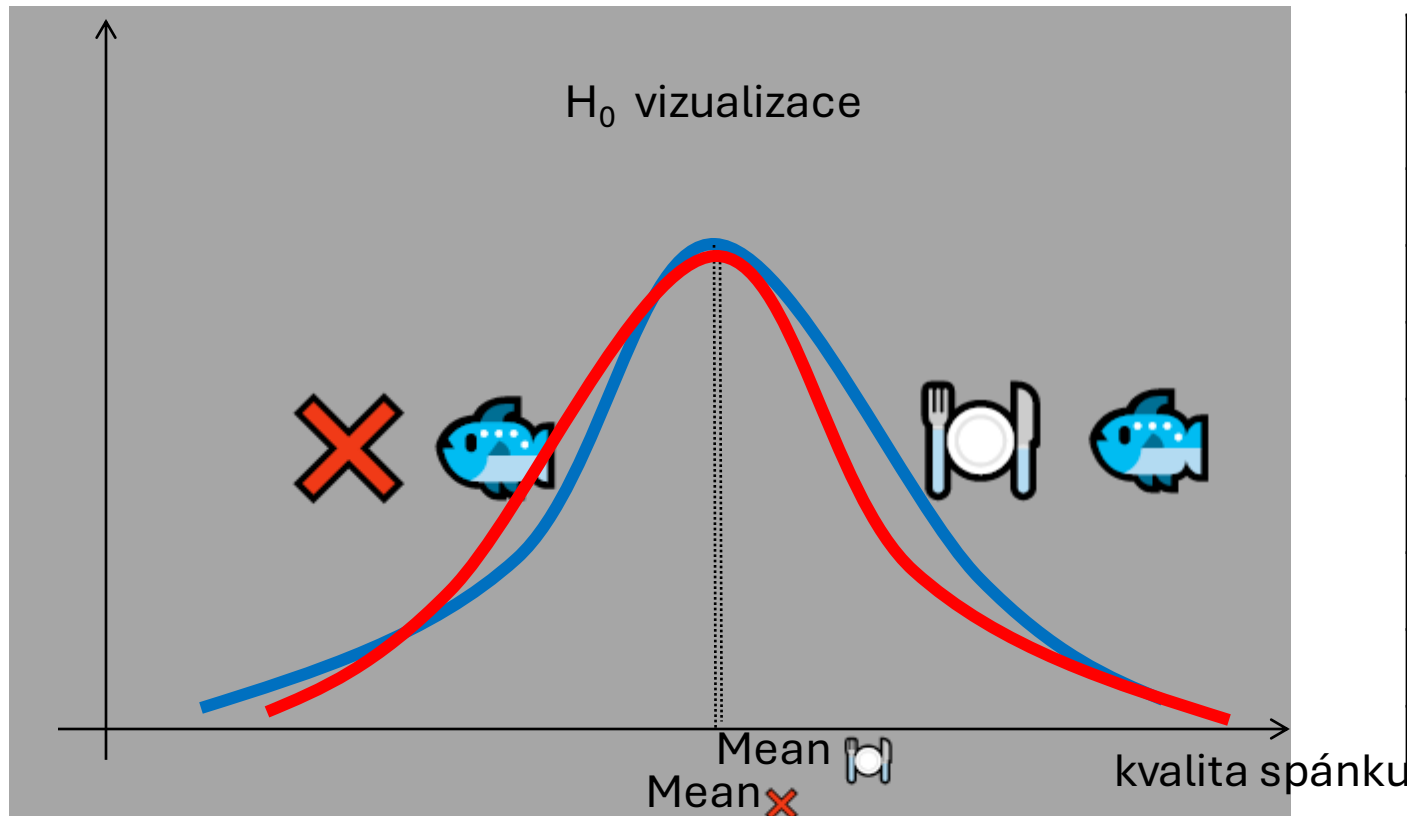
H_0 : Lidé, kteří jedí ryby nemají vyšší kvalitu spánku než lidi, kteří nejedí ryby



Nulová hypotéza

H_0 : Lidé, kteří jedí ryby nemají vyšší kvalitu spánku než lidi, kteří nejedí ryby

- Jaká data potřebujeme, abychom to vyvrátili?

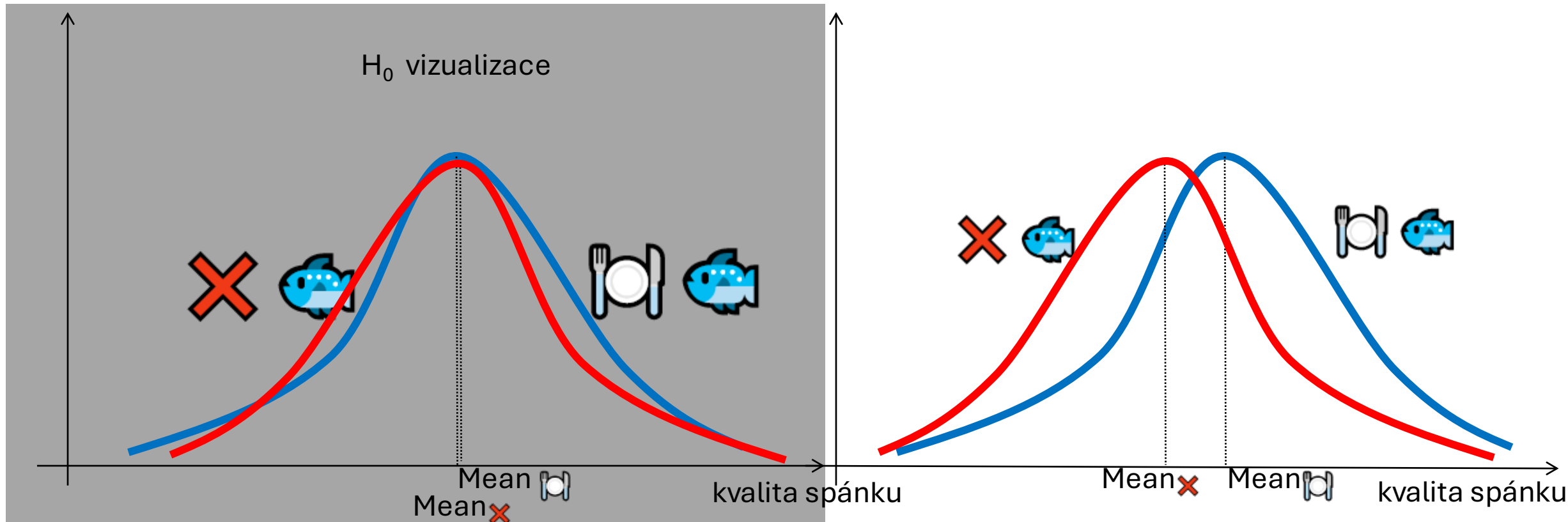


ID	Eats fish?	Sleep quality
001	yes	78
002	yes	66
003	no	69
004	no	50
005	no	95
006	yes	99
007	yes	45
...	...	...
999	no	77

Nulová hypotéza

H_0 : Lidé, kteří jedí ryby nemají vyšší kvalitu spánku než lidi, kteří nejedí ryby

- Jaká data potřebujeme, abychom to vyvrátili?

