Kategorická data

METODOLOGICKÝ PROSEMINÁŘ II

TÝDEN 7 | 4. DUBNA 2018

Typy proměnných

- nominální (nominal)
  - o dvou hodnotách lze říci pouze to, zda jsou stejné či různé
  - např. pohlaví, politická strana, typ politického režimu apod.

kvalitativní data

- ordinální (ordinal)
  - u hodnot je navíc možné určit jejich pořadí („co je více a co méně“)
  - např. úroveň dosaženého vzdělání, míra spokojenosti s politickým režimem apod.

lze užít jako oba typy dat

- intervalová (interval)
  - lze vypočítat, o kolik je jedna hodnota větší (menší) než druhá
  - např. míra růstu HDP, výše inflace apod.

kvantitativní data

- poměrová (ratio)
  - lze vypočítat, kolikrát je jedna hodnota větší (menší) než druhá – existuje absolutní nula, pouze kladné hodnoty
  - např. počet hlasů získaných ve volbách, počet poslanců hlasujících pro zákon apod.

Typy proměnných

Typ proměnné	Matematické operace	Lze vypočítat	Příklad
Nominální	=, ≠	Modus	Pohlaví
Ordinální	=, ≠, >, <	Modus, medián	Pořadí sportovců v závodě
Intervalová	=, ≠, >, <, +, -	Modus, medián, aritmetický průměr	Teplota ve stupních Celsia
Poměrová	=, ≠, >, <, +, -, *, /	Modus, medián, aritmetický průměr	Teplota ve stupních Kelvina, hmotnost

Kategorická data

- jiný název pro kvalitativní proměnné
- kvalitativní proměnné mají zejména nominální, ale často i ordinální podobu
  - záleží, jak k nim přistupujeme (jakou analýzu využíváme)
- možnosti práce s těmito typy proměnných jsou omezené
- proměnné proto stavíme základním způsobem vedle sebe a odhalujeme vzájemné vztahy k dalším datům
- příkladem analýzy kategorických dat jsou kontingenční tabulky

Kontingenční tabulka

- primární metoda využívaná pro analýzu vztahu nominálních a ordinálních proměnných
- má předepsanou podobu
  - v jednotlivých sloupcích jsou hodnoty nezávisle proměnné
  - v řádcích jsou hodnoty závisle proměnné
  - hodnoty proměnných jsou řazeny logicky (od nejnižší spokojenosti po nejvyšší, od nejvyšší důvěry po nejnižší apod.)
- samotné kontingenční tabulky uvádí hodnoty absolutních pozorovaných četností
- pro interpretaci a primární odhalení souvislostí se uvádí procentuální četnosti
- pozor na to, zda jsou uváděna sloupcová nebo řádková procenta!
- pro vytvoření kontingenční tabulky v Excelu je možné využít například funkci „Kontingenční tabulka“ (*Pivot Table*)

Kontingenční tabulka

CVVM, Naše společnost – říjen 2017

nezávisle proměnná

závisle proměnná

Pozorované četnosti	**Vzdělání**				
Důvěra prezidentovi	základní	střední bez maturity	střední s maturitou	vysokoškolské	**Total**
rozhodně důvěřuje	31	39	40	16	126
spíše důvěřuje	61	119	116	44	340
spíše nedůvěřuje	32	90	77	43	242
rozhodně nedůvěřuje	32	65	61	47	205
Total	156	313	294	150	913

Procentuální četnosti	**Vzdělání**			
Důvěra prezidentovi	základní	střední bez maturity	střední s maturitou	vysokoškolské
rozhodně důvěřuje	19,87	12,46	13,61	10,67
spíše důvěřuje	39,10	38,02	39,46	29,33
spíše nedůvěřuje	20,51	28,75	26,19	28,67
rozhodně nedůvěřuje	20,51	20,77	20,75	31,33
Total	100	100	100	100

Pearsonův chí-kvadrát test

- základní a nejpoužívanější test nezávislosti v kontingenční tabulce
- H_0: náhodné veličiny X a Y jsou nezávislé
  - pravděpodobnost nastání určité hodnoty proměnné X neovlivňuje nastání určité hodnoty proměnné Y
- při výpočtu v Excelu postupujeme následovně:
  1. zapíšeme pozorované četnosti (POZOR – do prázdných buněk je třeba zapsat nulu!)
  2. ve vedlejší tabulce vypočteme očekávané četnosti

