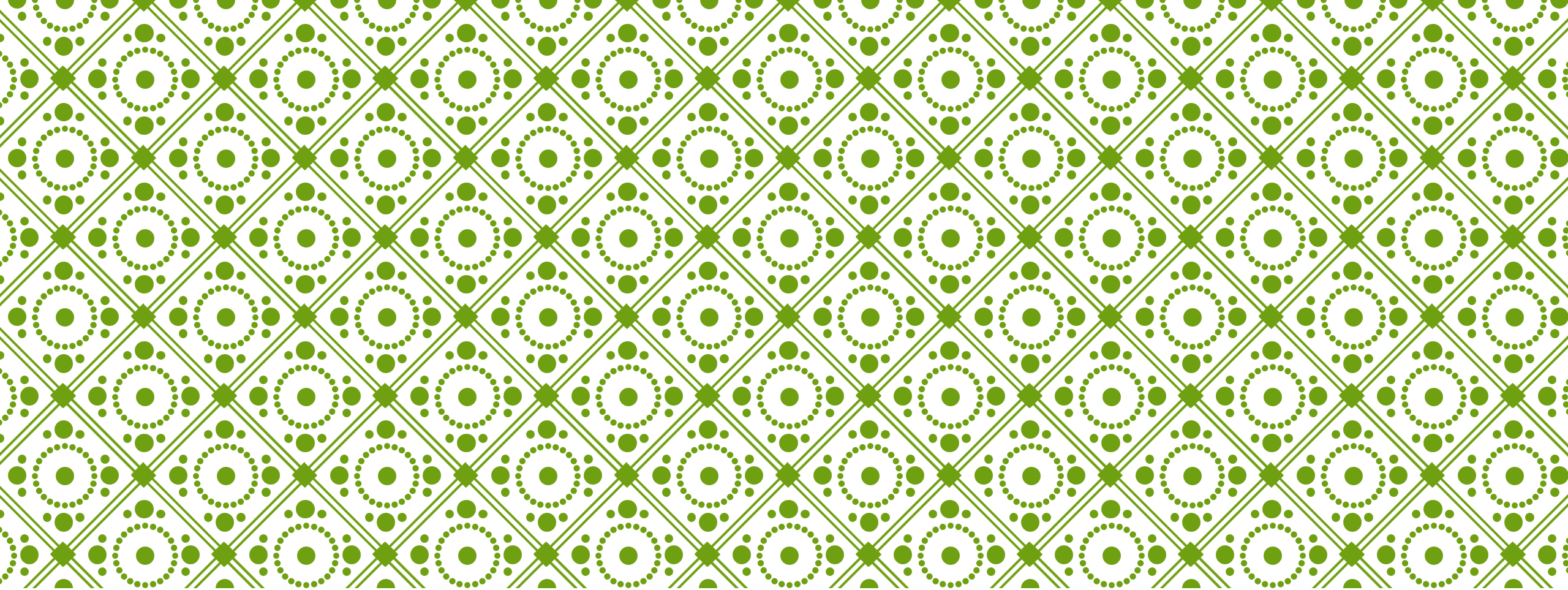


NORMÁLNÍ ROZLOŽENÍ, ZÁKLADY TESTOVÁNÍ HYPOTÉZ A STATISTICKÁ INFERENCE.

ZUR357 Statistická analýza dat --

9. listopadu 2017

AUTOMATIC RECODE

Transformace string values do numerických hodnot

CATEGORIZE VARIABLES

Kategorizuje kardinální znaky podle percentilů tak, aby každá kategorie obsahovala přibližně stejný počet případů

