

Základy datové analýzy

Lukáš Hejtmánek

lukas.hejtmánek@fhs.cuni.cz



Co je cílem statistiky

Statistika v psychologii nevyžaduje nutně počítání, ale porozumění

Po analýze dat bychom měli o světě vědět více, než před tím (nebo alespoň vědět, co nevíme)

Statistika řeší dva problémy

Svět je velký

Svět je chaotický

Svět je velký

Nemůžeme nikdy analyzovat data individuálně,
potřebujeme nějak agregovat

Nemůžeme nikdy změřit všechny
Musíme změřit jen pár osob a jejich výsledky
zobecnit



Svět je chaotický

Porozumět zákonitostem v chaosu

Porozumět, co je jen náhoda a co je pravidlo

Když se zeptám dvou lidí na jednu otázku, dostanu dvě různé odpovědi

Když se zeptám jednoho člověka na tutéž otázku, dostanu dvě různé odpovědi

Na škále od jedné do 10 ti, jak se ti líbí takhle písnička





Úkoly statistiky

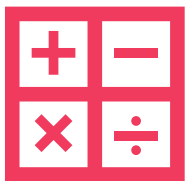
Popsat

Pochopit

Předpovědět

Cokoli nám umožní lépe popsat data než
kompletní náhodnost je považováno za dobré

Typy analýz



Deskriptivní analýza (exploratory)

Popisuje stav věcí – průměry, počty,
variabilita

Typická pro bussiness i akademické
reporty



Induktivní analýza (confirmatory)

Snaží v datech najít zákonitosti, odhalit
pravidla

Typická pro akademické reporty



Prediktivní analýza (predictive)

Jak z dat odvodit budoucnost
Není zcela nutné chápat zákonitosti, jen,
že predikce fungují

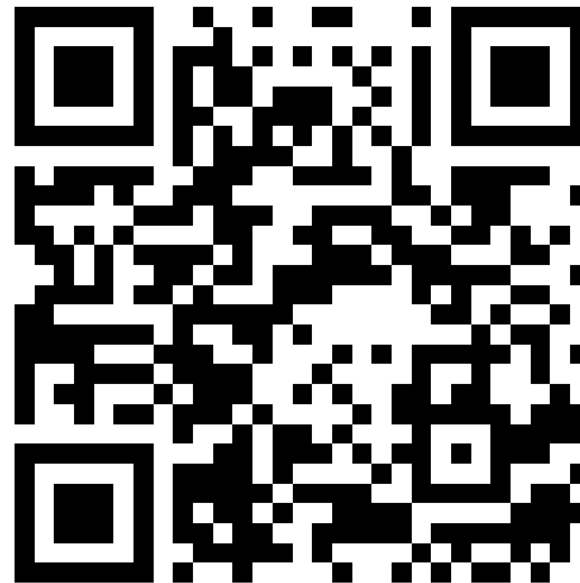
Typická pro bussiness reporty

Song demo

Vyplňte první stranu dotazníku

Poslechněte si skladbu

Vyplňte poslední stranu dotazníku.



<https://forms.gle/AZkTTgrmEvkYrnjQ6>

Typy statistické analýzy



Deskriptivní statistika

Popisuje jak věci jsou

Dává přehled o situaci

Surveys – sčítání lidu



Inferenční statistika

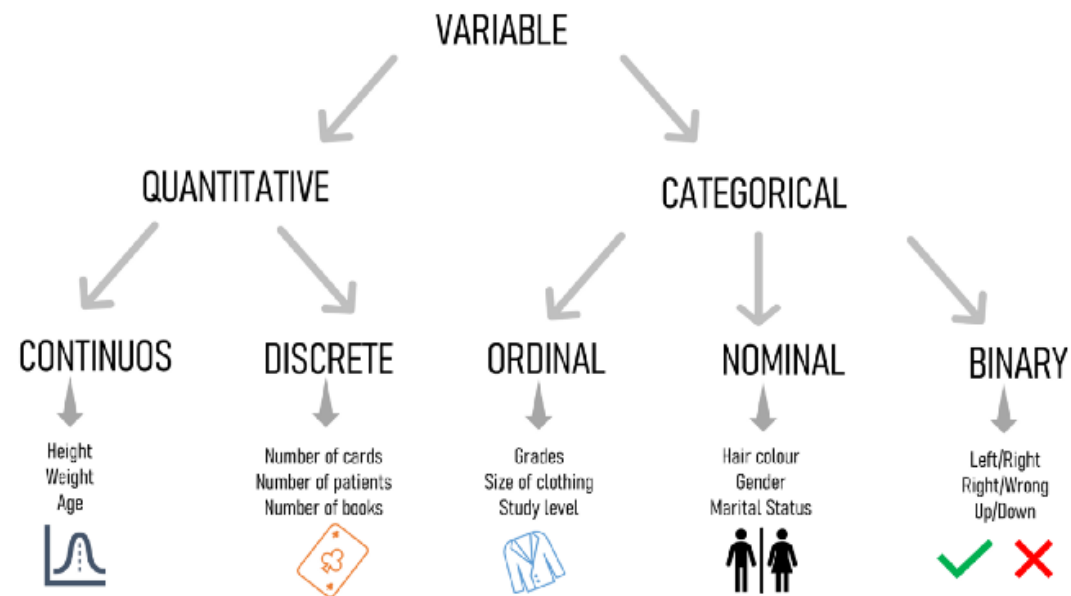
Hledá závislosti v datech, co s čím souvisí

Jejím cílem je snížit nejistotu

Popisuje zákonitosti, očekávání

Umožňuje predikce na základě dat

Typy dat



Kvalitativní/Kategorické

Ordinální

Existuje "lepší" varianta

Nominální

Všechny možnosti jsou stejně validní

Kvantitativní

Spojité X Nespojité

Parametrické x neparametrické

Parametrická vs neparametrická data

Parametrická data

- Numerické „ratio“ hodnoty, kardinální hodnoty

- Rozdíl mezi 1 – 2 je stejný jako 4 – 5

- Počet správně zodpovězených otázek v testu

Neparametrická data

- Ordinální hodnoty, pořadí

- Víme, že 2 je více než 1 a 5 více než 4, ale ne o kolik

- Známka z předmětu

Statistické testy jsou vhodné pouze na některé typy dat

- Pearsonova korelace vs Spearmanova korelace

- Studentův T-test vs Mann-Whitney Test

Věk

Tepová frekvence

Míra oblíbenosti (1-5)

Počet bodů v IQ testu (0-100)

SCIO skór

SCIO percentil

Parametrické či
neparametrické
hodnoty?

