MUNI ARTS

Základy štatistiky

Formální a experimentální sémantika II
Experimentální syntax a sémantika II
JS 2021

Mojmír Dočekal & Lucia Vlášková

Filozofická fakulta, Masarykova univerzita

Plán na semester

- čo všetko treba, aby som urobil/a experiment?
  všeobecný úvod
  prehľad procesu experimentu
- **pozorovanie zaujímavých dát, formulovanie hypotézy**
  základy štatistiky
  typy dát
- **Cimrmanův krok stranou (mimo procesu experimentu): základy Rka**
- získanie dát na analýzu
  - práca s korpusom
    Český národný korpus
    Zipfov zákon
    proporcie základných farebných termínov v subkorpusoch
  - vlastný experiment
    experimentálna lingvistika
    vytvorenie experimentu v L-rex

Plán na semester II

- úprava získaných dát
  - import a úprava dát v Rku
    exploratory data analysis
    balíček dplyr
  - čistenie dát
    outliery
    filtrovanie subjektov experimentu na základe fillerov
- analýza dát
  štatistické testy a kedy ich používať
  aplikácia na vlastné dáta
- vizualizácia dát a komunikácia výsledkov
  vizualizácia a grafy v Rku
  balíček ggplot2
  ucelené reportovanie experimentov
- prezentácia výsledkov
  individuálne výstupy študentov
  ukončenie predmetu

Plán na dnes

- pozorovanie zaujímavých dát, formulovanie hypotézy
  základy štatistiky
  typy dát
- Cimrmanův krok stranou (mimo procesu experimentu): základy Rka

Plán na dnes

- **pozorovanie zaujímavých dát, formulovanie hypotézy**
  základy štatistiky
  typy dát
  - čo je to štatistika + štatistická terminológia
  - typy hypotéz
  - *p*-value
  - výber štatistického testu
- **Cimrmanův krok stranou (mimo procesu experimentu): základy Rka**
  - kontrola úlohy 1: máme všetci Rko a RStudio?
  - kontrola úlohy 1: máme všetci L-rex?
  - áno $\rightarrow$ základy Rka
  - nie $\rightarrow$ riešime problémy

Základy štatistiky

Základy štatistiky

štatistika vs. štatistiky

Porovnanie tempa reči medzi belgickou a holandskou holandčinou

- populácia
- parameter
- vzorka
- štatistika, *-y*
- výberová chyba (*sampling error*)
  random sampling
  representative sampling
  convenience sampling

Porovnanie tempa reči medzi belgickou a holandskou holandčinou

- populácia
  všetci holandsky hovoriaci v Holandsku a Belgicku
- parameter
  priemerné tempo reči všetkých holandsky hovoriaci v Belgicku
- vzorka
  sto holandsky hovoriacich v Belgicku
- štatistika, *-y*
  priemerné tempo reči sto holandsky hovoriacich v Belgicku
- výberová chyba (*sampling error*)
  random sampling: zoznam všetkých holandsky hovoriacich v Holandsku a Belgicku
  representative sampling: rovnaké pomery mužov/žien, vekových skupín, etnických skupín, dialektov...
  convenience sampling: holandsky hovoriaci v najväčších mestách

Dva druhy štatistiky

- deskriptívna/popisná štatistika
  priemerné tempo reči sto holandsky hovoriacich v Belgicku
  priemer, medián, odchýlka...
  výsledky nemožno zovšeobecniť na populáciu alebo inú skupinu
- inferenčná štatistika
  porovnanie priemerného tempa reči sto holandsky hovoriacich v Belgicku a Holandsku
  lineárne regresné analýzy, logistické regresné analýzy, ANOVA, korelačné analýzy...
  testy významnosti: chí-kvadrát, t-test...
  výsledky sa zovšeobecňujú na populáciu

Hypotézy

- nulová vs. alternatívna
- direkcionálna vs. nedirekcionálna

Hypotézy

nulová vs. alternatívna

- H0: neexistuje žiaden vzťah medzi tempom reči a krajinou holandsky hovoriaceho
- H1: holandsky hovoriaci v Holandsku majú vyššie priemerné tempo reči ako holandsky hovoriaci v Belgicku