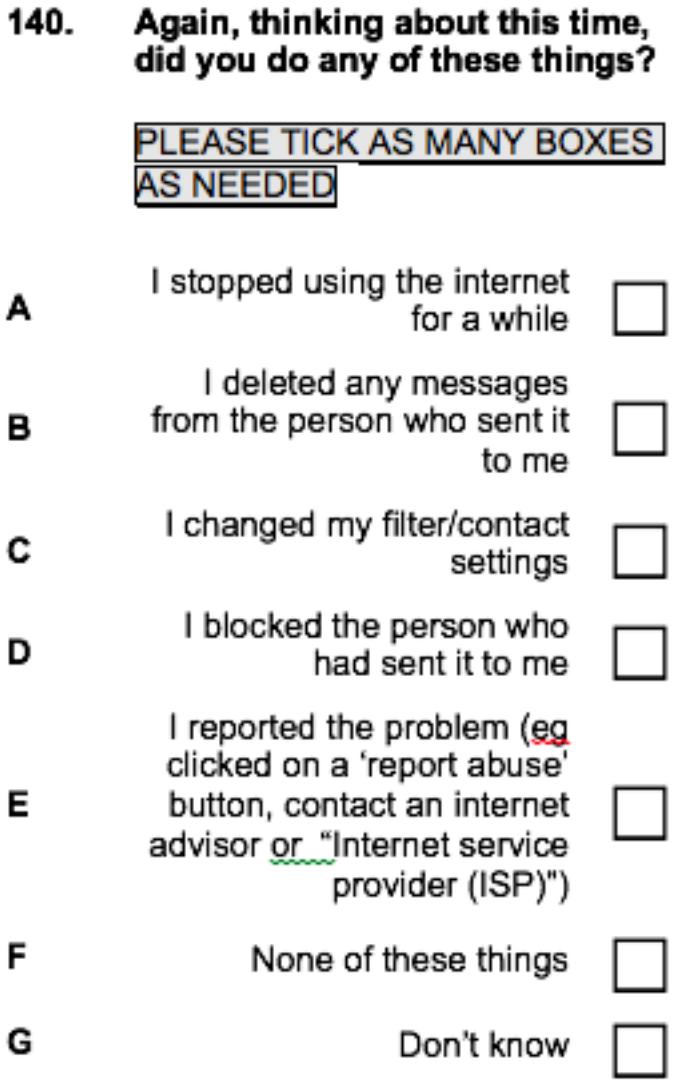
COUNT VALUES

Sčítání výskytu určité hodnoty

Např. Které tituly pravidelně čtete (0=nečte, 1čte)

- Respekt, Reflex, Euro, Týden….

Kolik strategií používají děti, aby se vypořádali s obtěžo



140. Again, thinking about this time, did you do any of these things?

PLEASE TICK AS MANY BOXES AS NEEDED

A	I stopped using the internet for a while	☐
B	I deleted any messages from the person who sent it to me	☐
C	I changed my filter/contact settings	☐
D	I blocked the person who had sent it to me	☐
E	I reported the problem (eg clicked on a 'report abuse' button, contact an internet advisor or "Internet service provider (ISP)")	☐
F	None of these things	☐
G	Don't know	☐

RANK CASES

Vytvoří novou proměnnou, která určuje pořadí případu podle původní proměnné

COMPUTE

Libovolné operace zadané strukturovaným příkazem

Podmínka IF…

Změna rozložení

Centrování

Sumační indexy

Komplexní operace

144. In the PAST 12 MONTHS, how often, have these things happened to you?

PLEASE TICK ONE BOX ON **EVERY** LINE

		Very often	Fairly often	Not very often	Never/ almost never
A	I have gone without eating or sleeping because of the internet	☐	☐	☐	☐
B	I have felt bothered when I cannot be on the internet	☐	☐	☐	☐
C	I have caught myself surfing when I'm not really interested.....	☐	☐	☐	☐
D	I have spent less time than I should with either family, friends or doing schoolwork because of the time I spent on the internet.....	☐	☐	☐	☐
E	I have tried unsuccessfully to spend less time on the internet	☐	☐	☐	☐