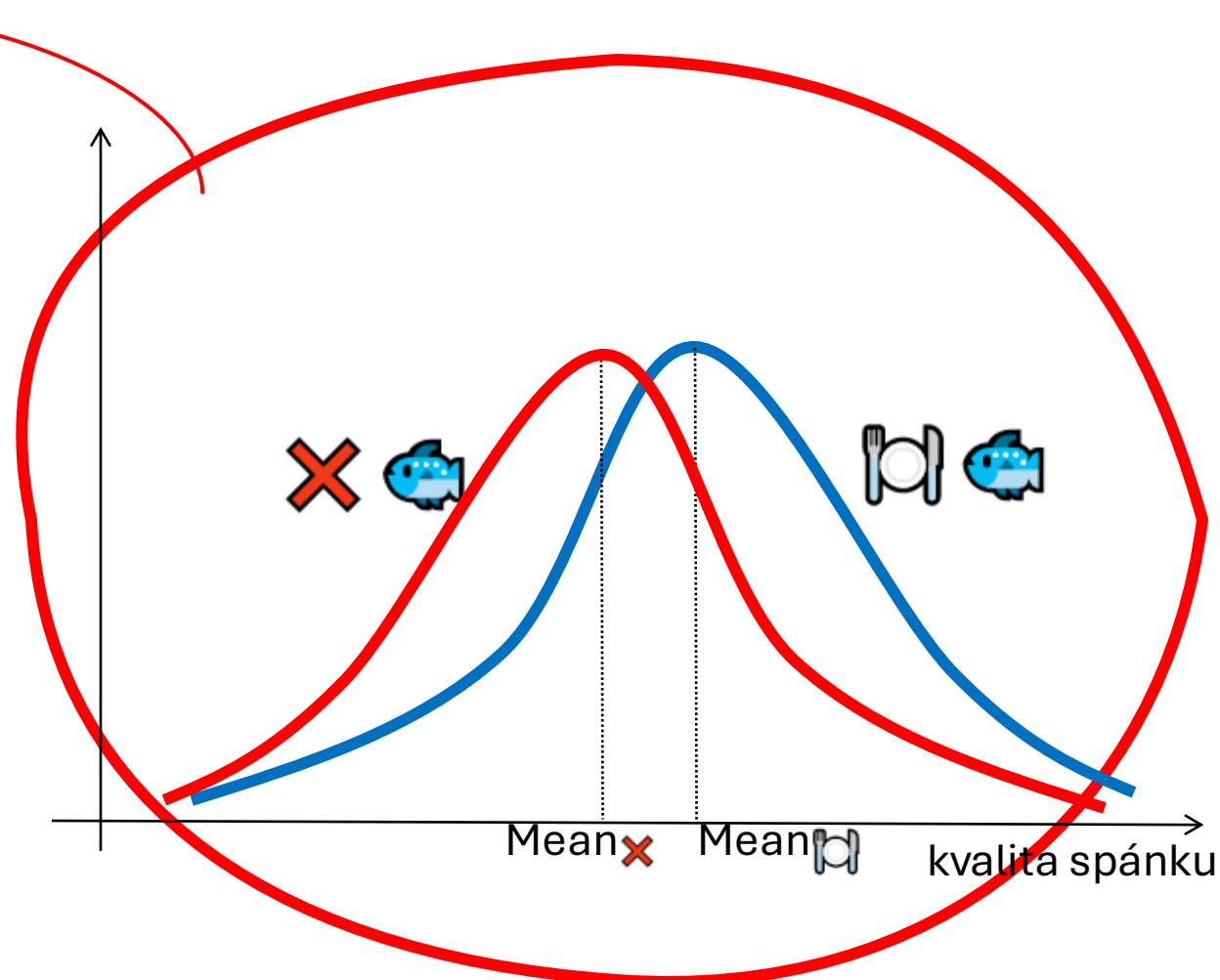


Sample



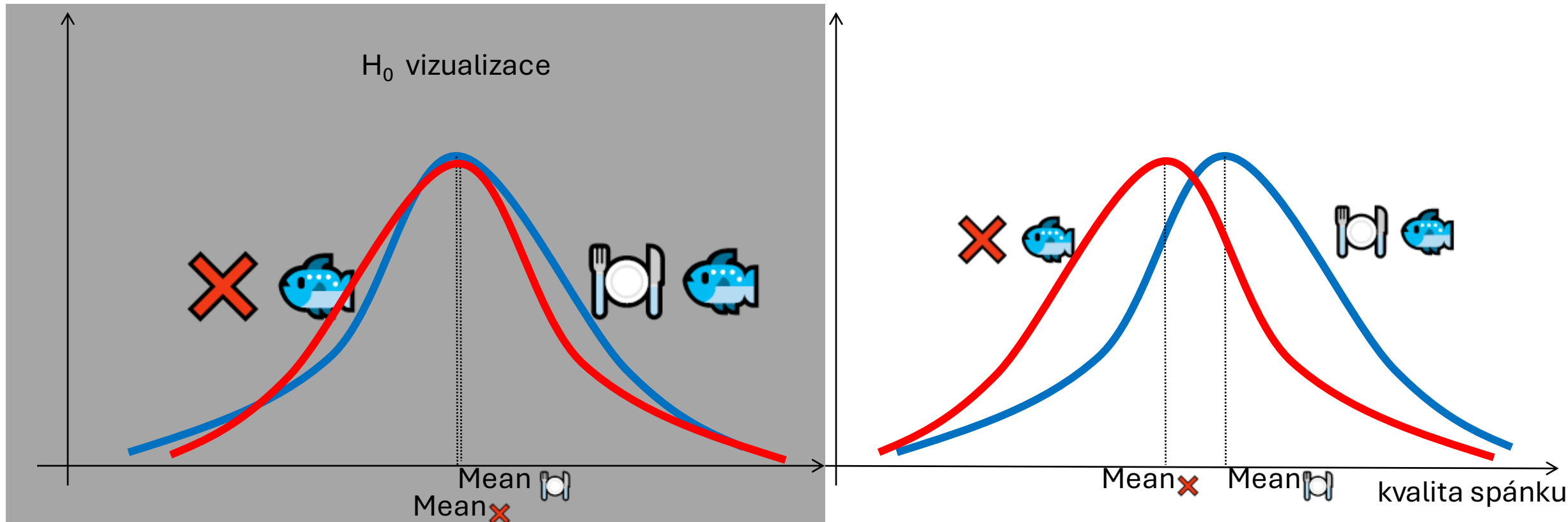
ID	Eats fish?	Sleep quality
001	yes	78
002	yes	66
003	no	69
004	no	50
005	no	95
006	yes	99
007	yes	45
...	...	...
999	no	77

Sample



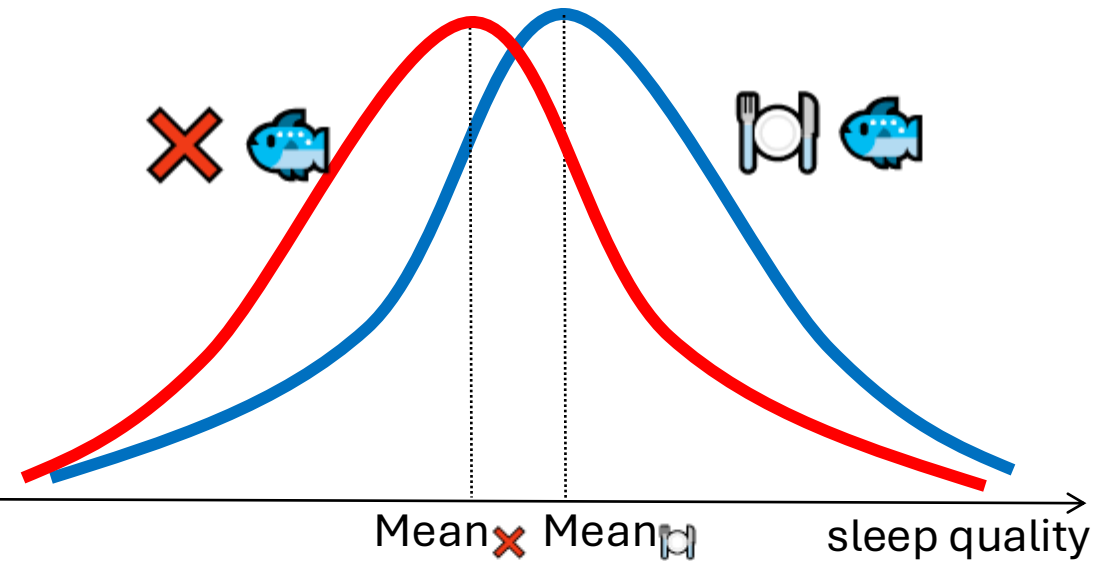
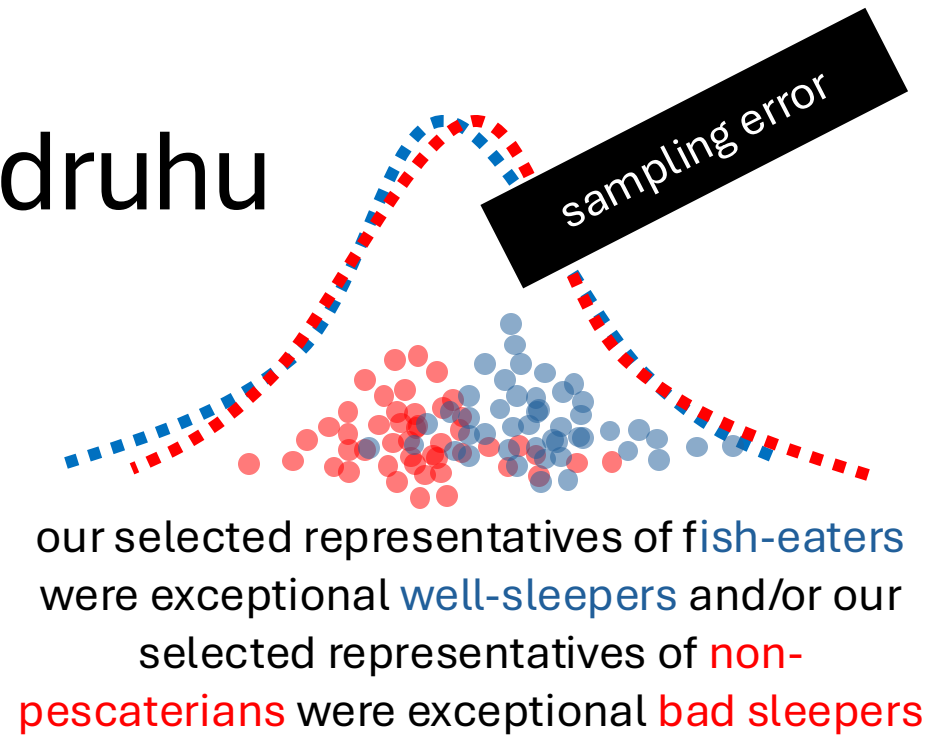
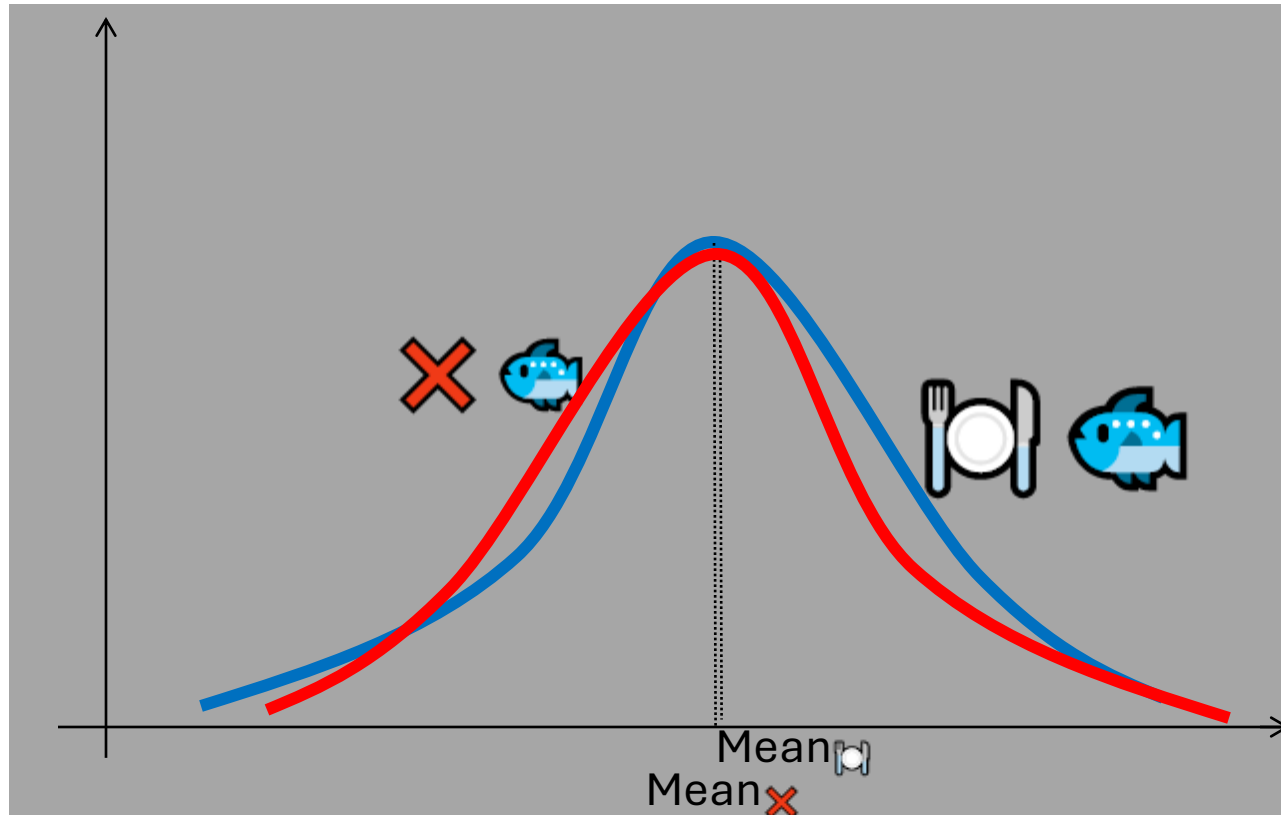
Alpha error / Chyba prvního druhu