Deskriptivní statistika

Cíle deskriptivní statistiky

Popsat stav věcí

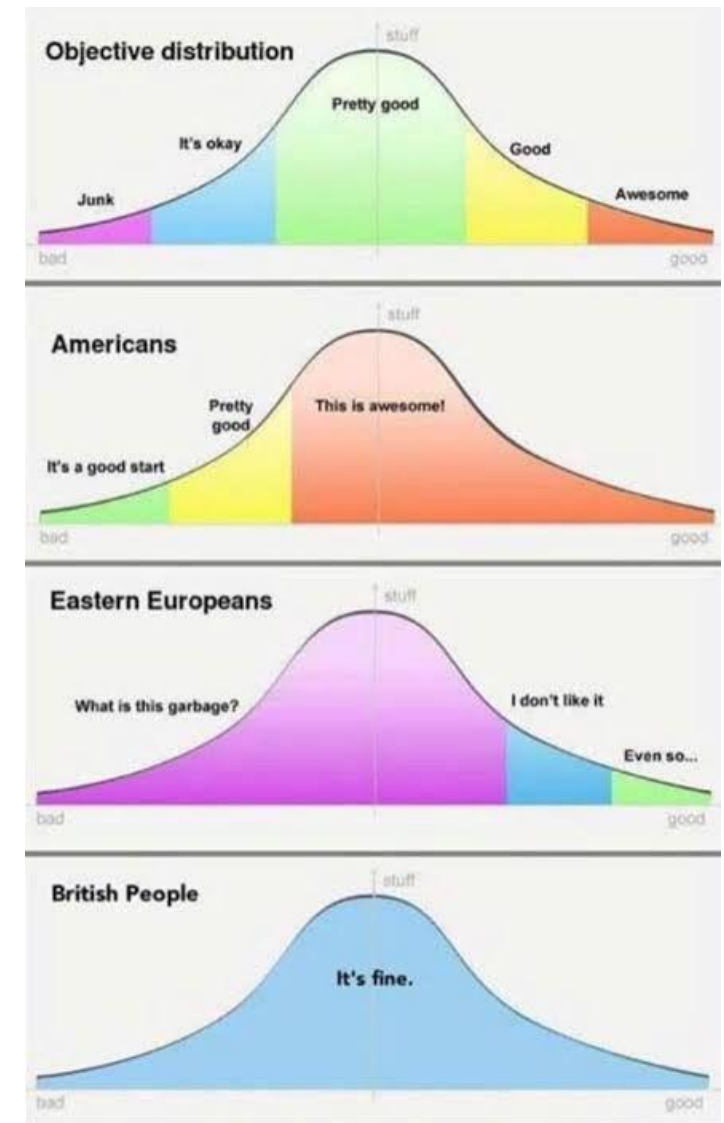
V případě dotazníků, průzkumů je deskriptivní statistika finální

Popsat vzorek tak, aby čtenář tušil, na jakých datech se pracuje

Nutné v případě induktivní statistiky

Bias = zkreslení

Je ve vzorku nějaký bias?



Frekvenční tabulky

Pro kategorická data

Frekvenční tabulky

Jak často se hodnoty vyskytují

Absolutní vs procentuální (ratio)

Pokud jsou čísla malá, preferujeme absolutní hodnoty

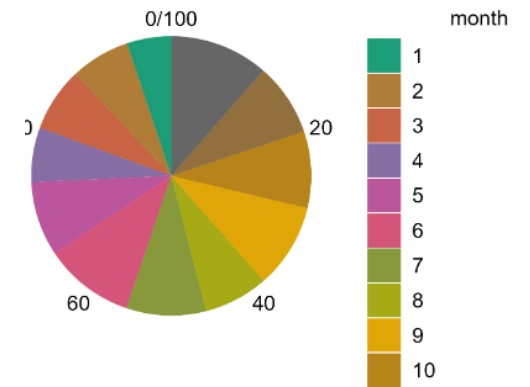
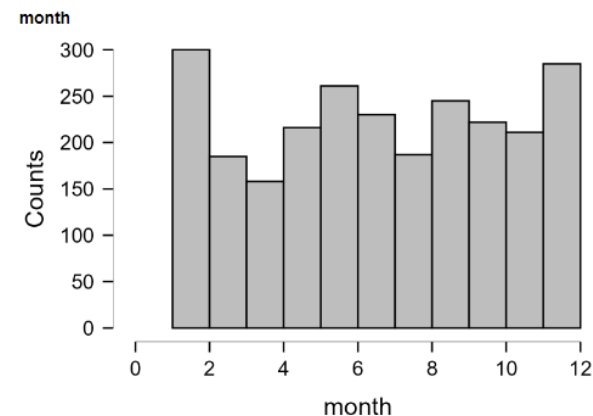
Nejlépe pak obojí

Frequency Tables ▼

Frequencies for month

month	Frequency	Percent	Valid Percent	Cumulative Percent
1	128	5.120	5.120	5.120
2	172	6.880	6.880	12.000
3	185	7.400	7.400	19.400
4	158	6.320	6.320	25.720
5	216	8.640	8.640	34.360
6	261	10.440	10.440	44.800
7	230	9.200	9.200	54.000
8	187	7.480	7.480	61.480
9	245	9.800	9.800	71.280
10	222	8.880	8.880	80.160
11	211	8.440	8.440	88.600
12	285	11.400	11.400	100.000
Missing	0	0.000		
Total	2500	100.000		