Hypotézy

direkcionálna vs. nedirekcionálna

- H0: neexistuje žiaden vzťah medzi tempom reči a krajinou holandsky hovoriaceho
- direkcionálna H1: holandsky hovoriaci v Holandsku majú vyššie priemerné tempo reči ako holandsky hovoriaci v Belgicku
- nedirekcionálna H1: holandsky hovoriaci v Holandsku majú rozdielne tempo reči od holandsky hovoriacich v Belgicku

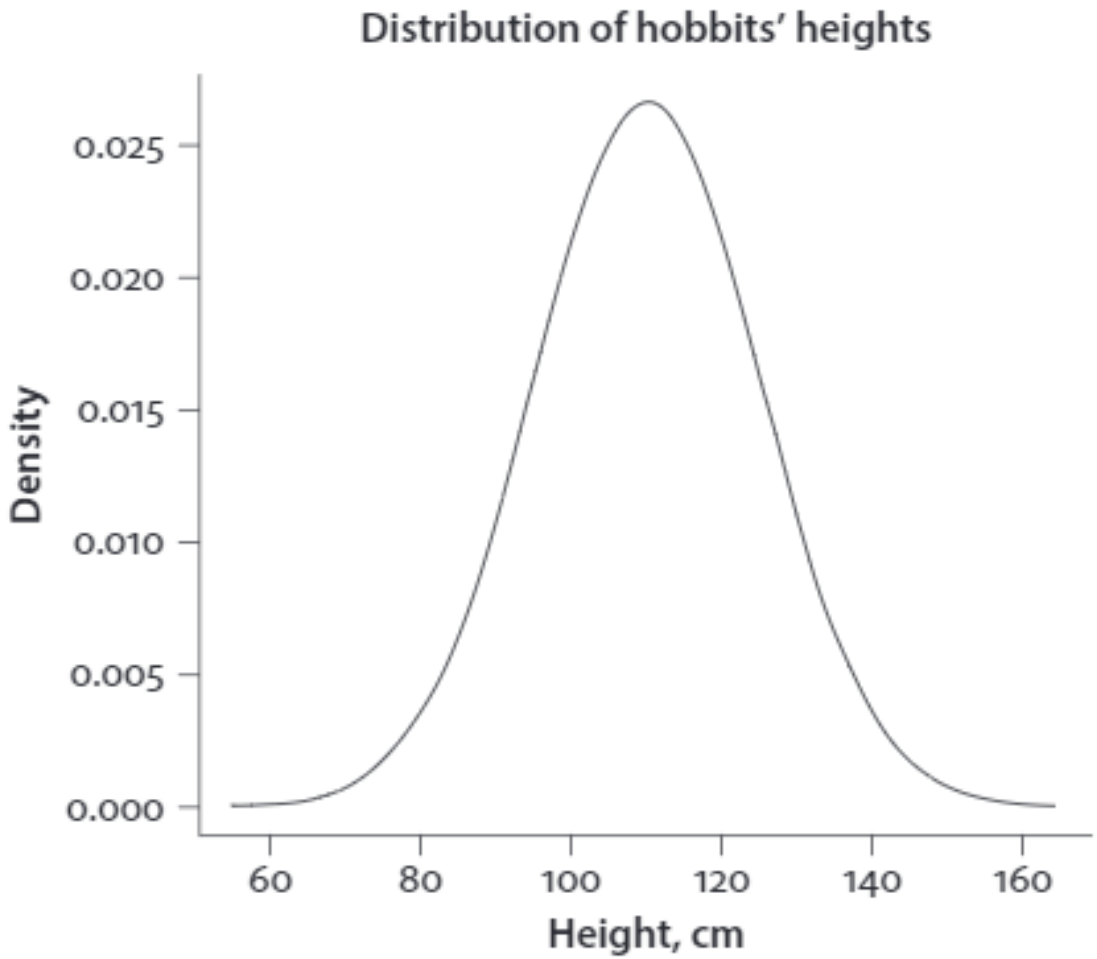
Hypotézy

cvičenie

- H0: There is no difference in the number of lexemes that denote snow in Eskimo and Yucatec Maya.
- H1: There are more lexemes that denote snow in Eskimo than in Yucatec Maya.
- H0: There is no relationship between the frequency of a word and how fast it is recognized in a lexical decision task.
- H1: The more frequent a word, the faster it is recognized in a lexical decision task.
- H0: There is no difference in the relative frequencies of metaphoric expressions used by men and women when they speak about sex.
- H1: There is a difference in the relative frequencies of metaphoric expressions used by men and women when they speak about sex.