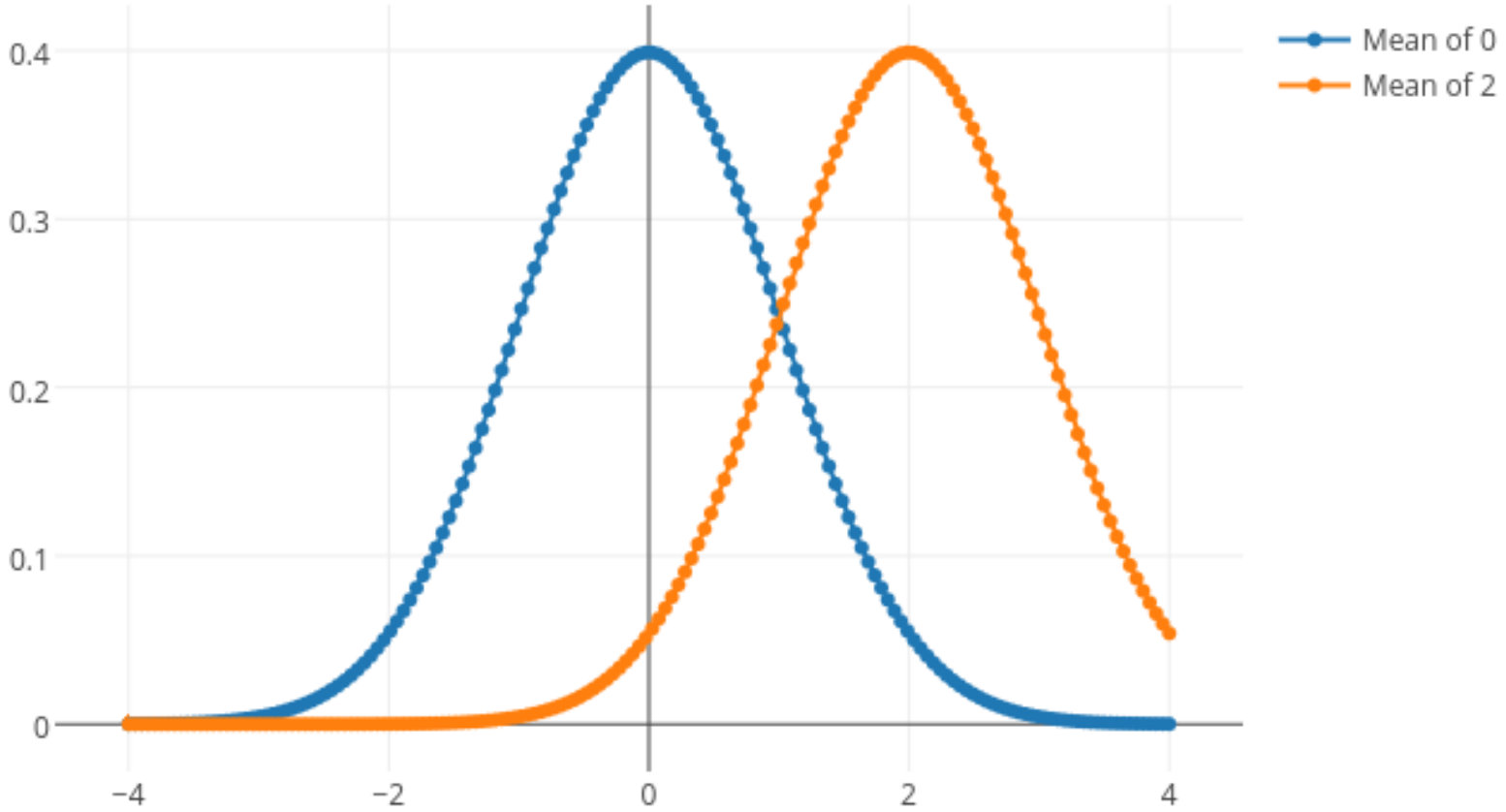
$$očekávaná\ četnost = \frac{součet\ všech\ četností\ v\ řádku * součet\ všech\ četností\ ve\ sloupci}{součet\ četností\ v\ celé\ tabulce}$$