Distribution Plots

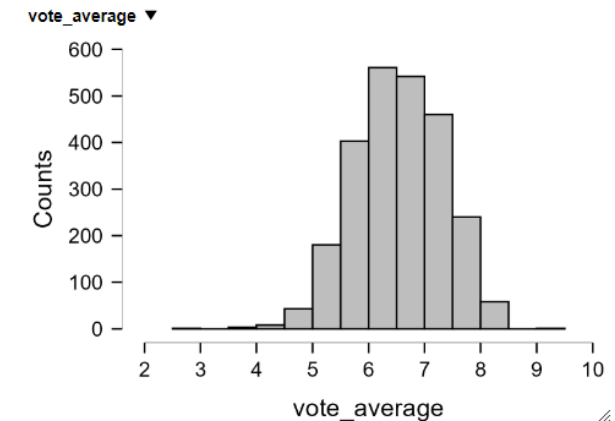
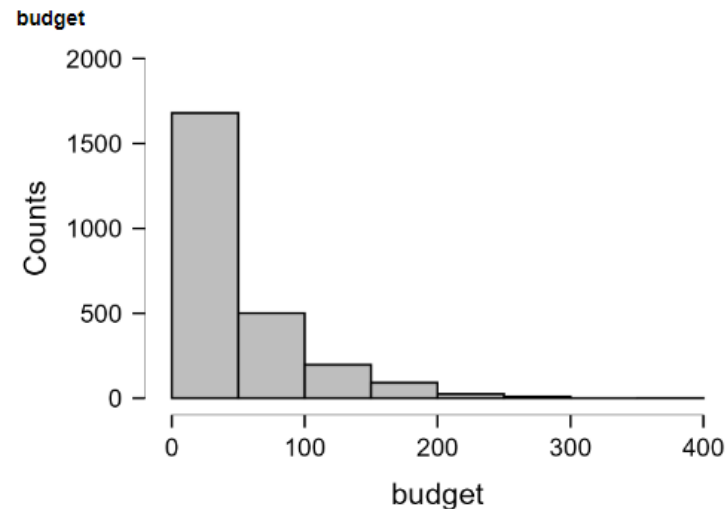
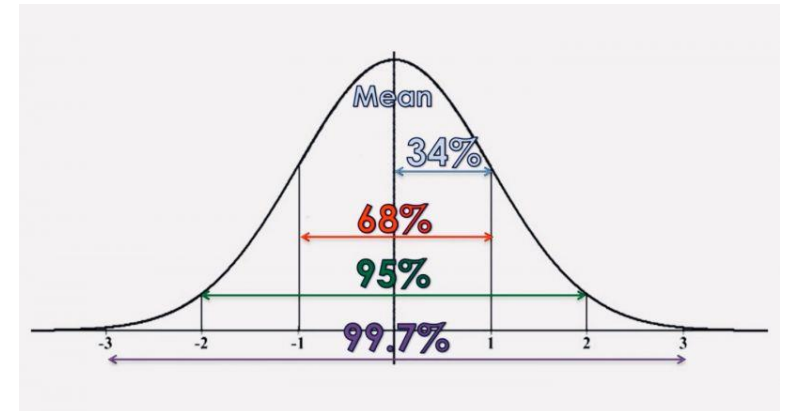
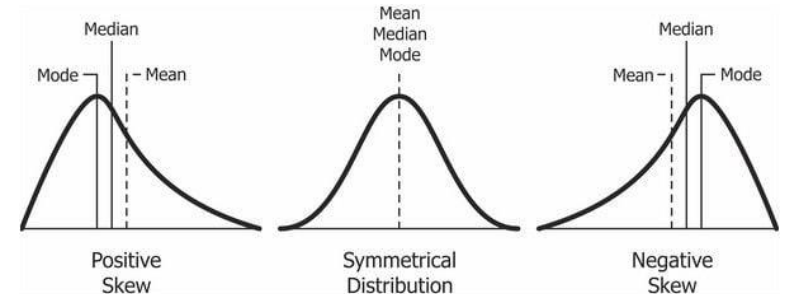


Distribuční rozdělení

Je vzorek “normální”?

Existuje bias/zkosení na jednu stranu?