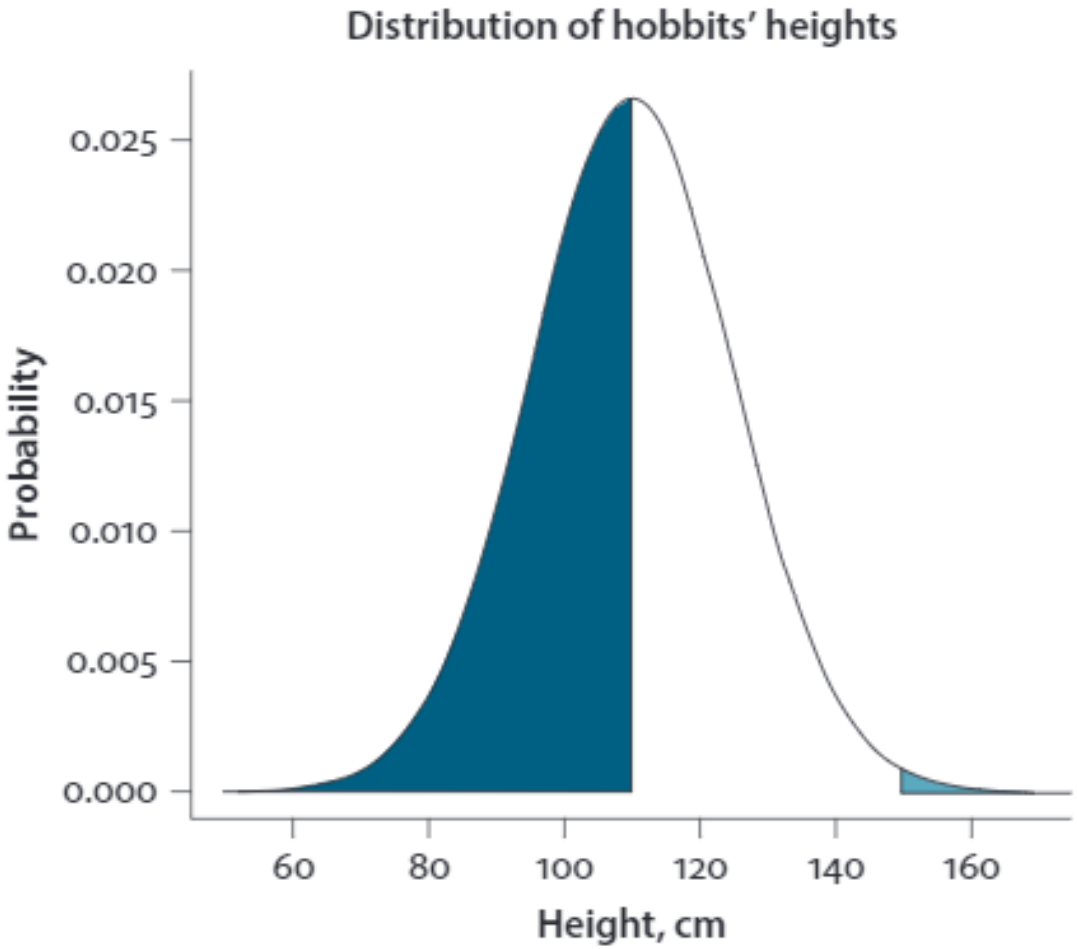
Hypotézy

rozhodovanie



Hypotézy

rozhodovanie



Hypotézy

rozhodovanie

p-value

The p-value shows the probability of obtaining a given test statistic value or more extreme values if the null hypothesis is true.