PRÁCE S PODSOUBORY

Procedura SELECT CASES

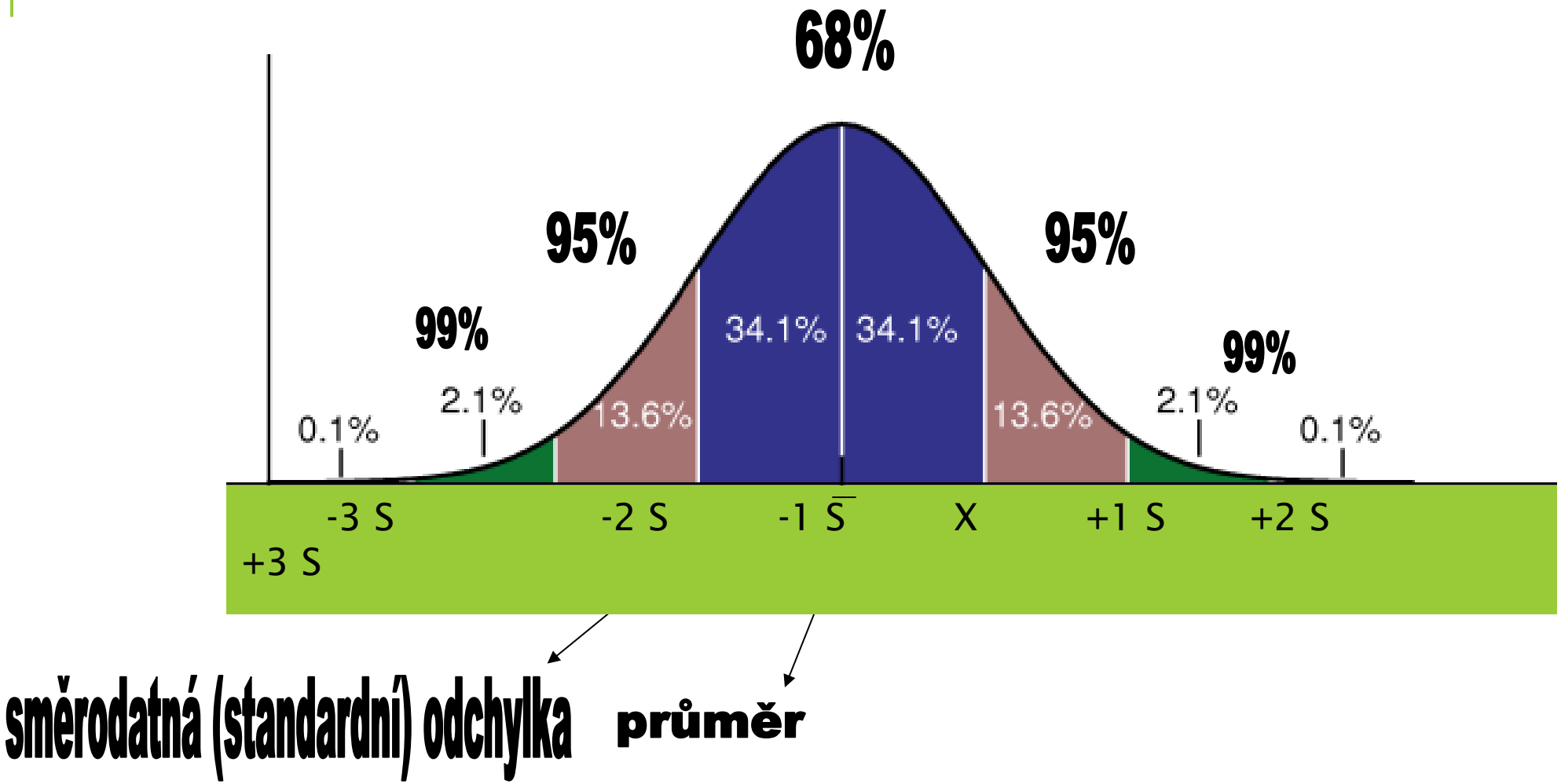
- Náhodně vybrané případy – redukce souboru (Random Sample of Cases)
- Výběr vedený výzkumnou otázkou (Select IF)

ÚKOL 1. Ověřte H1: V Lidových novinách vystupuje v roli hlavního aktéra častěji politik než odborník na dané téma.