Typicky reportujeme pouze u analýz, které to vyžadují
T-test, ANOVA, Korelace



Míry centrální tendence

Průměr

Aritmetický průměr

$\text{SUM}(x)/\text{COUNT}(x)$

Vhodný pro

Velké datasety

Symetrické datasety

Medián

Střední hodnota

Vhodný pro

Asymetrická data

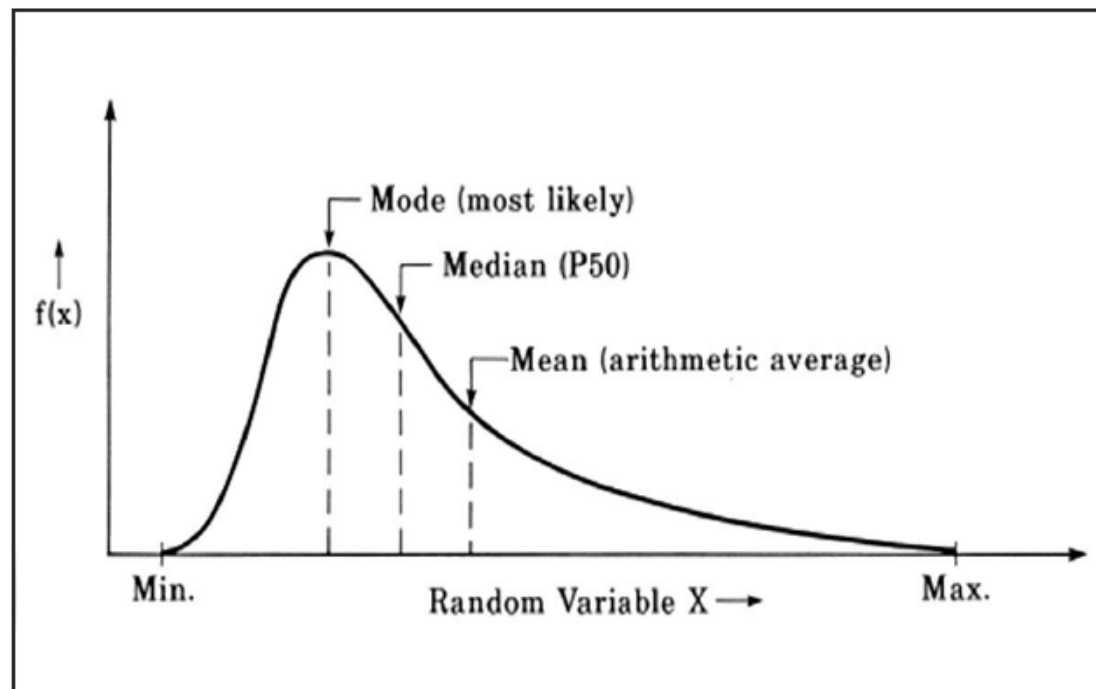
Menší datasety

(Modus)

Nejčastější hodnota

Vhodný pro

Kategorická data



Kde použít jakou míru centrální tendence?

Platové ohodnocení

Věkové rozložení

Odpovědi na škále 1-7

Souhrný skór na škále 1-100

Známka ve škole (1-5)

Váha v kilogramech

Počet hodin spánku minulou noc

Míry rozptylu

Rozptyl a směrodatná odchylka

Var = Variance

SD = standard deviation

Standardní odchylka průměru

SE = standard error

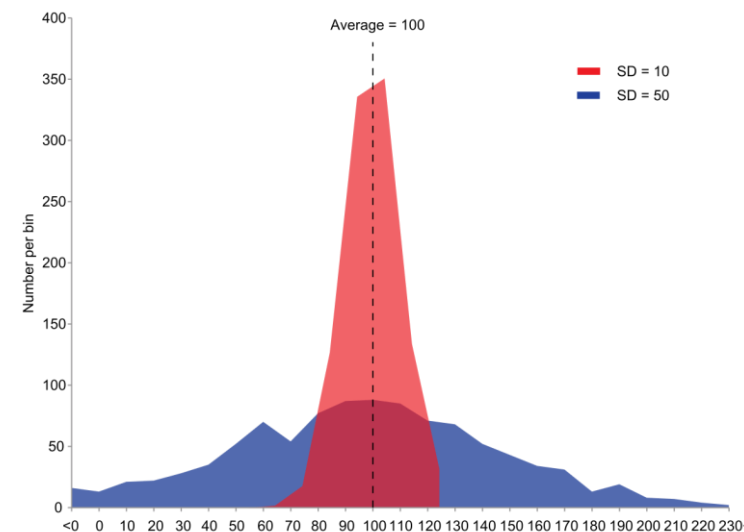
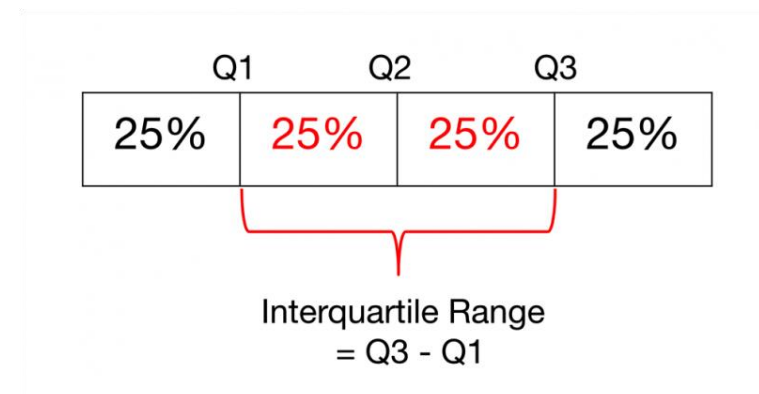
Míra popisující jak moc by se měnil průměr