- -> zamítneme nulovou hypotézu, přestože bychom neměli



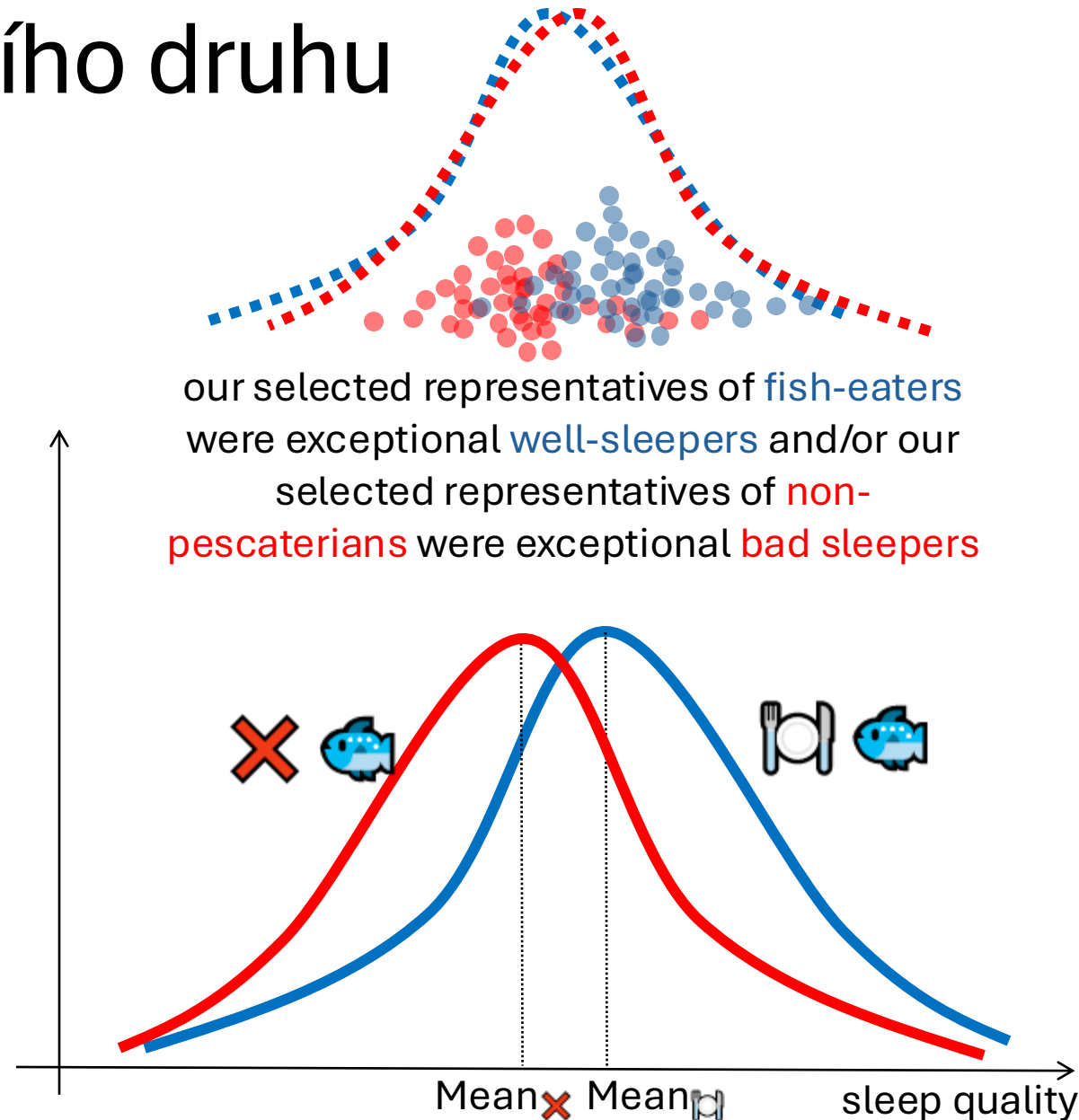
Alpha error / Chyba prvního druhu

- -> zamítneme nulovou hypotézu, přestože není nepravdivá



Alpha error / Chyba prvního druhu

- -> zamítneme nulovou hypotézu, přestože není nepravdivá
- Pravděpodobnost chyby prvního druhu / Alpha error probability
 - klesá s klesající pravděpodobností výběrové chyby (sample size and representativeness of target population) a celkovou kvalitou studie
- α err prob nižší než 5% / 1% / 0.1% se považuje za přijatelnou pro zamítnutí nulové hypotézy

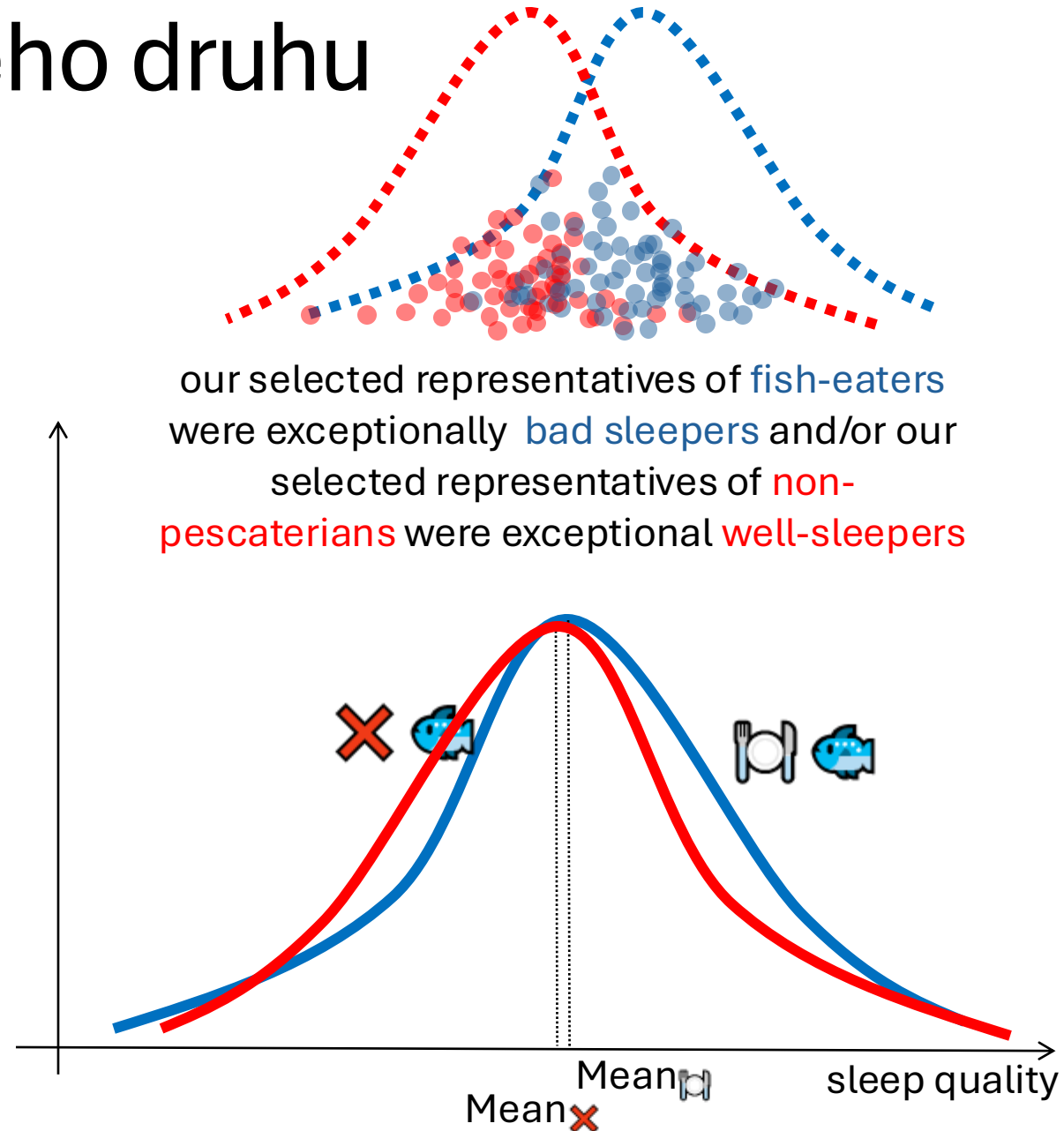


p value / p hodnota a statistická významnost

- pravděpodobnost získání takových dat, jaká byla získána, ***při platnosti nulové hypotézy***
 - $\cong$ pravděpodobnost, že jsme získali data naznačující neplatnost nulové hypotézy ***omylem*** (*sampling error and other potential errors*)
- pokud je p dostatečně nízké, můžeme z toho vyvodit závěr (conclude), že nulová hypotéza je neplatná a tedy naše alternativní hypotéza je statisticky významně podpořena (supported)
- tři hladiny statistické významnosti $\rightarrow$ určují, kdy můžeme zamítnout nulovou hypotézu – váží se k α err probability:
 - $p < .05$ (probability lower than 5%)
 - $p < .01$ (probability lower than 1%)
 - $p < .001$ (probability lower than 0.1%)