- miera významnosti (*significance level*)
  - 0.05 alebo 0.01
  - určený dopredu
- $p > 0.05 \rightarrow$ H0 nezamietnutá, H1 neplatí
- $p < 0.05 \rightarrow$ H0 zamietnutá, H1 platí

Hypotézy

chyby

- chyba typu I
  - falošne pozitívny
  - pozitívny antigén, ale nemám COVID
- chyba typu II
  - falošne negatívny
  - negatívny antigén, ale mám COVID

Štatistické testy

výber testu

Štatistické testy

výber testu

- škála premennej
- distribúcia premennej: parametrický vs. neparametrický
- typ premenných: závislé vs. nezávislé
- typ alternatívnej hypotézy: direkcionálna vs. nedirekcionálna
- počet porovnávaných skupín: dve vs. viac

Štatistické testy

škála premennej

- nominálne (kategorické)
  - zaradenie do skupín, bez kvantitatívnej informácie, bez poradia
  - pohlavie, národnosť, native vs. non-native, materský jazyk, pád NP...
- ordinálne
  - poradie
  - Likertova škála 1-5
- intervalové
  - vzdialenosť medzi dvoma „susediacimi“ premennými je rovnaká
  - poradie aj veľkosť rozdielov
  - dĺžka slov v písmenách/slabikách
- pomerové
  - absolútna nula, multiplikácia
  - počet napr. adverbií v texte

Štatistické testy

škála premennej

Nominal scale	Ordinal scale	Interval scale	Ratio scale
Chi-square	Median	Mean	Coefficient of variation
Phi coefficient	Interquartile range	Standard deviation	Sign test
Cramér's *V*	Spearman's correlation coefficient	Pearson correlation coefficient	Median test
Contingency coefficient	Kendall's tau	*t*-test	
Uncertainty coefficient	Kolmogorov Smirnov test	Analysis of variance	
Kappa	Kendall coefficient of concordance	Multivariate analysis of variance	
Likelihood ratio	Friedman two-way anova	Factor analysis	
Goodman & Kruskal tau	Mann–Whitney *U*-test	Regression	
McNemar	Wald–Wolfowitz	Multiple correlation	
Cochran *Q*	Kruskal–Wallis	Sign test	
	Sign test	Median test	
	Median test		