  3. následně srovnáme četnosti pozorované a očekávané v rámci chí-kvadrát testu - funkce CHITEST()
  4. výsledek uvádí pravděpodobnost platnosti nulové hypotézy
     - čím je číslo nižší, tím pravděpodobnější je, že veličiny X a Y jsou na sobě vzájemně závislé
     - rozhodující je pro nás úroveň 0,05 – ta určuje hranici pravděpodobnosti 95 % (pokud je výsledek testu nižší než 0,05, máme minimálně 95% jistotu, že veličiny X a Y jsou na sobě vzájemně závislé)

t-test

- proměnné, u kterých už máme možnost smysluplně spočítat například průměr (intervalové, poměrové), můžeme zkoumat dalšími více pokročilými metodami
- jednou z nich je např. t-test
- umožňuje ověřit, zda dvě normální rozložení, z nichž pocházejí dva nezávislé výběry, mají stejné střední hodnoty (průměry)
- H_0: průměry dvou populací jsou stejné
  - H_0: $\mu_1 = \mu_2$
- alternativní hypotézou je, že průměry dvou populací se signifikantně liší (liší se tedy tyto dvě populace v určité hodnotě)

t-test



plot.ly

t-test

- postup při výpočtu je následující
  1. vypočítat průměry a standardní odchylky obou vzorků
  2. vypočítat jednotlivé standardní chyby obou vzorků
  3. vypočítat celkovou standardní chybu vzorků

$$SE_d = \sqrt{SE_1^2 + SE_2^2}$$

  4. vypočítat t-skóre

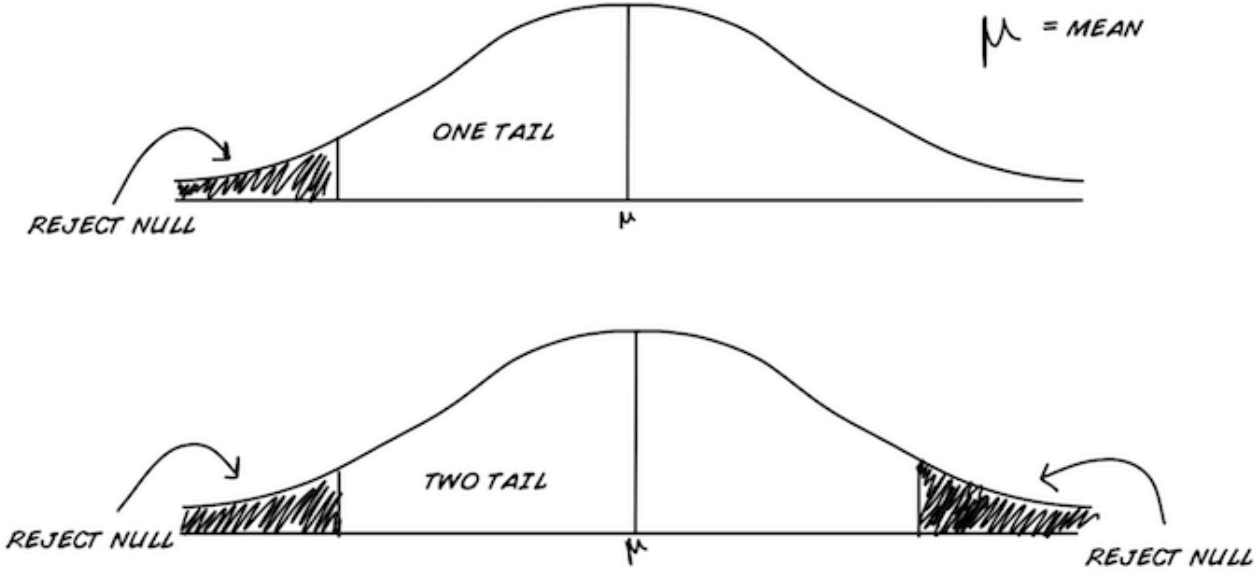
$$t = \frac{\overline{X_1} - \overline{X_2}}{celková\ standardní\ chyba}$$

  5. porovnat t-skóre s t-tabulkou za účelem odhalení pravděpodobnosti platnosti nulové hypotézy

- při výpočtu t-testu v Excelu získáme přímo hodnotu pravděpodobnosti platnosti nulové hypotézy
- pozor na různé t-testy vzhledem k vzájemné závislosti vzorků
  - v případě nejistoty vždy používat nejvíce konzervativní nepárový test pro vzorky s rozdílnými rozptyly

Test jednostranný nebo oboustranný?