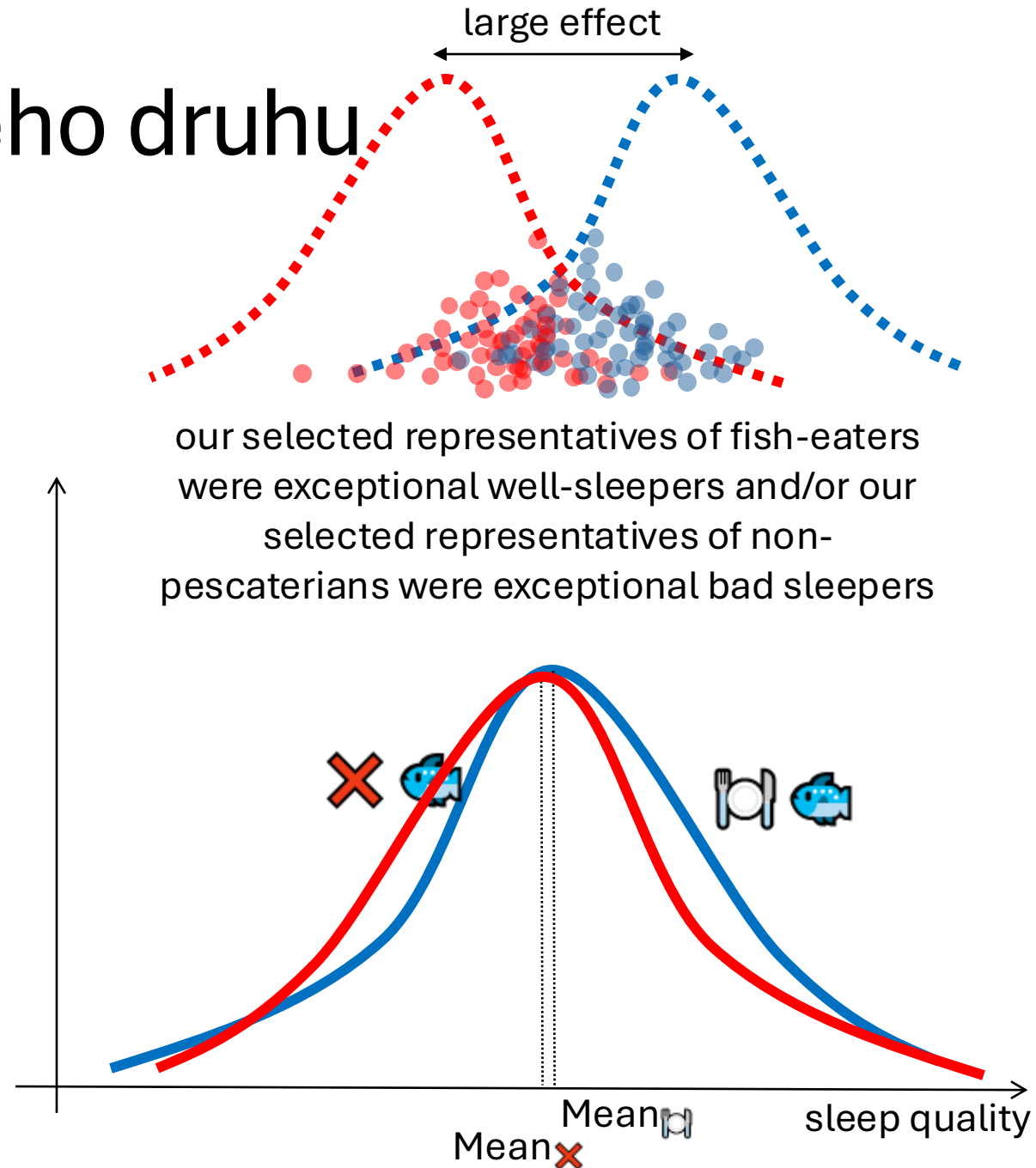
Beta error / Chyba druhého druhu

- -> nezamítneme nulovou hypotézu, přestože je nepravdivá
- Pravděpodobnost chyby druhého druhu / Beta error probability
 - klesá s klesající pravděpodobností výběrové chyby (sample size and representativeness of target population)



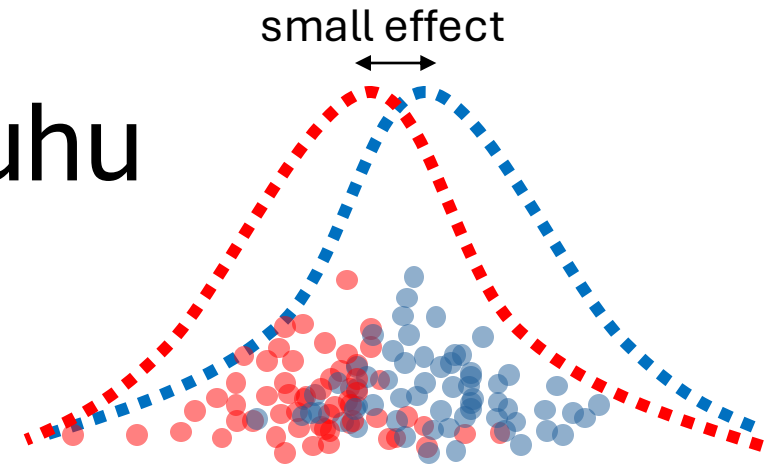
Beta error / Chyba druhého druhu

- -> nezamítneme nulovou hypotézu, přestože je nepravdivá
- Pravděpodobnost chyby druhého druhu / Beta error probability
 - klesá s klesající pravděpodobností výběrové chyby (sample size and representativeness of target population)
 - klesá s rostoucí velikostí efektu

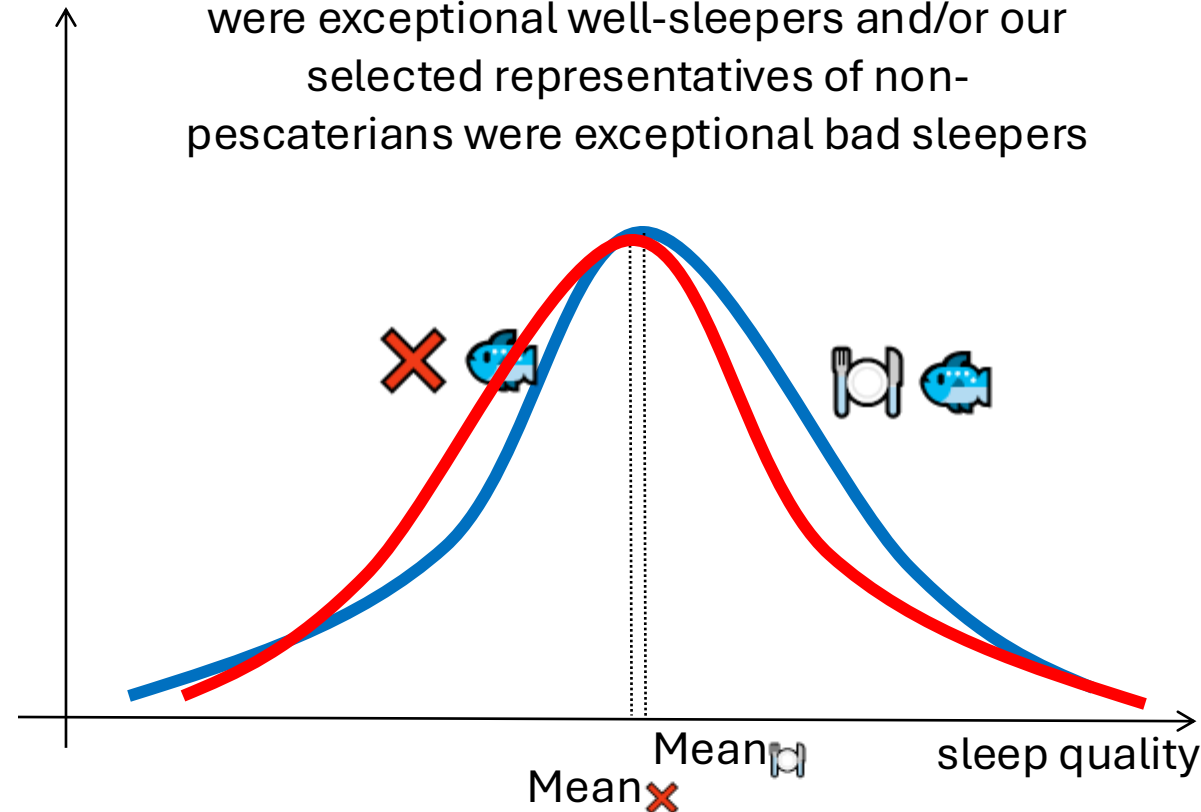


Beta error / Chyba druhého druhu

- -> nezamítneme nulovou hypotézu, přestože je nepravdivá
- Pravděpodobnost chyby druhého druhu / Beta error probability
 - klesá s klesající pravděpodobností výběrové chyby (sample size and representativeness of target population)
 - klesá s rostoucí velikostí efektu
- β err prob nižší než 5% se obvykle považuje za přijatelnou...->



our selected representatives of fish-eaters were exceptional well-sleepers and/or our selected representatives of non-pescaterians were exceptional bad sleepers



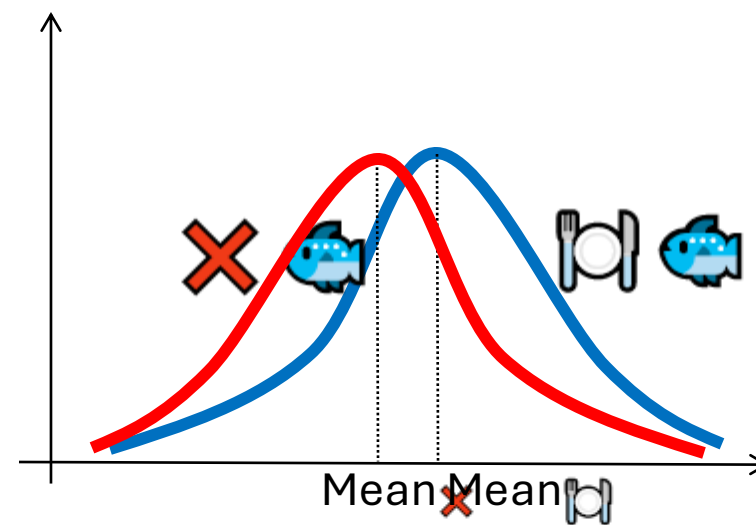
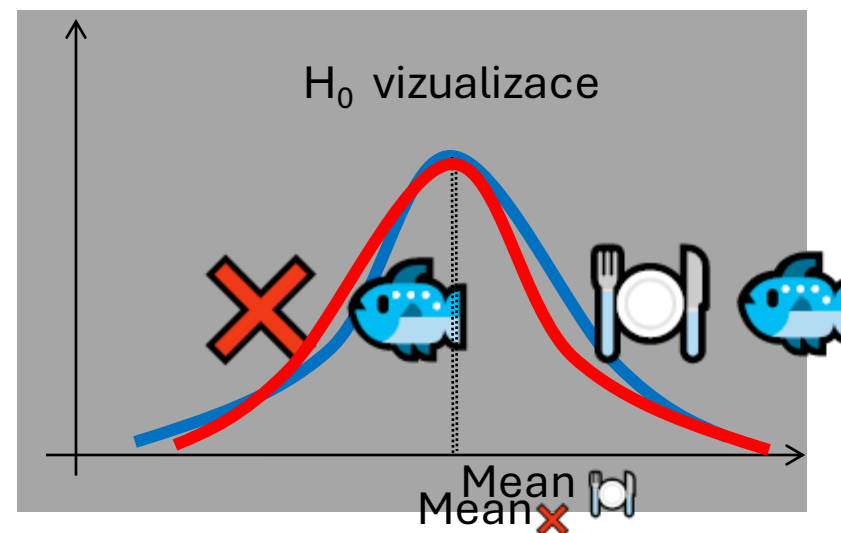
Statistical power / Statistická síla

- Power = 1 minus β err prob
- i.e., β err prob = 5% ... Power = 95%
- máme-li vysokou sílu, máme malé riziko, že nedokážeme nulovou hypotézu zamítnout ***z jiného důvodu než, že není nepravdivá*** (např. sampling error)
- Závisí na
 - naši ochotě akceptovat vyšší pravděpodobnost chyby alpha (to není cesta!)
 - na reálné velikosti efektu (to nemáme pod kontrolou...)
 - na velikosti a kvalitě výběrového souboru (to někdy pod kontrolou máme 😊) – larger sample = lower sampling error = lower probability of β error = větší síla
- G*Power software – nástroj pro výpočet optimální velikosti souboru, aby výzkumníci neprodukovali *under-powered studies*
 - dá se použít taky pro výpočet síly (β err prob) již proběhlých studií

Typy statistických testů

Důležitý úvod

- každý test má svůj předpoklad, jak by data vypadala při platnosti nulové hypotézy
- měří se, jak moc se reálná data odlišují od tohoto předpokladu



Testy založené na porovnávání středních hodnot (průměrů, mediánů)