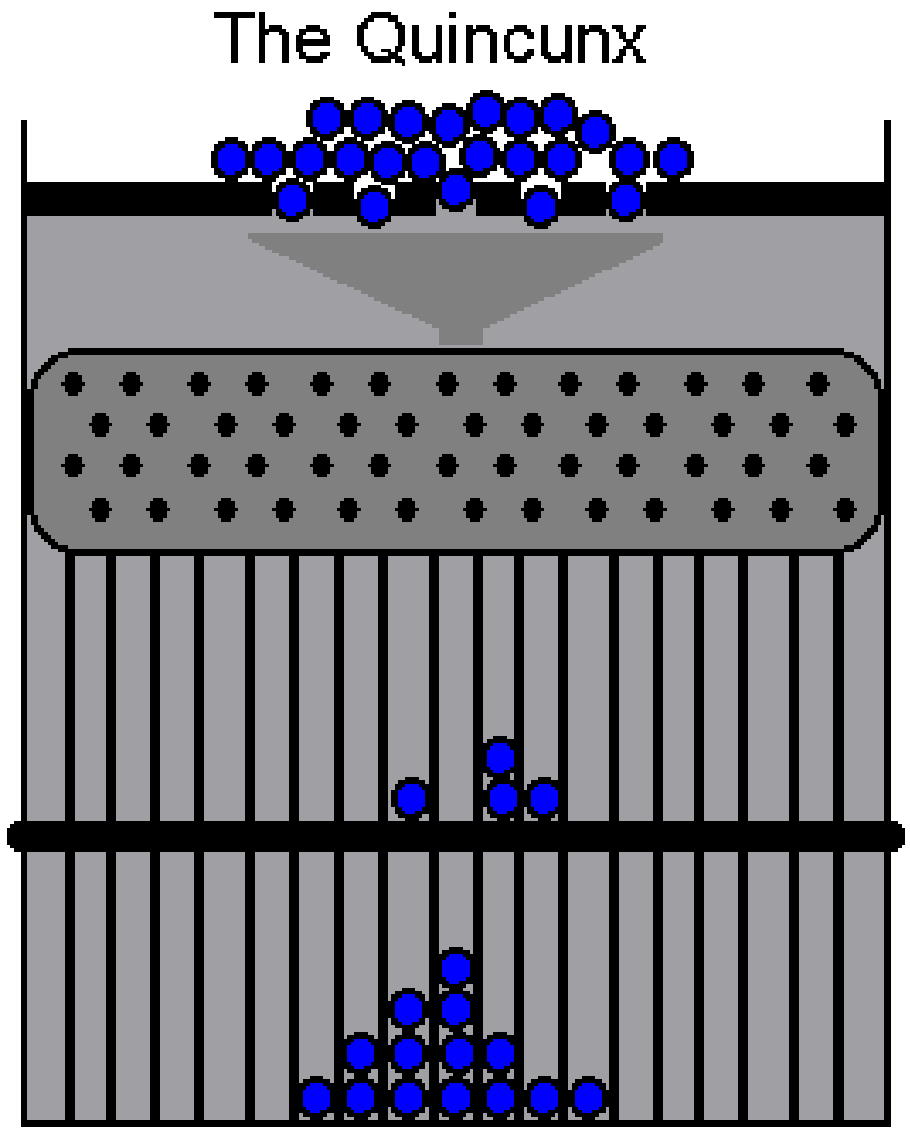
Procedura SPLIT FILE

ÚKOL 2. Porovnejte výskyt odborníků a politiků jako aktérů v jednotlivých denících (H2: MFD bude využívat odborníků častěji než LN.)

NORMÁLNÍ ROZLOŽENÍ (NORMAL DISTRIBUTION; GAUSSOVA KŘIVKA)