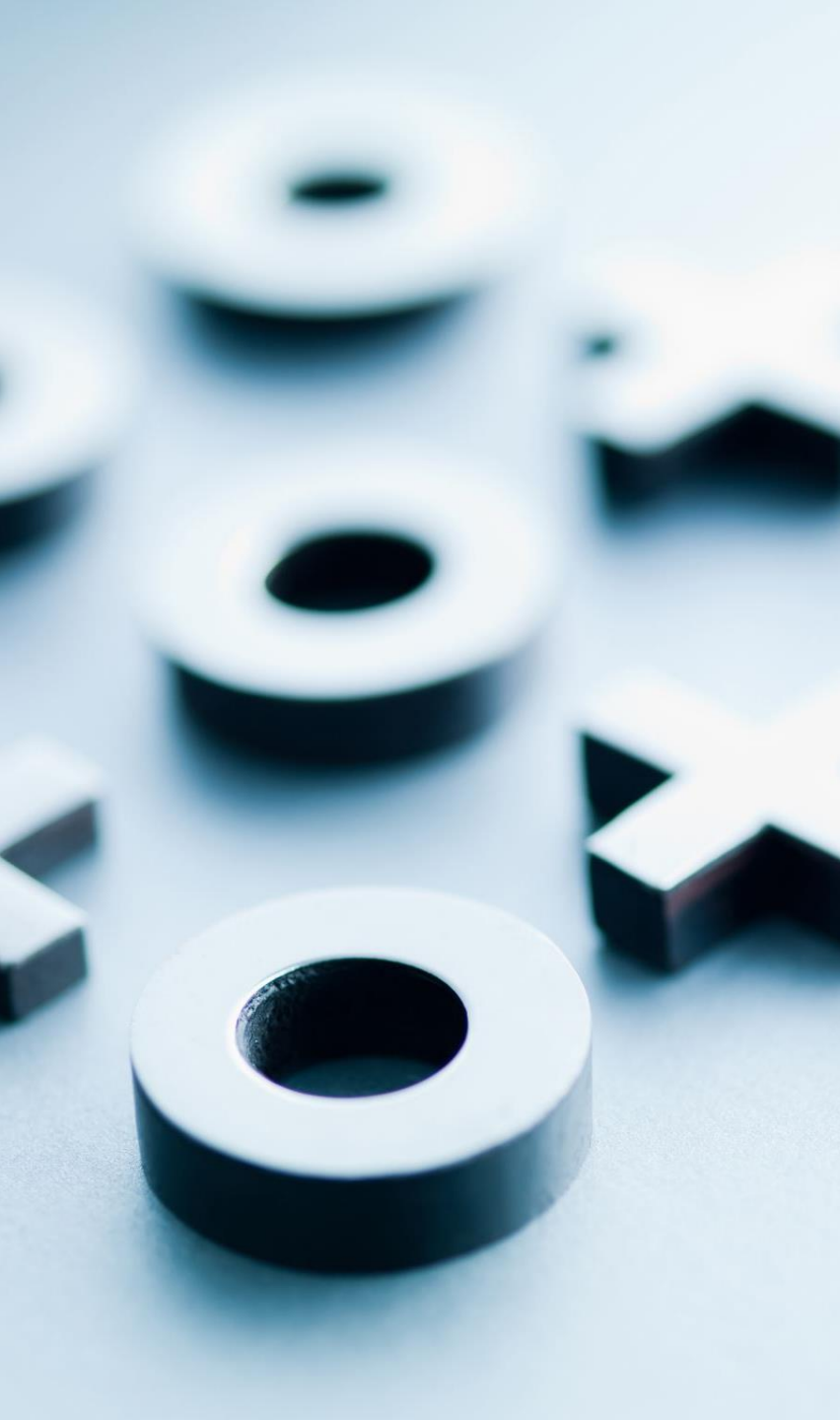
$SD/\sqrt{N-1}$

Minimum a maximum (range)

Quartily

Hodnoty pro 0.25, 0.5, 0.75





Co nám deskriptivní statistika neřekne?

Třída s profesorem Klouckým má výsledky v matematice v průměru 40 bodů, třída u paní Kloucké má 46 bodů. Učí paní Kloucká lépe?

Průměrná hodnota v našem vzorku písniček je XY pro Rockery a XY pro Hoperý. Mají Rockeři skladbu radši?

Jaký je vztah mezi energičností a oblibou skladby?

Systematická vs nesystematická chyba

Systematická

- systematicky zkreslená na jednu stranu
- statistika jí vůbec nevidí (jsou zkrátka jiné, než by měla být)

Náhodná chyba

- občas tak a občas jinak
- někdy zkrátka vyberu vzorek s nadanějšími studenty/depresivnějšími pacienty atd.
- Statistika modeluje a „odstraňuje“ náhodnou chybu výběru/měření

Statistika jako metoda odstraňování náhodné chyby

Průměrná hodnota v našem vzorku písniček je XY

Rozhodli byste tak, že rockeři skladbu preferují a měli bychom marketovat vůči nim?

Dvě možné chyby rozhodnutí

Marketujeme na rockery, přestože bychom mohli na všechny

Marketujeme na všechny, přestože bychom měli na rockery

Nulová hypotéza

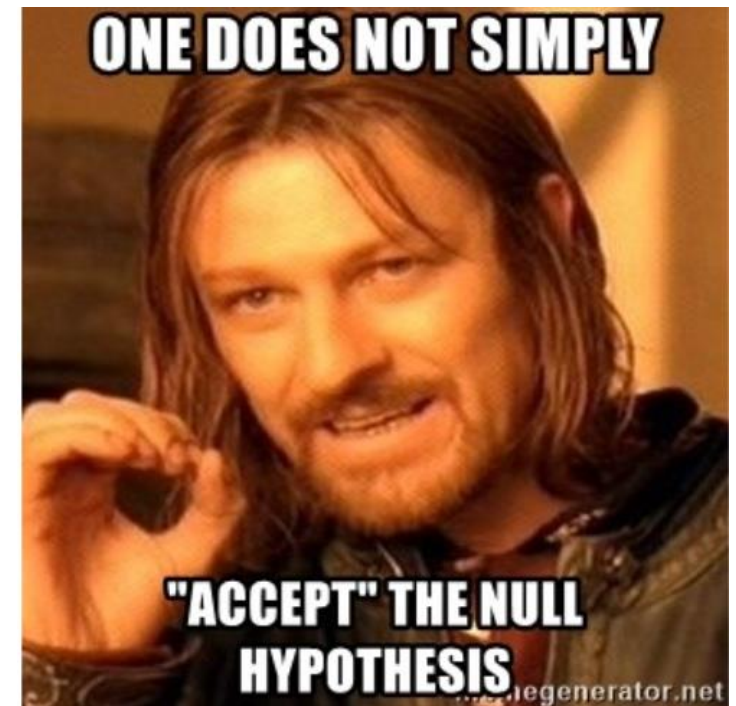
Typicky hypotéza o tom, že status quo platí

Mezi skupinami není rozdíl

Rockeři i Hopeři mají skladu stejně rádi

Závislost mezi proměnnými neexistuje

Mezi prožívanou energií a oblíbeností skladby není žádný vztah



Chyba prvního a druhého řádu

Chybné zamítnutí nulové hypotézy

Přestože to není pravda, data tvrdí, že ano

Důležitější ve vědeckém poznání

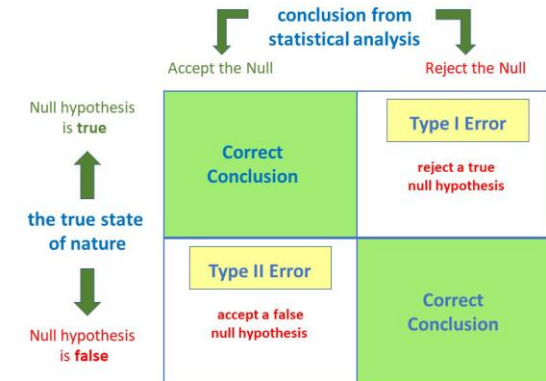
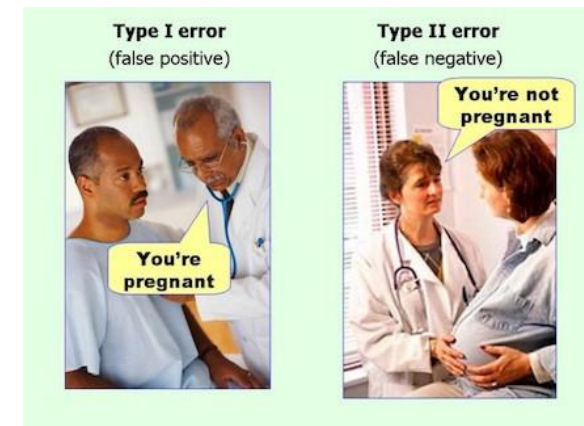
Raději se nedozvíme nic, než bychom se dozvěděli blbost