- t-test

- ANOVA

Testy založené na porovnávání středních hodnot (průměrů, mediánů)

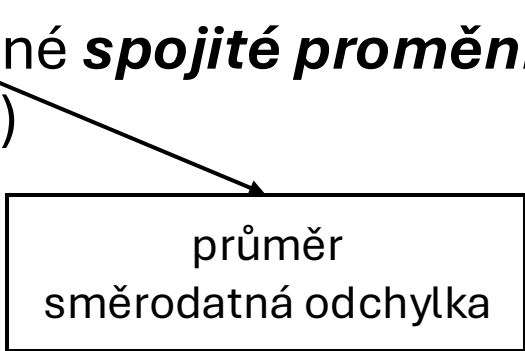
- t-test
 - Student/Welch: porovnáváme distribuce hodnot jedné **spojité proměnné** mezi **dvěma skupinami** (tj. binominální proměnnou)
- ANOVA

Testy založené na porovnávání středních hodnot (průměrů, mediánů)

- t-test

- Student/Welch: porovnáváme distribuce hodnot jedné ***spojité proměnné*** mezi ***dvěma skupinami*** (tj. binominální proměnnou)

- Student/Welch jsou tzv. parametrické
(předpokládají normální rozložení)



průměr
směrodatná odchylka

- ANOVA

Testy založené na porovnávání středních hodnot (průměrů, mediánů)

- t-test

- Student/Welch: porovnáváme distribuce hodnot jedné **spojité proměnné** mezi **dvěma skupinami** (tj. binominální proměnnou)

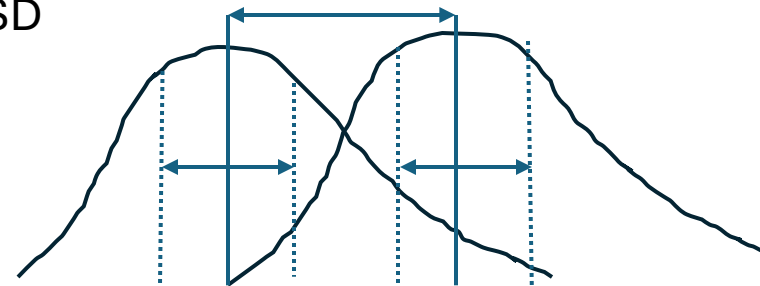
- Student/Welch jsou tzv. parametrické
(předpokládají normální rozložení)

- $t(df)$, p , Cohenovo d

poměr mezi rozdílem mezi
průměry a průměrnou SD

průměr
směrodatná odchylka (SD)

- ANOVA



Testy založené na porovnávání středních hodnot (průměrů, mediánů)

- t-test

- Student/Welch: porovnáváme distribuce hodnot jedné **spojité proměnné** mezi **dvěma skupinami** (tj. binominální proměnnou)

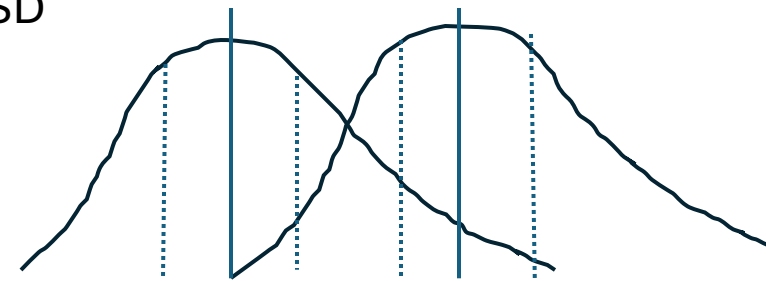
- Student/Welch jsou tzv. parametrické
(předpokládají normální rozložení)

- $t(df)$, p , Cohenovo d

poměr mezi rozdílem mezi
průměry a průměrnou SD

průměr
směrodatná odchylka (SD)

- ANOVA



Cohenovo $d \cong 2$

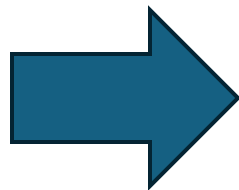
Testy založené na porovnávání středních hodnot (průměrů, mediánů)

- t-test
 - Student/Welch: porovnáváme distribuce hodnot jedné **spojité proměnné** mezi **dvěma skupinami** (tj. binominální proměnnou)
 - Student/Welch jsou tzv. parametrické
(předpokládají normální rozložení)
 - $t(df)$, p , Cohenovo d
- ANOVA
 - porovnáváme distribuce hodnot jedné **spojité proměnné** mezi **více skupinami** (One-way ANOVA, 2x2 ANOVA...)
 - testujeme přítomnost celkového (omnibus) rozdílu (F , p , η^2/ω^2) a dílčí rozdíly mezi skupinami (post-hoc testy s t , korigovanou p hodnotou, Cohenovým d)

Testy založené na porovnávání četností

- hodí se pro případy, kdy ***obě proměnné zájmu jsou diskrétní*** (nominální) (např. jí/nejí ryby + má/nemá depresi)
- χ^2 test asociace

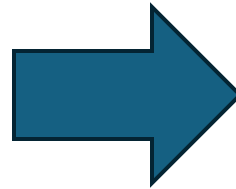
ID	ryby	deprese
01	ano	ano
02	ano	ne
03	ne	ano
04	ne	ano
05	ne	ne
06	ano	ne
07	ne	ne
08	ano	ano
09	ano	ne



Testy založené na porovnávání četností

- hodí se pro případy, kdy ***obě proměnné zájmu jsou diskrétní*** (nominální) (např. jí/nejí ryby + má/nemá depresi)
- χ^2 test asociace – **kontingenční tabulka**

ID	ryby	deprese
01	ano	ano
02	ano	ne
03	ne	ano
04	ne	ano
05	ne	ne
06	ano	ne
07	ne	ne
08	ano	ano
09	ano	ne

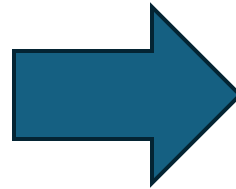


	Ryby - ano	Ryby - ne	
Deprese ano			50
Deprese ne			150
	100	100	200

Testy založené na porovnávání četností

- hodí se pro případy, kdy ***obě proměnné zájmu jsou diskrétní*** (nominální) (např. jí/nejí ryby + má/nemá depresi)
- χ^2 test asociace – **kontingenční tabulka**

ID	ryby	deprese
01	ano	ano
02	ano	ne
03	ne	ano
04	ne	ano
05	ne	ne
06	ano	ne
07	ne	ne
08	ano	ano
09	ano	ne



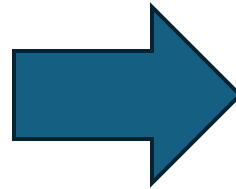
	Ryby - ano	Ryby - ne	
Deprese ano			50
Deprese ne			150
	100	100	200

přepoklad nulové hypotézy?

Testy založené na porovnávání četností

- hodí se pro případy, kdy ***obě proměnné zájmu jsou diskrétní*** (nominální) (např. jí/nejí ryby + má/nemá depresi)
- χ^2 test asociace – **kontingenční tabulka**

ID	ryby	deprese
01	ano	ano
02	ano	ne
03	ne	ano
04	ne	ano
05	ne	ne
06	ano	ne
07	ne	ne
08	ano	ano
09	ano	ne



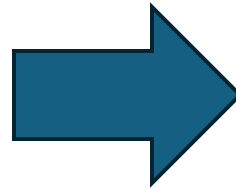
	Ryby - ano	Ryby - ne	
Deprese ano	25	25	50
Deprese ne	75	75	150
	100	100	200

přepoklad nulové
hypotézy?

Testy založené na porovnávání četností

- hodí se pro případy, kdy ***obě proměnné zájmu jsou diskrétní*** (nominální) (např. jí/nejí ryby + má/nemá depresi)
- χ^2 test asociace – **kontingenční tabulka**

ID	ryby	deprese
01	ano	ano
02	ano	ne
03	ne	ano
04	ne	ano
05	ne	ne
06	ano	ne
07	ne	ne
08	ano	ano
09	ano	ne



	Ryby - ano	Ryby - ne	
Deprese ano	25 25%	25 25%	50
Deprese ne	75	75	150
	100	100	200

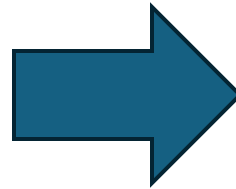
přepoklad nulové
hypotézy?

Testy založené na porovnávání četností

- hodí se pro případy, kdy ***obě proměnné zájmu jsou diskrétní*** (nominální) (např. jí/nejí ryby + má/nemá depresi)
- χ^2 test asociace – **kontingenční tabulka**

přepoklad nulové
hypotézy versus
skutečné četnosti

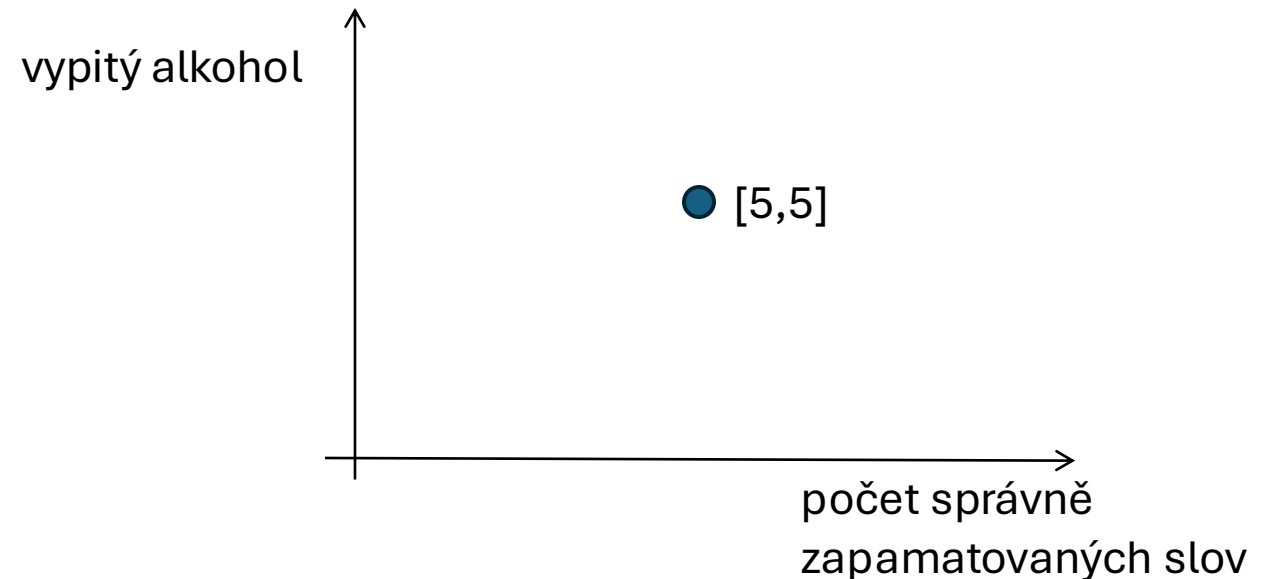
ID	ryby	deprese
01	ano	ano
02	ano	ne
03	ne	ano
04	ne	ano
05	ne	ne
06	ano	ne
07	ne	ne
08	ano	ano
09	ano	ne