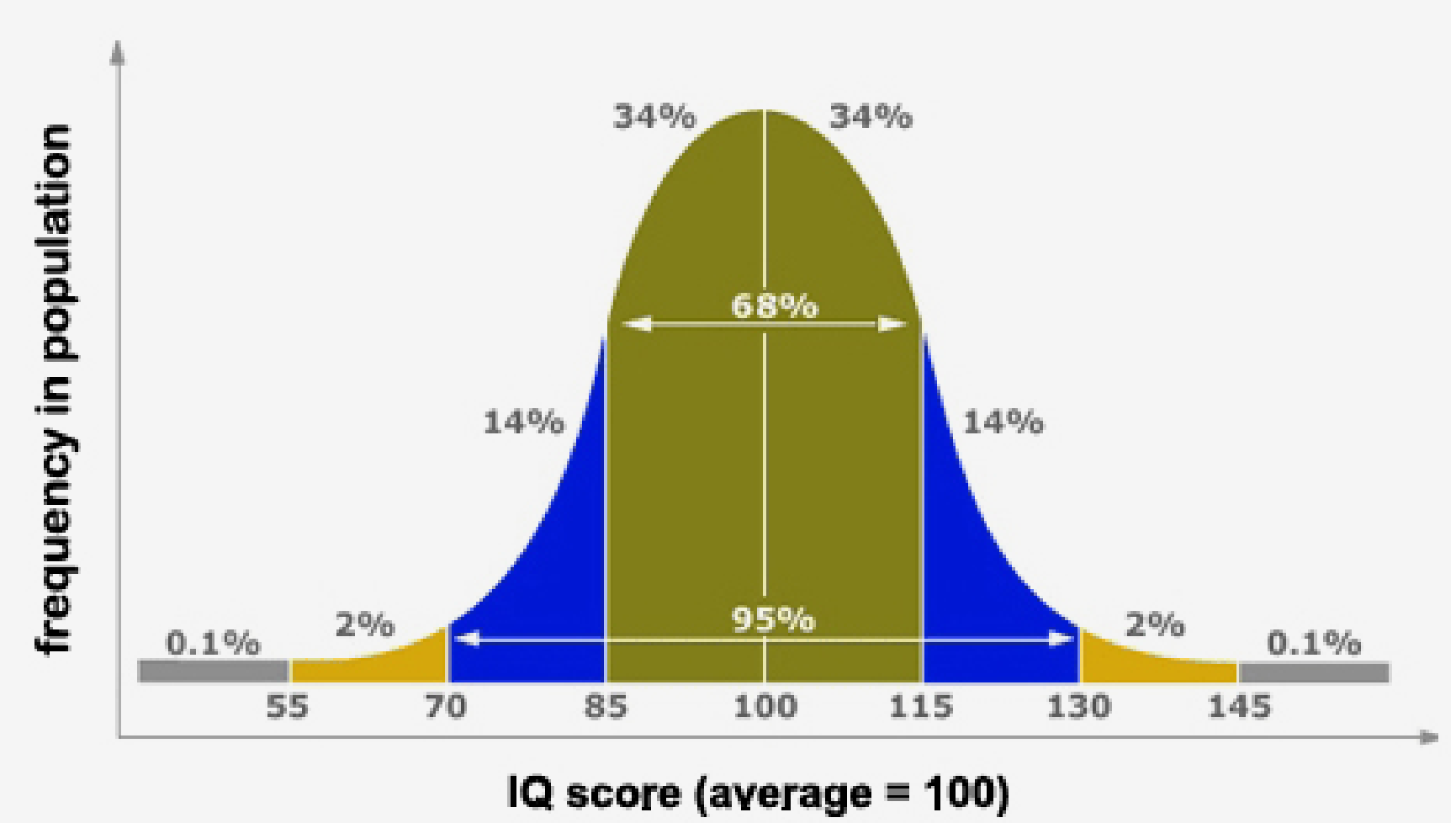
SIR FRANCIS GALTON (1822-1911)



1. Beads are put into the hopper
2. The beads drop through the funnel one by one
3. The beads bounce randomly through a series of pins
4. Each bead drops into a numbered slot
5. The results of one run can be saved and compared with the results of a later trial

https://www.youtube.com/watch?v=AjI_LcQOO

IQ TEST


frequency in population
34%
34%
68%
14%
14%
95%
0.1%
2%
2%
0.1%
55
70
85
100
115
130
145
IQ score (average = 100)

TESTY NORMÁLNÍHO ROZDĚLENÍ

Histogram s křivkou normálního rozdělení

Posouzení šikmosti a špičatosti <0; 1>

- 0 = normální rozdělení
- 1 = rozdělení není normální
- ——————— | | (neplatí pro N > 200)

Kolmogorovův-Smironovovův test (K-S test): Procedura Explore

- H0: Rozložení proměnné se neliší od normálního.
- Kritická hranice sig. > 0,05

Důsledky:

- použití neparametrického testu
- trasnformace proměnných
- Použití parametrického testu