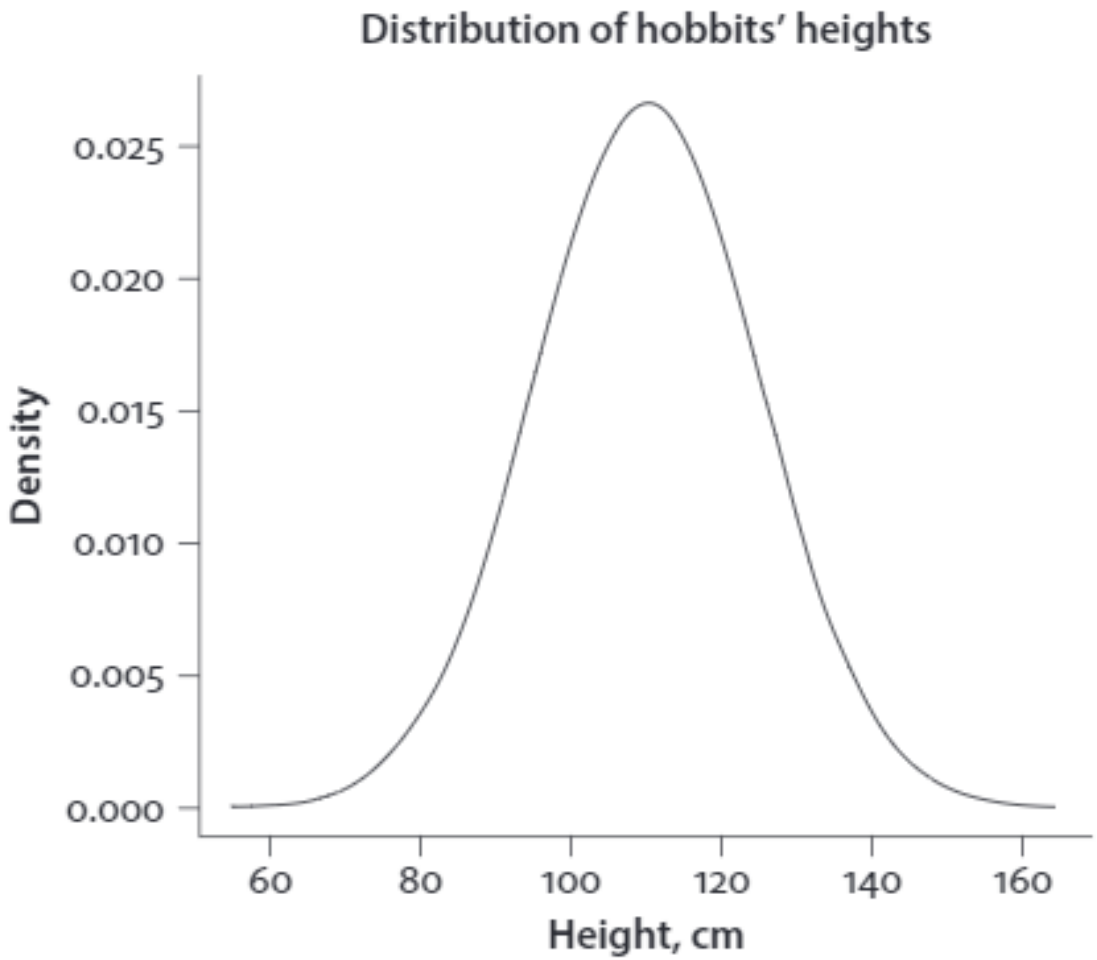
Štatistické testy

distribúcia premennej

- parametrický test
  - známa parametrická distribúcia pravdepodobnosti premennej
  - najčastejšie normálna (Gaussova krivka; hobiti)
- neparametrický test
  - pracuje s populáciou mimo nejakej známej parametrickej distribúcie pravdepodobnosti
  - nepredpokladá nič o dátach
  - odporúča sa pri malých vzorkách