	Ryby - ano	Ryby - ne	
Deprese ano	10 25%	40 25%	50
Deprese ne	90	60	150
	100	100	200

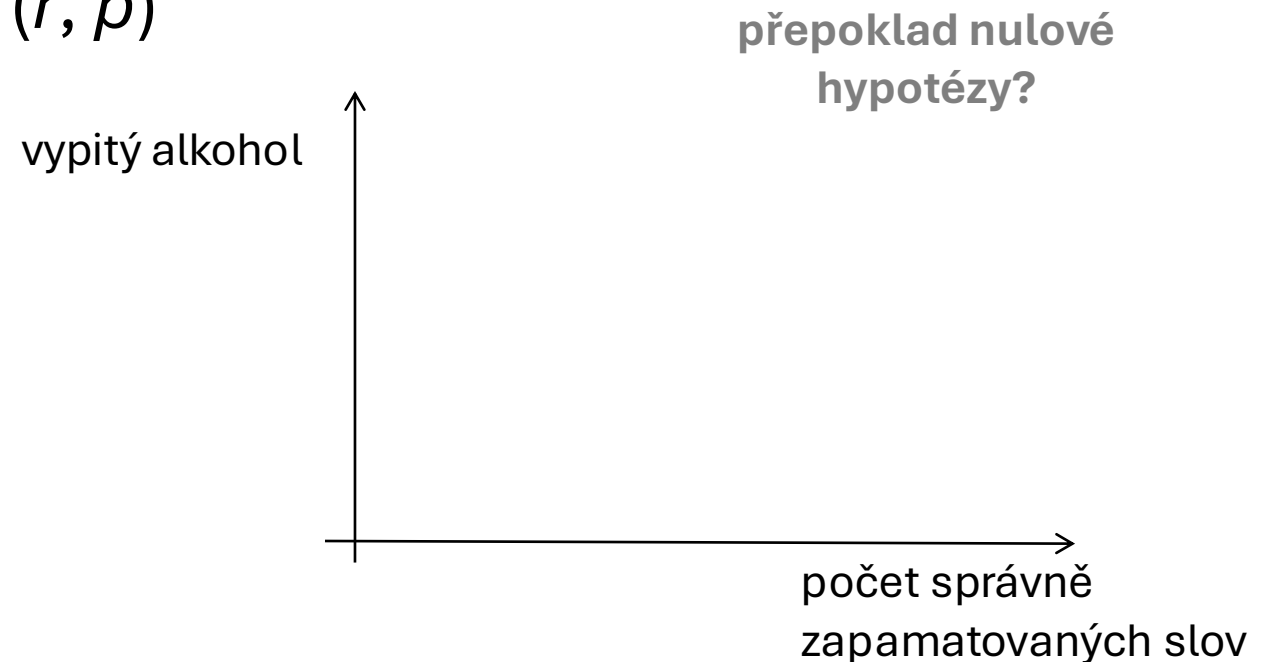
Korelační test a lineární regrese

- korelační test se používá se v případě zjišťování **(lineární) souvislosti mezi dvěma spojitými proměnnými** (např. množství vypitého alkoholu a kvalita paměťového výkonu)
- Pearsonův korelační koeficient (r , p)



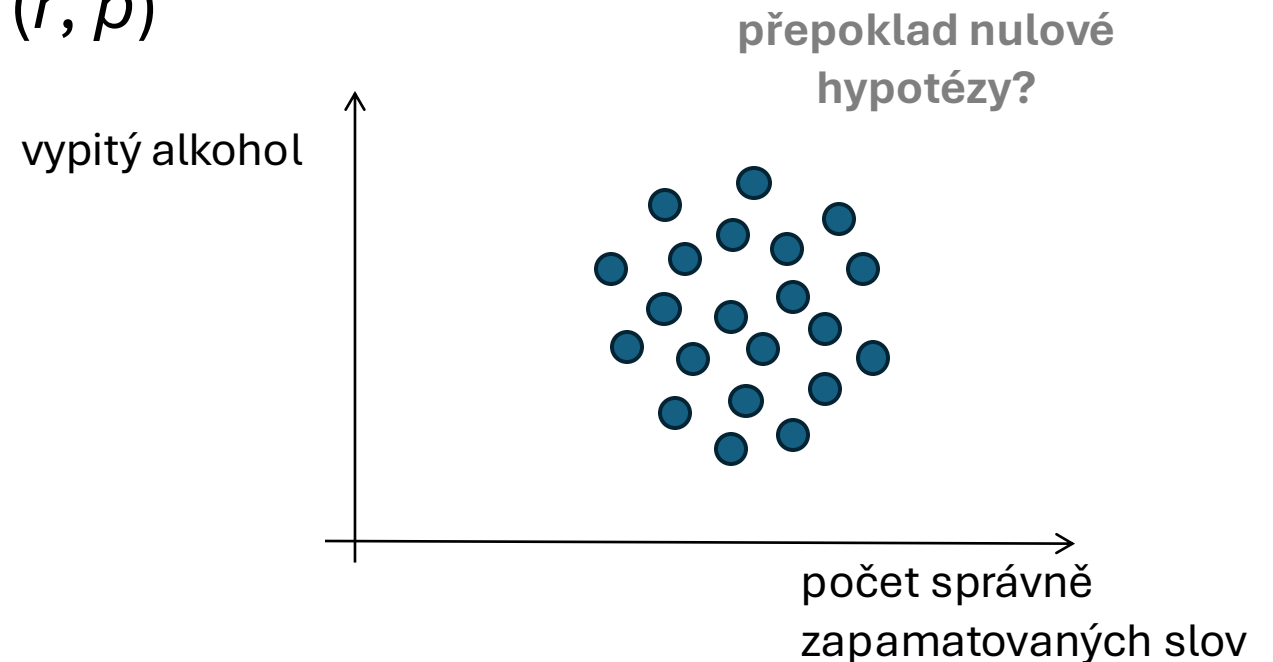
Korelační test a lineární regrese

- korelační test se používá se v případě zjišťování **(lineární) souvislosti mezi dvěma spojitými proměnnými** (např. množství vypitého alkoholu a kvalita paměťového výkonu)
- Pearsonův korelační koeficient (r , p)



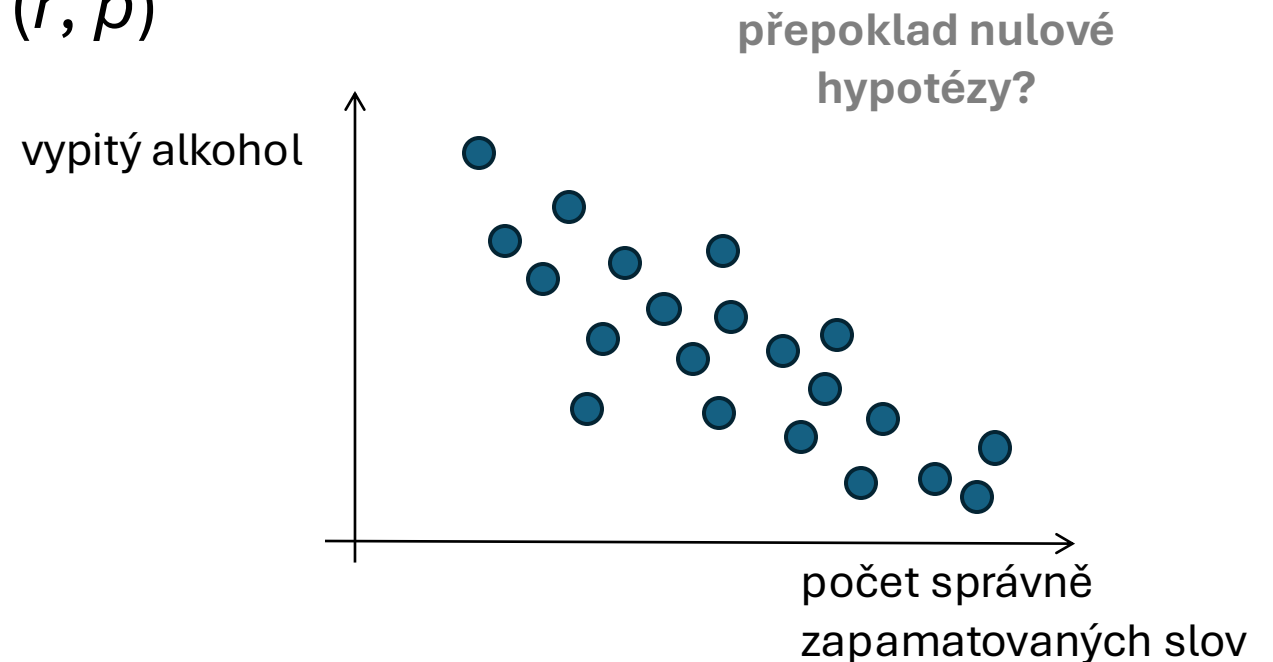
Korelační test a lineární regrese

- korelační test se používá se v případě zjišťování **(lineární) souvislosti mezi dvěma spojitými proměnnými** (např. množství vypitého alkoholu a kvalita paměťového výkonu)
- Pearsonův korelační koeficient (r , p)