STANDARDIZOVANÉ (NORMOVANÉ) NORMÁLNÍ ROZDĚLENÍ

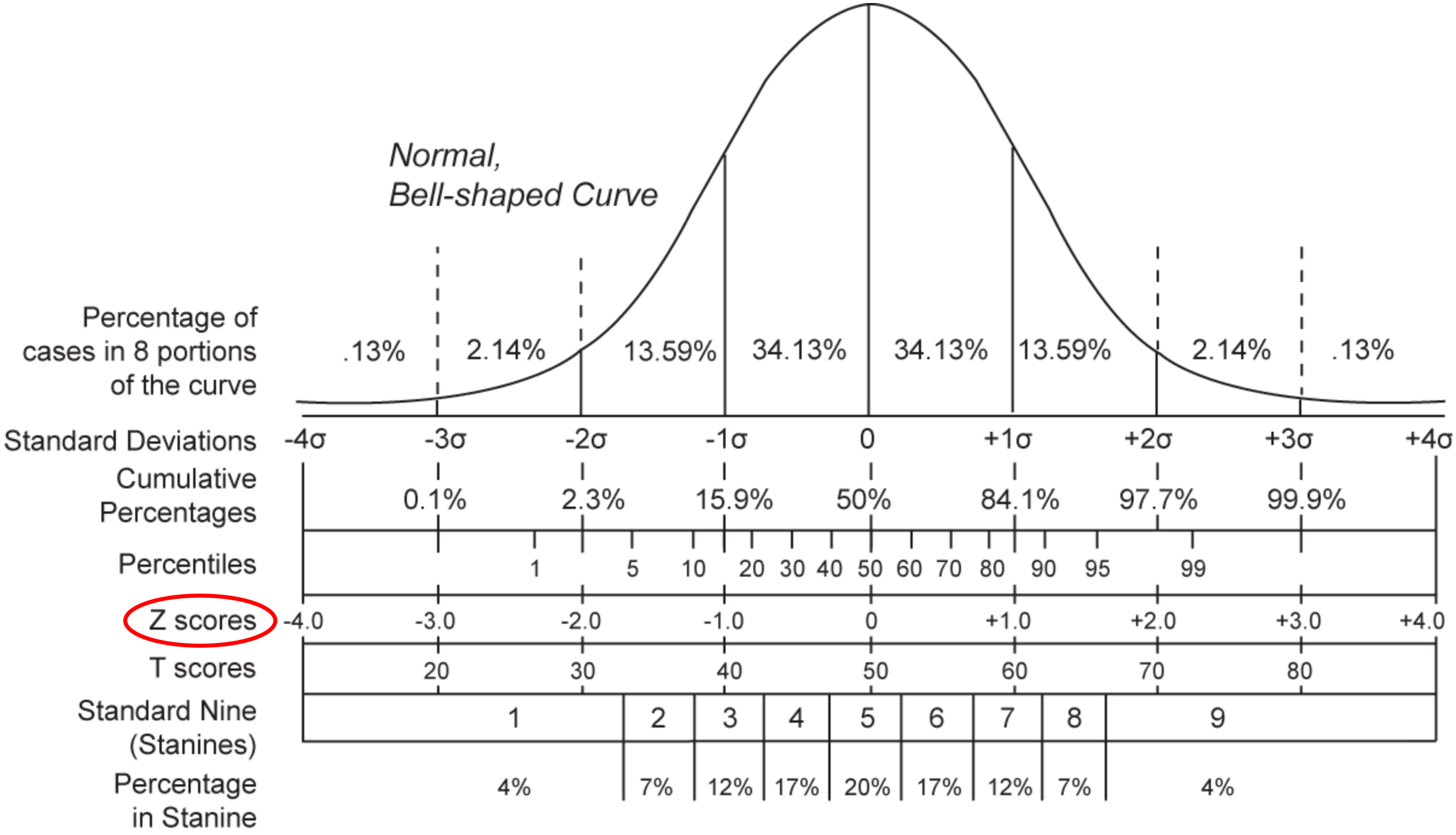
Základem pro inferenční statistiky

Převádí hodnoty na tzv. z-skóry, tzn. vyjadřuje hodnotu jako vzdálenost od průměru ve směrodatných odchylkách

—— (Procedura DESCRIPTIVES)

Užití:

- Umožňuje porovnávat proměnné měřené na různých škálách a v různých jednotkách
- Standardizace dat


Normal,
Bell-shaped Curve
Percentage of cases in 8 portions of the curve
.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
.13%
Standard Deviations
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative Percentages
0.1%
2.3%
15.9%
50%
84.1%
97.7%
99.9%
Percentiles
1
5
10
20
30
40
50
60
70
80
90
95
99
Z scores
-4.0
-3.0
-2.0
-1.0
0
+1.0
+2.0
+3.0
+4.0
T scores
20
30
40
50
60
70
80
Standard Nine (Stanines)
1
2
3
4
5
6
7
8
9
Percentage in Stanine
4%
7%
12%
17%
20%
17%
12%
7%
4%

INFERENČNÍ (VÝBĚROVÁ) STATISTIKA

Inference = statistické usuzování (z výběrového souboru na základní soubor, z parametru na statistiku)

Odpovídají data v našem vzorku populaci? S jakou pravděpodobností?