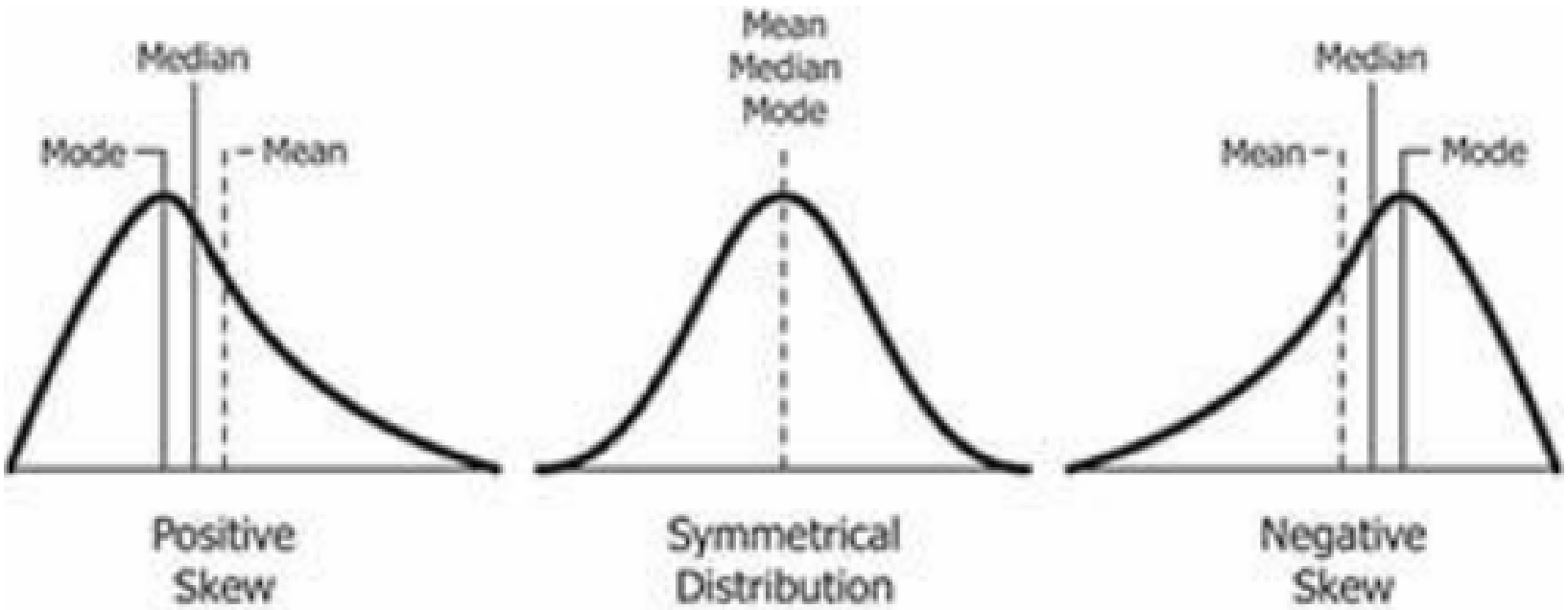
Štatistické testy

parametrická distribúcia



Štatistické testy

neparametrická distribúcia



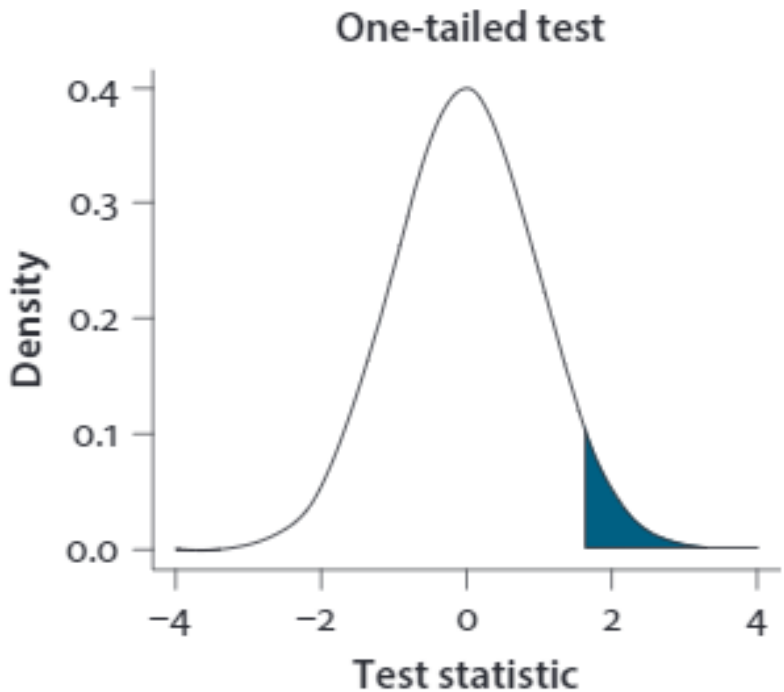
Štatistické testy

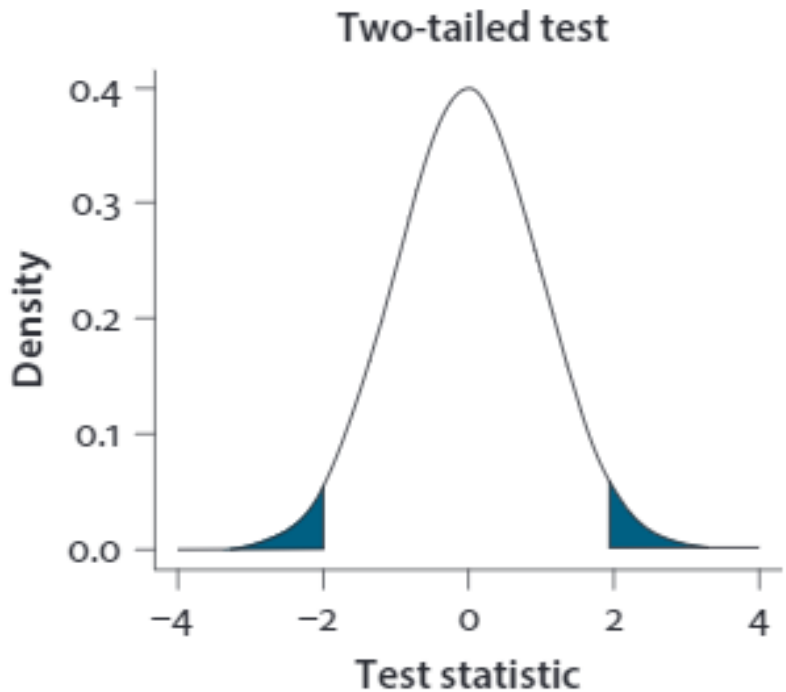
typ premenných

- závislé premenné $\rightarrow$ dependent (paired) test
  - pozorovanie z jednej skupiny má ekvivalent/súvislosť v druhej, napr. rovnaký subjekt alebo stimulus
  - body v teste pred a po novej metóde učenia (spoločný subjekt)
  - veľkosť slovnej zásoby pred a po novej metóde učenia (spoločný subjekt)
- nezávislé premenné $\rightarrow$ independent (unpaired) test
  - hodnoty z dvoch skupín bez spojenia/súvislosti
  - body v teste pred ALEBO po novej metóde učenia
  - veľkosť slovnej zásoby na konci semestra

Štatistické testy

typ alternatívnej hypotézy





Štatistické testy

typ alternatívnej hypotézy

- direkcionálna alternatívna hypotéza $\rightarrow$ one-tailed test
- nedirekcionálna alternatívna hypotéza $\rightarrow$ two-tailed test

Statistické testy

Table 3.2 Statistical tests described.

	Tests of differences between groups (independent samples)		Tests of differences between variables (dependent samples)	Tests of relationships between variables
	Two groups	Multiple groups		
Parametric statistics	*t-test for independent samples* §3.2.1	*ANOVA* §3.2.3	*t-test for paired samples* §3.2.2	*Pearson correlation coefficient* §3.2.4
Non-parametric statistics	*Mann–Whitney U-test* §3.3.1	*Median test* §3.3.4	*Sign test* §3.3.2	*Spearman rank correlation coefficient* §3.3.5 *Chi-square test* §3.3.3