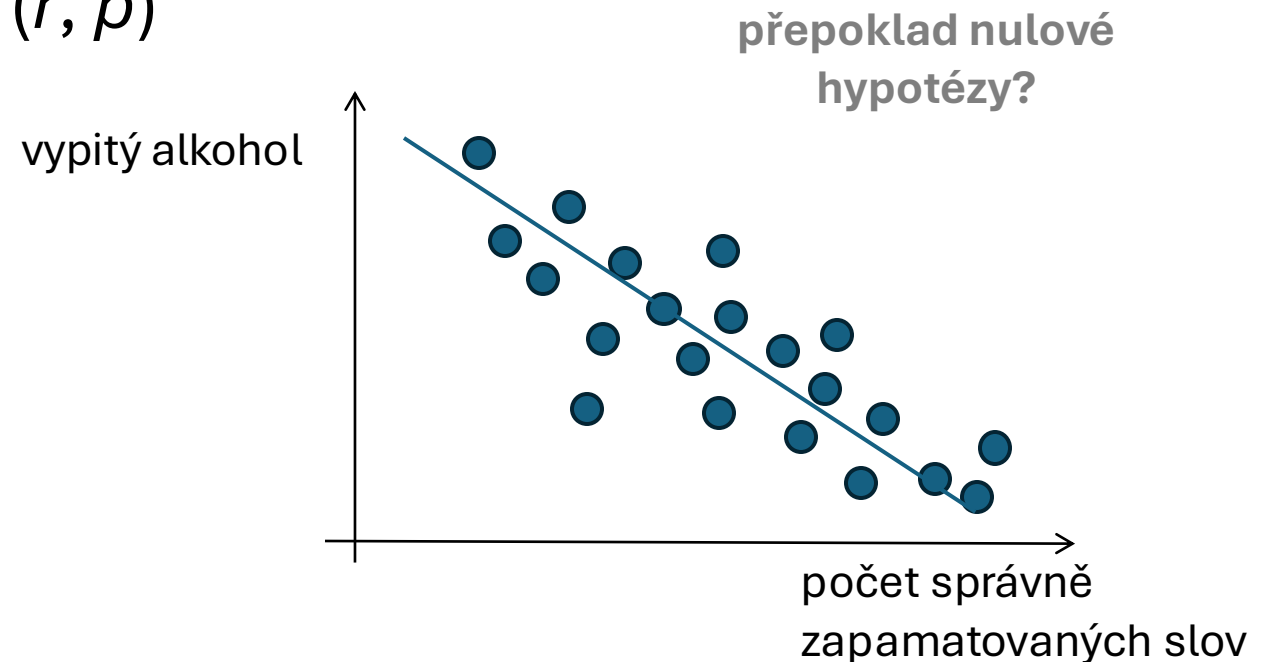
Korelační test a lineární regrese

- korelační test se používá se v případě zjišťování **(lineární) souvislosti mezi dvěma spojitými proměnnými** (např. množství vypitého alkoholu a kvalita paměťového výkonu)
- Pearsonův korelační koeficient (r , p)



Korelační test a lineární regrese

- korelační test se používá se v případě zjišťování **(lineární) souvislosti mezi dvěma spojitými proměnnými** (např. množství vypitého alkoholu a kvalita paměťového výkonu)
- Pearsonův korelační koeficient (r , p)



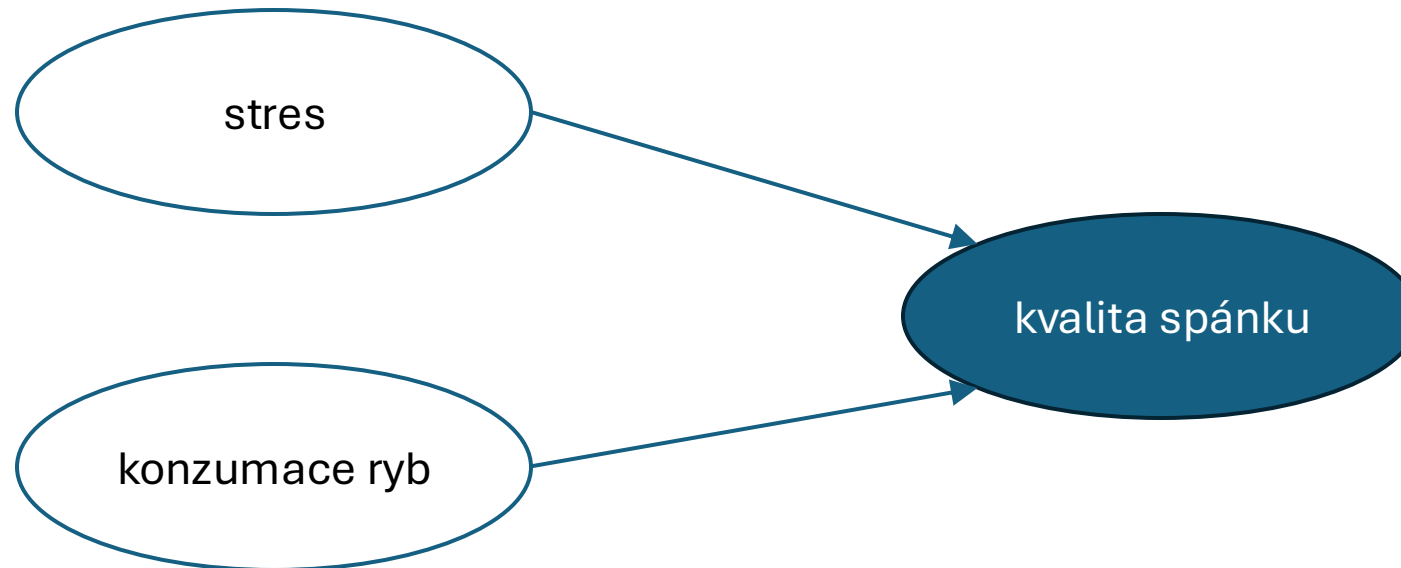
Korelační test a lineární regrese

Pokud je závisle
proměnná diskrétní →
logistická regrese

- korelační test se používá se v případě zjišťování **(lineární) souvislosti mezi dvěma spojitými proměnnými** (např. množství vypitého alkoholu a kvalita paměťového výkonu)
 - Pearsonův korelační koeficient (r, p)
- lineární regrese umožňuje testovat **efekt více proměnných (prediktorů) na spojitou závisle proměnnou (outcome)**
 - Jak se zlepší můj odhad závisle proměnné, když budu znát hodnotu prediktoru?
 - Regresní koeficient (β, p) pro každý prediktor (je statisticky významný?)
 - Koeficient determinace (R^2) – explanační síla celého modelu

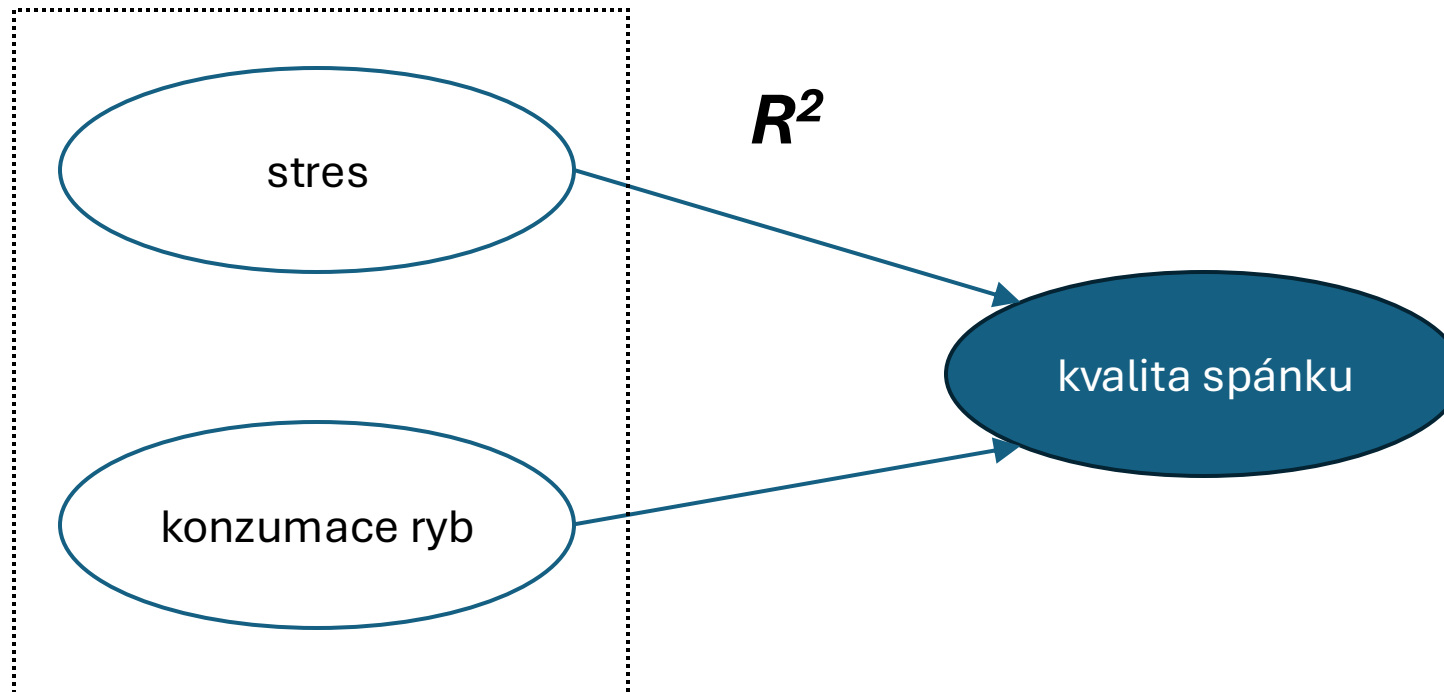
Lineární regrese – příklad

- závisle proměnná (musí být spojitá): kvalita spánku
- prediktory (spojité nebo diskrétní proměnné): stres, konzumace ryb



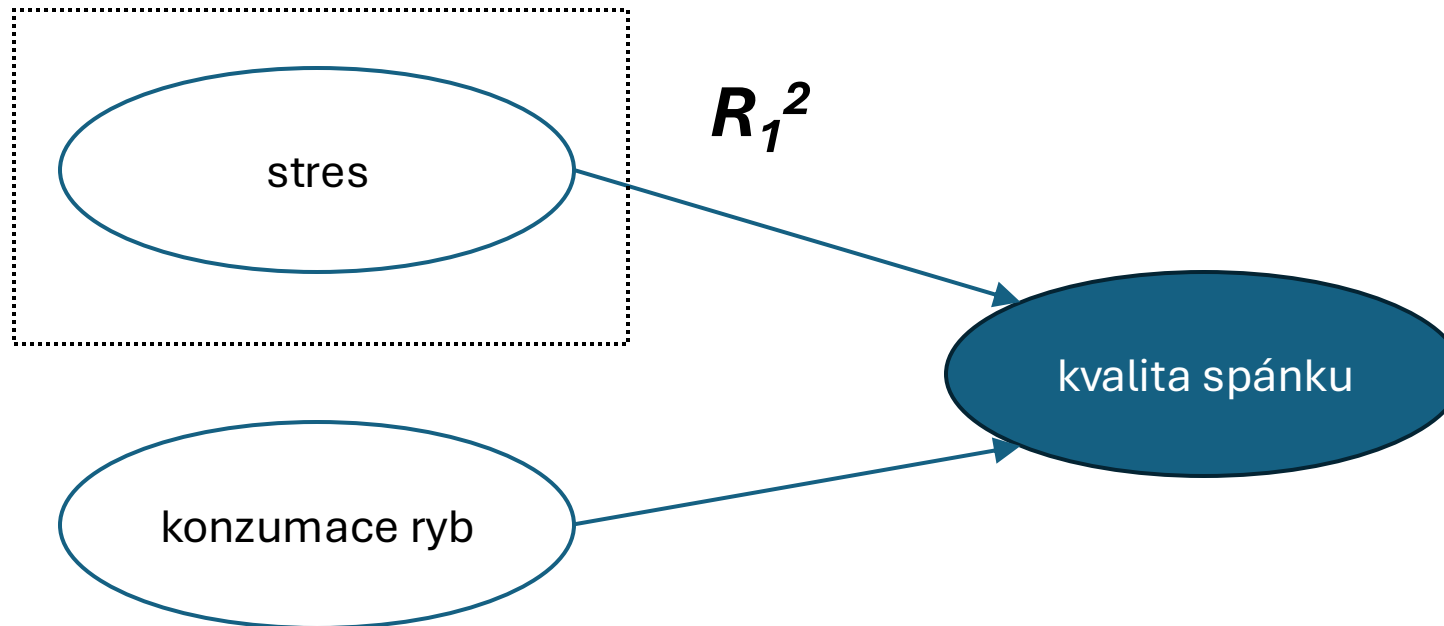
Lineární regrese – příklad

- závisle proměnná (musí být spojitá): kvalita spánku
- prediktory (spojité nebo diskrétní proměnné): stres, konzumace ryb