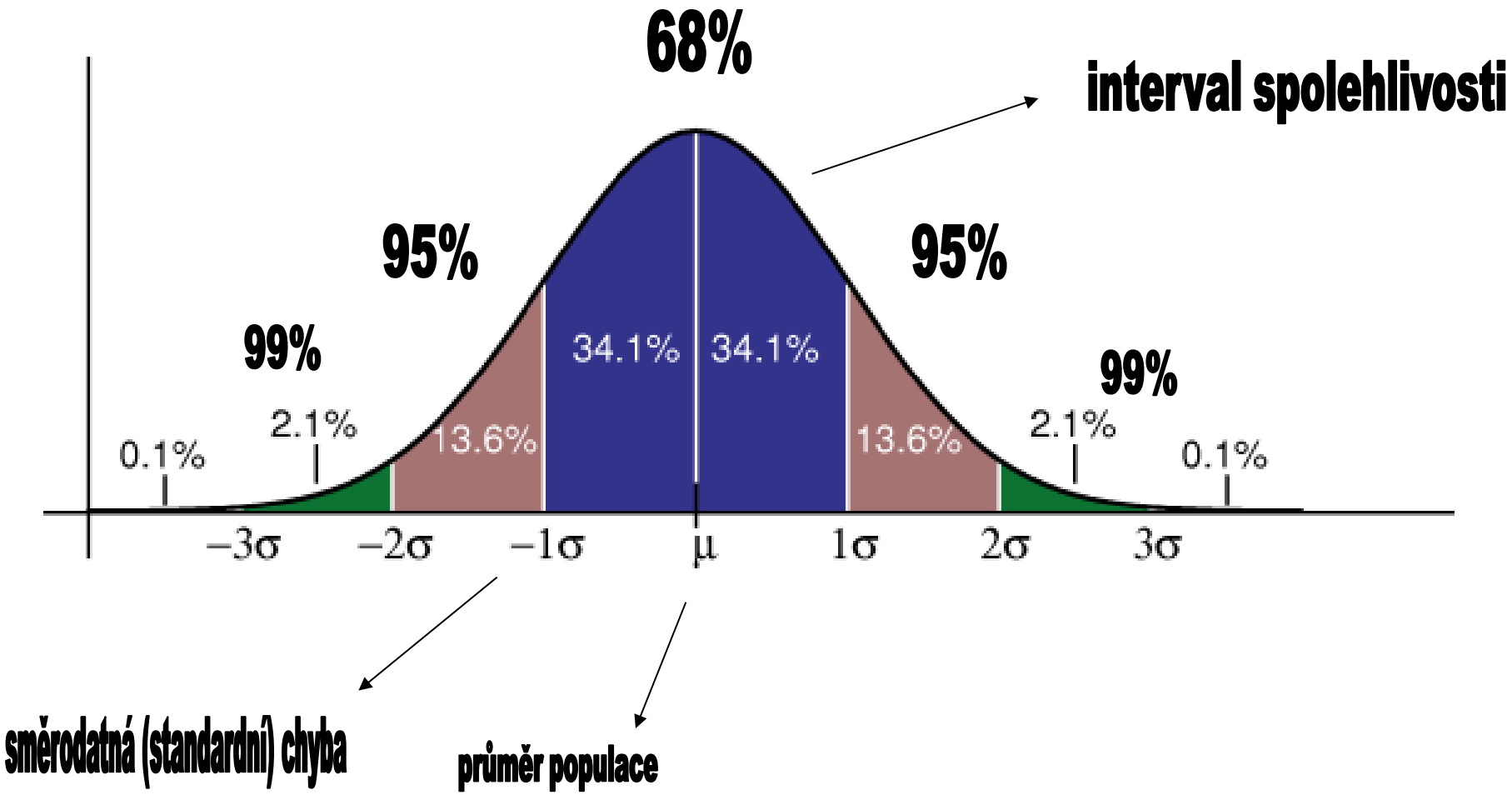
Zahrnuje

- (1) odhad parametrů – metody odhadu pro bodové a intervalové odhady
- (2) testování hypotéz – testy statistické významnosti