Krok stranou: Rko

Krok stranou: Rko

kontrola úlohy 1:

- máme všetci Rko a RStudio?
- máme všetci L-rex?

Plán na dnes

- pozorovanie zaujímavých dát, formulovanie hypotézy
  základy štatistiky
  typy dát
  - čo je to štatistika + štatistická terminológia
  - typy hypotéz
  - *p*-value
  - výber štatistického testu
- Cimrmanův krok stranou (mimo procesu experimentu): základy Rka
  - kontrola úlohy 1: máme všetci Rko a RStudio?
  - kontrola úlohy 1: máme všetci L-rex?
  - áno $\rightarrow$ základy Rka
  - nie $\rightarrow$ riešime problémy

Plán na semester

- čo všetko treba, aby som urobil/a experiment?
  všeobecný úvod
  prehľad procesu experimentu
- pozorovanie zaujímavých dát, formulovanie hypotézy
  základy štatistiky
  typy dát
- **Cimrmanův krok stranou (mimo procesu experimentu): základy Rka**
- získanie dát na analýzu
  - práca s korpusom
    Český národný korpus
    Zipfov zákon
    proporcie základných farebných termínov v subkorpusoch
  - vlastný experiment
    experimentálna lingvistika
    vytvorenie experimentu v L-rex

Plán na semester II

- úprava získaných dát
  - import a úprava dát v Rku
    exploratory data analysis
    balíček dplyr
  - čistenie dát
    outliery
    filtrovanie subjektov experimentu na základe fillerov
- analýza dát
  štatistické testy a kedy ich používať
  aplikácia na vlastné dáta
- vizualizácia dát a komunikácia výsledkov
  vizualizácia a grafy v Rku
  balíček ggplot2
  ucelené reportovanie experimentov
- prezentácia výsledkov
  individuálne výstupy študentov
  ukončenie predmetu

MASARYKOVA
UNIVERZITA