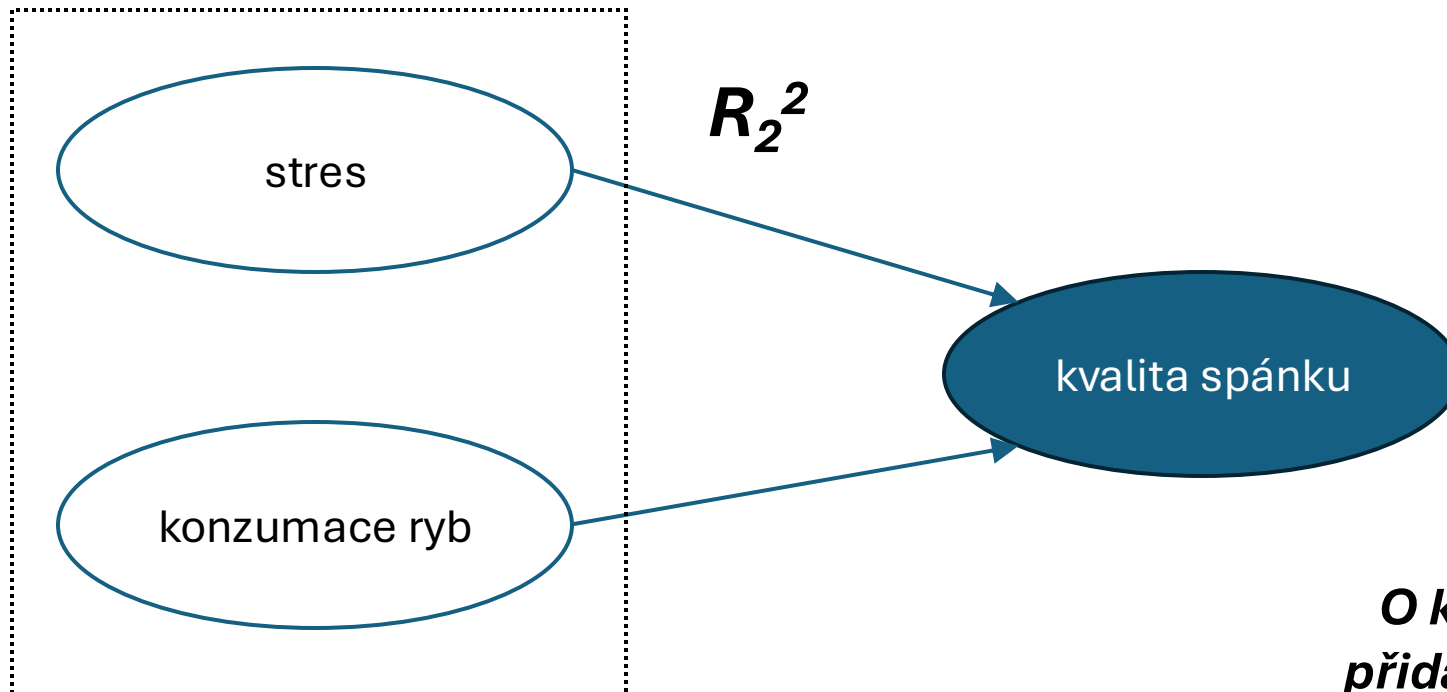
Lineární regrese – příklad

- závisle proměnná (musí být spojitá): kvalita spánku
- prediktory (spojité nebo diskrétní proměnné): stres, konzumace ryb



Lineární regrese – příklad

- závisle proměnná (musí být spojitá): kvalita spánku
- prediktory (spojité nebo diskrétní proměnné): stres, konzumace ryb



*O kolik se zvýší R^2 když
přidám další prediktor(y)?*

104 minut



Predictor	Estimate	SE	95% Confidence Interval		t	p
			Lower	Upper		
Intercept ^a	103.80	0.65	82.9074	124.7	9.746	< .001
Number of children's own devices ¹						
1 device	46.72	8.32	30.4057	63.0	5.617	< .001
2 devices	80.66	9.44	62.1395	99.2	8.541	< .001
3 devices	90.35	11.77	67.2662	113.4	7.676	< .001
4 or 5 devices	147.48	14.02	119.9827	175.0	10.519	< .001
Grade ² :						
2nd	4.26	6.49	-8.4682	17.0	0.657	0.511
3rd	21.26	6.98	7.5737	34.9	3.047	0.002
Child's sex ³ :						
Boy	21.27	5.42	10.6394	31.9	3.924	< .001
Having younger sibling(s) ⁴ :						
No	16.61	5.52	5.7877	27.4	3.010	0.003
Intact family ⁵ :						
No	2.35	7.16	-11.6889	16.4	0.328	0.743
Family income ⁶ :						
Less than 1200 EUR	19.00	10.87	-2.3116	40.3	1.749	0.081
1200-1799 EUR	16.24	8.27	0.0201	32.5	1.964	0.050
1800-2399 EUR	11.76	7.11	-2.1920	25.7	1.653	0.098
Responding parent's education ⁷ :						
Elementary/Practical	55.08	9.54	36.3648	73.8	5.772	< .001
High school	23.60	6.32	11.2053	36.0	3.735	< .001
Country sample:						
Czechia – at-home schooling	24.37	7.71	9.2537	39.5	3.162	0.002
Czechia	3.19	7.69	-11.8993	18.3	0.415	0.678
Finland	13.83	8.52	-2.8724	30.5	1.624	0.105

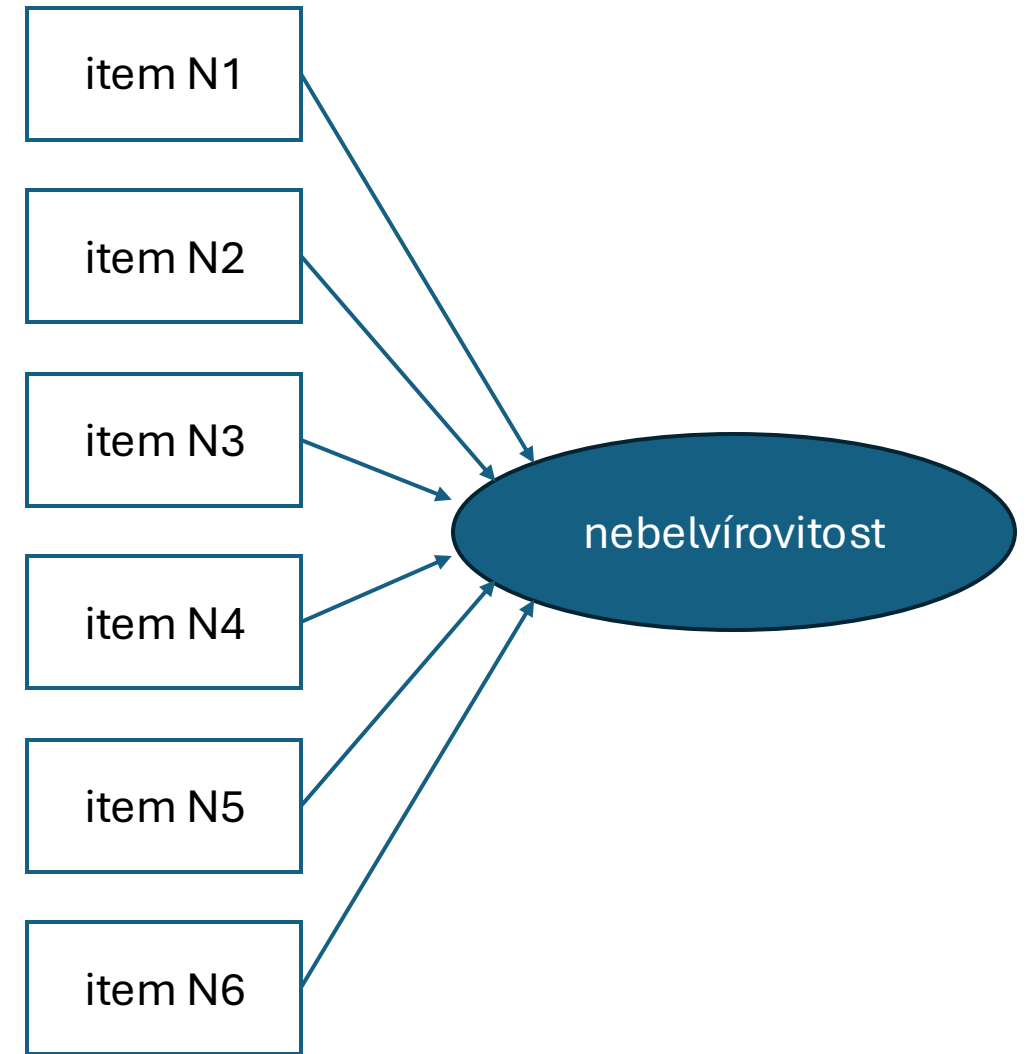
^a Represents reference level. Reference levels for predictors were: ¹0 devices, ²1st, ³Girl, ⁴Yes, ⁵Yes, ⁶More than 1200 EUR, ⁷University, ⁸Slovakia.

365 minut



Psychometrické analýzy

- Konfirmační faktorová analýza
 - regresní model, kde celkový skór je závisle proměnná a dílčí položky jsou prediktory
 - ověření faktorové struktury
 - regresní koeficient a p pro každou položku, parametry “model fit” (shoda modelu s daty)...RMSEA, CFI, TLI



Psychometrické analýzy

- Konfirmační faktorová analýza
 - regresní model, kde celkový skór je závisle proměnná a dílčí položky jsou prediktory
 - ověření faktorové struktury
 - regresní koeficient a p pro každou položku, parametry “model fit” (shoda modelu s daty)...RMSEA, CFI, TLI
- Analýza vnitřní konzistence
 - korelace mezi položkami
 - Cronbach α , McDonald ω

(Cronbach $\alpha = 0.67$)
McDonald $\omega = 0.69$

Item	Mean	SD	Item-rest correlation	(α)/ ω if dropped
N1	3.15	1.12	0.67	0.58
N2	2.89	0.75	0.32	0.64
N3	1.74	0.90	0.11	0.73
N4	3.06	1.50	0.45	0.62
N5	4.56	1.20	-0.07	0.74
N6	3.80	1.44	0.49	0.63