Chybné nezamítnutí nulové hypotézy

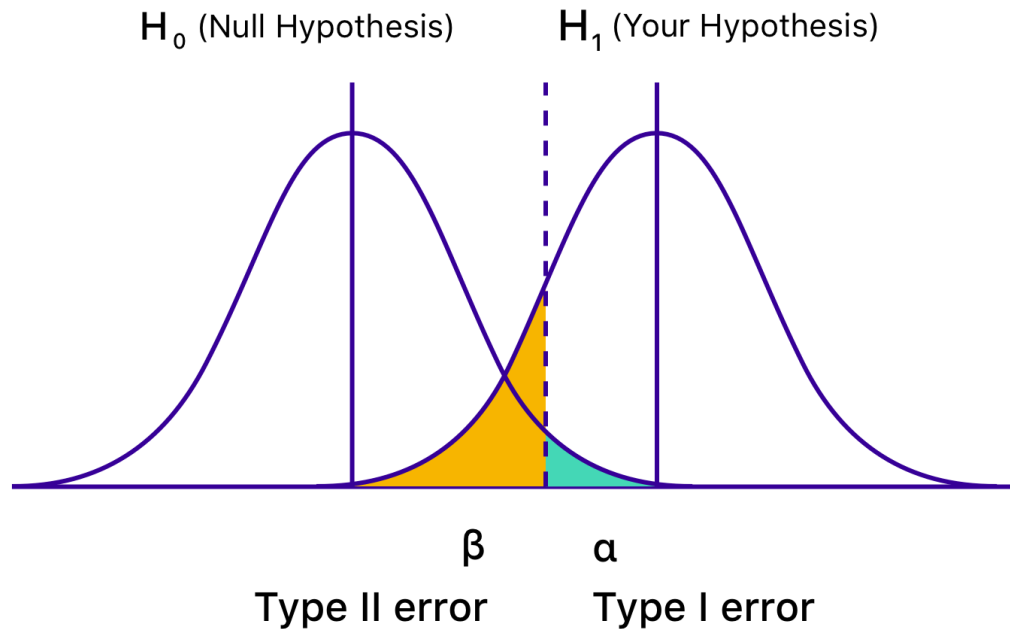
Méně problematické

Nejčastěji proto, že nemáme dost dat – trend existuje, ale není dost silný

Řeší se u tzv „power analýzy“



P hodnota



P hodnota značí pravděpodobnost, že došlo k chybě prvního řádu

Chybně zamítáme nulovou hypotézu

Hladina významnosti

Typicky 0.05, ale bezpečnější je 0.01 a menší

Studenti Kloucké jsou lepší v matematice než ti Klouckého

$P = 0.03 \Rightarrow$ existuje 3 procentní pravděpodobnost, že tento jev nastal i za předpokladu, že jsou v matematice stejně dobří

Statistické testy

Určují dva parametry

Pravděpodobnost, s jakou jste obdrželi výsledky náhodou

P-hodnota

Síla vypořádaného efektu

Jak moc vám znalost o dané proměnné pomůže v popsání dat

Každý test má jasné podmínky použití

Nutné dohledat a ověřit

Volba testu

Volba se odvíjí od:

Typu závislé proměnné

Typu nezávislé proměnné

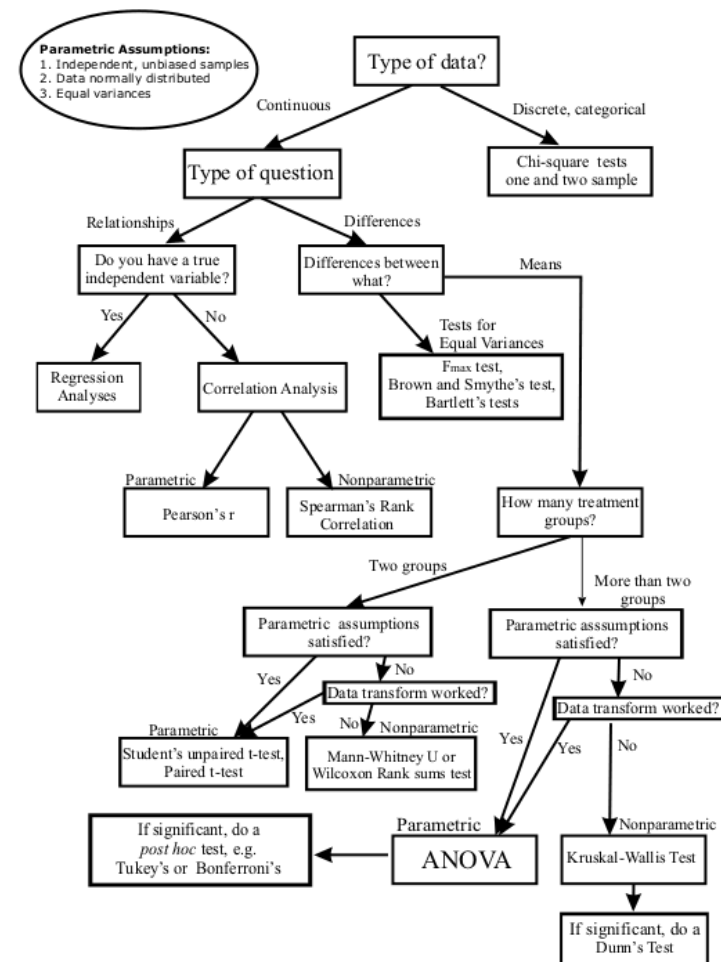
Množství dat

Výzkumné otázky

Typu dat (parametrické vs neparametrické)

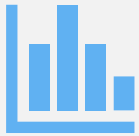
<https://stats.oarc.ucla.edu/other/mult-pkg/whatstat/>

Flow Chart for Selecting Commonly Used Statistical Tests





Podávají studenti po šálku kávy
lepší výsledky v testech
pozornosti?