- proti nulové hypotéze stavíme alternativní hypotézu
- ta může být buď jednostranná nebo oboustranná
- pokud je alternativní hypotéza H_A: $\mu_1 \neq \mu_2$, je možné, že $\mu_1 > \mu_2$, nebo $\mu_1 < \mu_2$ a musíme proto použít dvoustranný test
  - například současní poslanci nechybí stejně často jako jejich kolegové v 90. letech
- pokud je ale alternativní hypotéza například jen H_A: $\mu_1 > \mu_2$, použijeme jednostranný test
  - například současní poslanci chybí méně často než jejich kolegové v 90. letech



backyardbrains.com

t-tabulka

- při výpočtu t-skóre se podíváme do t-tabulky, co hodnota značí
- najdeme si příslušný řádek podle stupňů volnosti (degrees of freedom)
  - pokud řádek s příslušnými stupni volnosti chybí, použijeme nejbližší konzervativnější hodnotu (např. pokud chybí df = 36, použijeme df = 30)
- určíme dva sloupce, mezi kterými se hodnota t-skóre nachází
- sloupec s nižší hodnotou jistoty značí naši minimální pravděpodobnost odlišnosti srovnávaných dat
- vzhledem ke konvenci nás obecně zajímá, zda je pravděpodobnost vyšší či nižší než 95 %

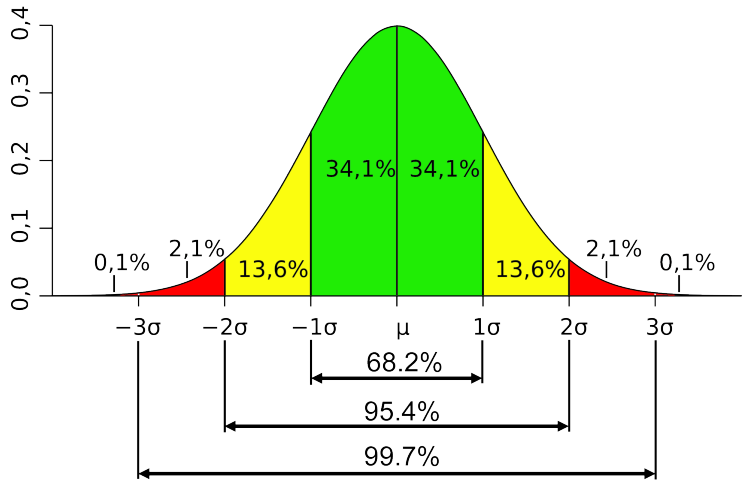
cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

ttable.org

z-skóre

- se zvyšujícím se počtem df se rozložení blíží normálnímu
- v normálním rozložení z-skóre

$$z = \frac{x - \mu}{\sigma}$$



kanbanize.com

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

ttable.org

Výpočet v Excelu

- při výpočtu pomocí Excelu využíváme funkci T.TEST()
- musíme zde určit několik parametrů
  - první vzorek dat
  - druhý vzorek dat
  - jednostranný (1) vs. dvoustranný test (2)
  - typ testu – párový (1), stejné rozptyly (2), různé rozptyly (3)
- výsledek uvádí pravděpodobnost platnosti nulové hypotézy
  - t-skóre je spočítáno v mezikroku a nemusíme se zabývat tabulkami
  - čím je číslo nižší, tím pravděpodobnější je, že dva
  - rozhodující je pro nás úroveň 0,05 – ta určuje hranici pravděpodobnosti 95 % (pokud je výsledek testu nižší než 0,05, máme minimálně 95% jistotu, že vzorky pocházejí z různých populací)
- https://www.evanmiller.org/ab-testing/t-test.html

Shrnutí

- kvalitativní (kategorická) data lze zkoumat omezeným množstvím statistických nástrojů
- k jejich zobrazení a porovnání využíváme např. kontingenční tabulky
- v rámci těch aplikujeme chí-kvadrát test určující přítomnost vztahu mezi proměnnými
- pro porovnání dvou vzorků používáme t-test
- při ručním výpočtu srovnáme t-skóre s t-tabulkou a odhalíme pravděpodobnost rozdílnosti obou populací
- při výpočtu v programovacím prostředí získáme rovnou výslednou pravděpodobnost, že vzorky pocházejí z identických populací