Předpoklady:

- Pravděpodobnostní výběr
- Vysoká návratnost (response rate)

TEORETICKÉ NORMÁLNÍM ROZLOŽENÍ (POPULACE)



CENTRÁLNÍ LIMITNÍ VĚTA

Centrální limitní věta (central limit theorem) říká, že když provedeme mnoho výběrů o určité velikosti založených na pravděpodobnostním principu, pak se rozložení (distribuce) výběrových průměrů přiblíží normálními rozdělení a celkový průměr těchto průměrů se bude podobat průměru v populaci. A to nezávisle na tom, jak jsou hodnoty proměnné rozloženy v populaci.

Standardní chyba průměru je pak směrodatnou odchylkou rozdělení tohoto rozdělení průměrů, vyjadřuje výběrovou chybu

STANDARDNÍ CHYBA PRŮMĚRU (STANDARD ERROR OF THE MEAN)

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

← **standardní odchylka**

← **velikost vzorku**

Pro nominální data se používá **standardní chyba**

proporce (standard error of proportion),

$$SE(p) = \sqrt{\frac{p \cdot q}{n}}$$

Výběrová chyba (sample error) na hladině p=0.05

1.96 * SE

Výpočet horního limitu konfidenčního intervalu na 95% hladině významnosti: **CI = m + (SE * 1.96)**

Výpočet spodního limitu konfidenčního intervalu na 95% hladině významnosti: **CI = m – (SE**

INTERVAL SPOLEHLIVOSTI (CONFIDENCE INTERVAL)

Parametry odhadujeme s určitou mírou pravděpodobnosti/rizika, tzv. hladina významnosti

- typicky 95% HV → 5% riziko, příp. 99%HV
- Odvozeno od normálního rozdělení (±4 → 99% případů, ±3 → 95% případů)
- Vyšší míra jistoty vede k širšímu CI

- - výběrový průměr
- - z-skór požadované úrovně pravděpodobnosti, pro je to 1.96
- - standardní směrodatná chyba, SD výběrových průměrů

Výpočet pro průměr: , kde je standardní chyba průměru
(PROCEDURA DESCRIPTIVES)

Výpočet pro relativní četnost (%): , kde p je pozorovaná
relativní četnost

TESTOVÁNÍ HYPOTÉZ

Testujeme, zda:

- Vzorek pochází z populace s určitým rozdělením (reprezentativita)
- Zda dva výběry pocházejí z téže populace (např. rozdíly mezi muži a ženami)
- Zda ne/existuje vztah mezi proměnnými

Nulová hypotéza H_0 – předpokládá neexistenci rozdílu, buď ji zamítnout lze nebo nelze

- Teoretická nulová hypotéza (Např. Neexistuje rozdíl mezi platy žen a mužů.)
- Statistická nulová hypotéza (Např. Rozdíl mezi průměrným platem mužů a průměrným platem žen je roven nule.)

Alternativní hypotéza H_1 – předpokládá rozdíl

- Oboustranná (two-tailed) - Mezi proměnnými pohlaví a příjem bude vztah, příjem se bude lišit.
- Jednostranná (one-tailed) – Ženy budou v průměru vydělávat méně.

URČETE H_0 NEBO H_1

H_0: Respekt a Reflex se nebudou odlišovat v hodnocení (pozitivní/negativní) prezidenta Zemana.

H_0: Mezi mírou prokrastinace a vlastnictvím účtu na Facebooku nebude žádný vztah.

H_1: Děti, jimž rodiče kontrolují telefon budou méně často obětmi kyberšikany.

H_1: Mezi mírou narcismu a počtem zveřejněných fotografií na sociálních sítích bude vztah. .

POSTUP TESTOVÁNÍ

Volba testu

Volba testovacího kritéria se známým rozložením

Výpočet hodnoty testového kritéria (data)

Interpretace výsledku

HV < 0,05 → zamítáme nulovou hypotézu, neboť pravděpodobnost, že bychom získali tako data, kdyby platila $H_{0,}$ je malá

- Sig. v SPSS
- *p* - v odborné literatuře
- Arbitrární charakter HV
- Statistická významnost ≠ věcná významnost

Např. Kolmogorovův-Smirnovovův test normality rozložení, srovnání CI pro dva populační průměry