Analyzují data lépe studenti z FF
nebo FHS?



V jaké hodině studenti diskutují
nejvíce? Sociální psychologie,
Metodologie či Datová analýza?

Porovnání
průměrů

Parametrické testy rozdílů

T test

- Porovnává průměry dvou skupin

- Párový či nepárový (within subject, between subject)

ANOVA

- Více skupin

- Between subject design

Repeated ANOVA

- Více skupin a více pozorování pro jednoho jedince

- Within-subject design

T-test

Testuje rozdíl mezi skupinami

Jednovzorkový (One-sample)

Porovnáváme vůči známé hodnotě

Dvouvzorkový (independent, two sample)

Porovnáváme dvě nezávislé skupiny

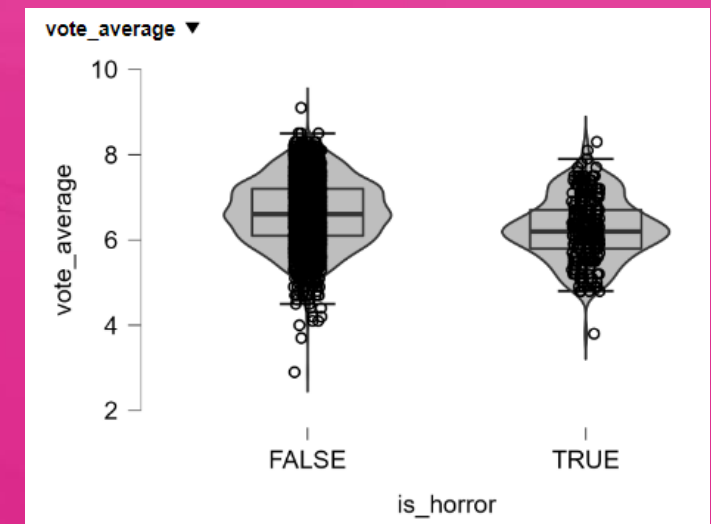
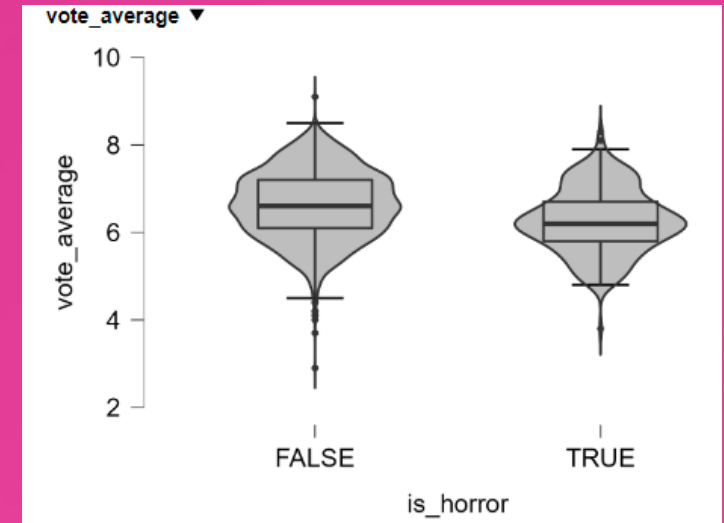
Párový (Paired)

Porovnáváme dvě podmínky (test ráno a večer)

Jeden člověk byl měřený vícekrát, měříme “změnu”

One-tailed/two-tailed

Síla efektu = Cohenovo D



T-test předpoklady

Pouze dvě skupiny

- nezávislé skupiny (independent t test)

- opakované měření (paired t test)

Kontinuální závislá proměnná

Žádní outliers – mohou posouvat výsledky

Závislá proměnná je rozložená normálně

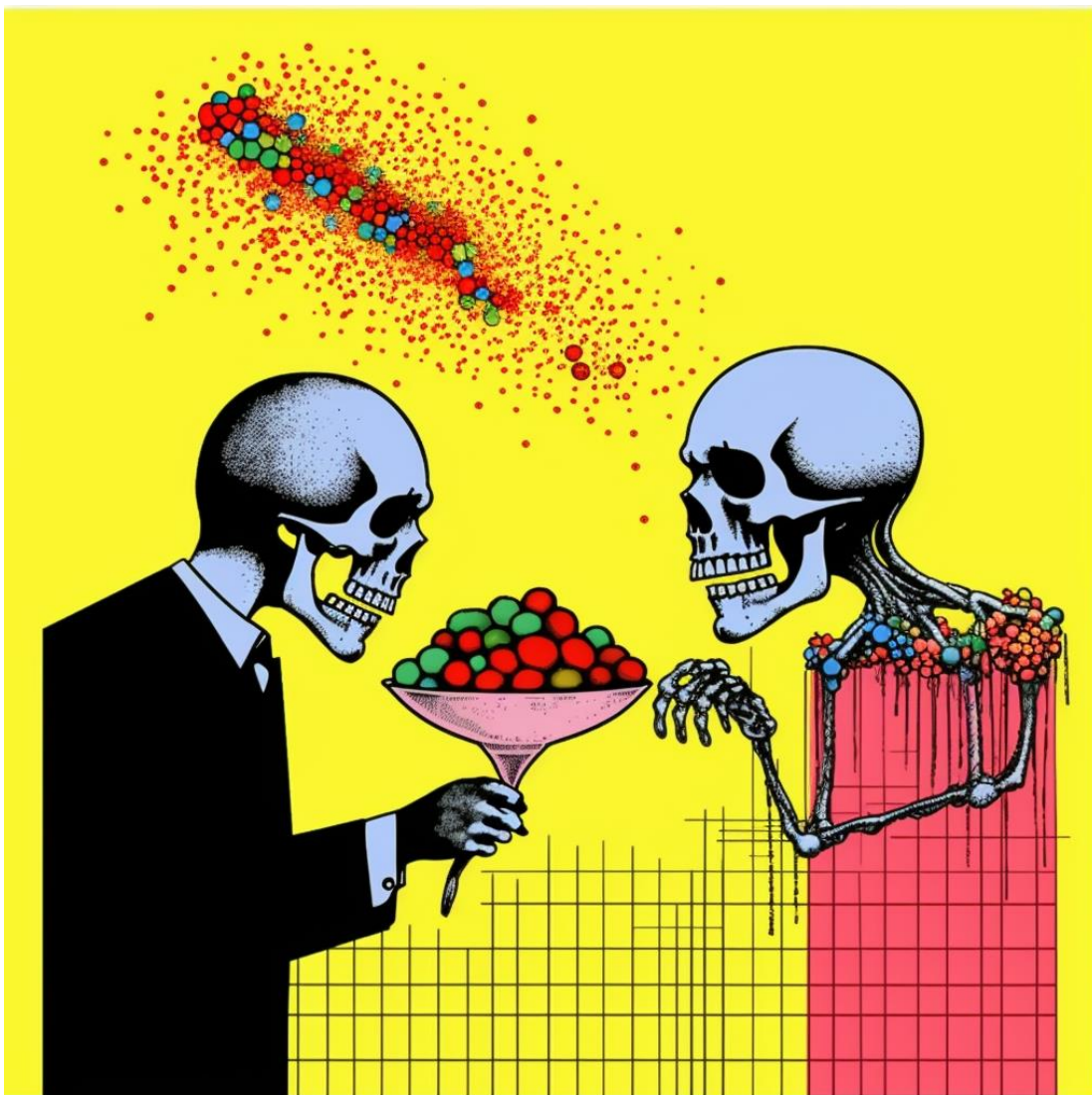
- Shapiro Wilk normality test

- Kouknu vidím

Skupiny mají stejnou variabilitu

- Možná korekce menších rozdílů

- Levene`s test for homogeneity of variances



Analýzy závislostí

Snažíme se využít jedné kontinuální proměnné a ukázat na její vliv na jinou proměnnou

Technicky i ANOVA je analýza závislostí, ale závislá na kategorickém prediktoru

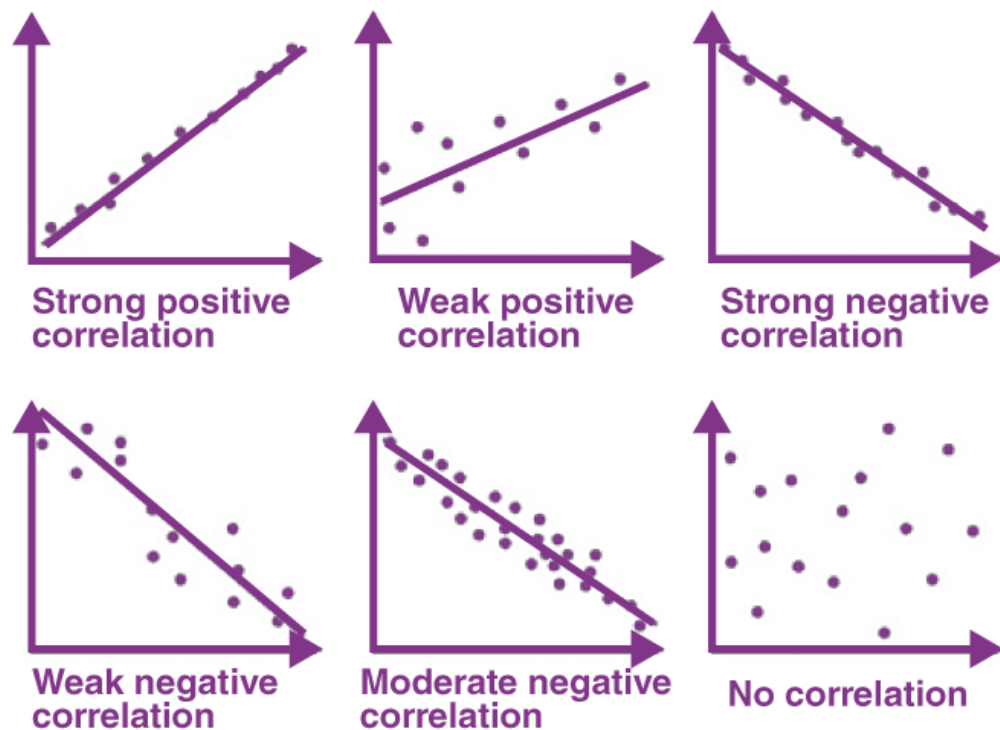
Typické přístupy jsou

Korelace

Pearsonova, Spearmanova

Regrese

Lineární, logistická, GLM, hierarchické modely, SEM etc.



Korelační analýza

Míra vztahu mezi dvěma proměnnými

Neukazuje na kauzalitu

Nevyjadřuje jak moc jedna proměnná
ovlivňuje druhou

Jen, že jdou “spolu”

Korelace je mezi -1 a 1

0 značí žádný vztah

Předpoklady korelace

Ověřit si, že máme opravdu kontinuální proměnné

Nejen na papíře, ale doopravdy

Pokud na škále 1-10 všichni odpovídají 7-9, tak to není kontinuální

Histogramy

Ověřit si, zdali jsou proměnné opravdu na sobě závislé

Scatter plots/Bodové grafy

Ověřit si, zdali jsou na sobě závislé lineárně

Pokud ano, regrese, pokud ne, Spearmanova korelace