LEKCE 04a
STATISTICKÁ INFERENCE ANEB ZOBECŇOVÁNÍ VÝSLEDKŮ Z VÝBĚROVÉHO NA ZÁKLADNÍ SOUBOR.

Ve většině případů pracujeme s výběrovým souborem a výběrové výsledky zobecňujeme na základní soubor. Smysluplné je to ale jen:

- Jde-li skutečně o VÝBĚROVÝ SOUBOR (při vyčerpávajícím šetření to nemá smysl).
- Jde-li o NÁHODNÝ VÝBĚR kdy každá jednotka dané populace má stejnou pravděpodobnost, že bude vybrána.
- Jde-li o NEZÁVISLÝ VÝBĚR (výběr žádné jednotky nezvyšuje ani nesnižuje pravděpodobnost výběru jiných jednotek).

  Příklad: Opisují-li studenti v testu, jejich výsledky nejsou nezávislé).

Tak jako se při měření musíme vyrovnat s měřícími chybami, musíme se při inferenci vyrovnat s výběrovou chybou (výběr je jen částí základního souboru).

ZÁKLADNÍ OTÁZKA: JAK JSOU POZOROVANÉ VÝSLEDKY PRAVDĚPODOBNÉ?

Pozorovaný výsledek představuje

- STATISTIKU (jak je to v našem výběrovém souboru),

z níž usuzujeme na

- PARAMETR (jak je to v populaci, z níž byl soubor vybrán).

PARAMETR

Neznámá (pokud nemáme vyčerpávající šetření) vlastnost základního souboru

- μ = průměr základního souboru
- σ = standardní odchylka základního souboru
- σ^2 = variance základního souboru

PARAMETRY ZÁKLADNÍHO SOUBORU obvykle neznáme, ale můžeme je odhadovat z VÝBĚROVÝCH STATISTIK.

STATISTIKA

Známá vlastnost výběrového souboru

- $\bar{X}$ = průměr výběrového souboru
- s = standardní odchylka výběrového souboru
- s^2 = variance výběrového souboru

Smysl otázky "JAK JSOU POZOROVANÉ VÝSLEDKY PRAVDĚPODOBNÉ?":

- Lze, se zvolenou pravděpodobností předpokládat, že STATISTIKA jakožto pozorovaný výsledek reprezentuje nepozorovatelný PARAMETR?
- Není STATISTIKA v důsledku výběrové chyby přece jen příliš vzdálená PARAMETRU?
- V jakém intervalu kolem STATISTIKY můžeme s danou pravděpodobností očekávat výskyt PARAMETRU?

INFERENCE ZE STATISTIKY NA PARAMETR

- BODOVÝ ODHAD jako číslo, jehož hodnota je v nějakém (teoreticky) stanoveném smyslu optimálně určena.
- INTERVALOVÝ ODHAD, kdy hledáme interval (spolehlivosti), v kterém s určitou, předem zvolenou pravděpodobností neznámý populační parametr leží.

NA PŘEDCHOZÍ OTÁZKY LZE ODPOVĚDĚT DÍKY VLASTNOSTEM NORMÁLNÍHO RESPEKTIVE STANDARDIZOVANÉHO NORMÁLNÍHO ROZLOŽENÍ.

STANDARDNÍ ODCHYLKA VE VÝBĚROVÉM SOUBORU

$$\sigma = \sqrt{\frac{\sum (\bar{x}_i - x)^2}{N}}$$

STANDARDNÍ ODCHYLKA V ZÁKLADNÍM SOUBORU

$$\sigma = \sqrt{\frac{\sum (\bar{x}_i - \mu)^2}{M}}$$

STANDARDNÍ CHYBA PRŮMĚRU

$$\sigma = \sqrt{\frac{\sum (\bar{x}_i - \mu)^2}{n_s}}$$

μ ← populační průměr

n_s ← počet provedených výběrů

$\bar{x}_i$ ← průměr z provedených výběrů

Příklad různých náhodných výběrů

VÝBĚROVÉ SOUBORY	průměr	std. odchylka
1. výběr (N = 892)	56,5	13,35
2. výběr (N = 892)	56,8	13,52
3. výběr (N = 892)	56,5	13,34
4. výběr (N = 892)	56,5	13,26
5. výběr (N = 892)	56,7	13,33
PRŮMĚR	56,6	13,36
ZÁKLADNÍ SOUBOR (N=1191)	56,4	13,33
ROZDÍL (při 5 výběrech)	0,2	0,03

PROČ JE STANDARDNÍ/SMĚRODATNÁ CHYBA PRŮMĚRU DŮLEŽITÁ?

S 95% pravděpodobností (5% riziko chyby) můžeme tvrdit, že:

průměr základního souboru (parametr)
=
průměr výběrového souboru (statistika)
± 1,96 směrodatná chyby
(často se zaokrouhluje na dvojnásobek)

S 99% pravděpodobností (1% riziko chyby) můžeme tvrdit, že:

průměr základního souboru (parametr)
=
průměr výběrového souboru (statistika)
± 2,96 směrodatná chyby
(často se zaokrouhluje na trojnásobek)

DOSTÁVÁME SE K POJMU INTERVAL SPOLEHLIVOSTI

Protože pracujeme s výběrovými soubory, můžeme vypočítat statistiky, ale nevíme, jak tyto statistiky korespondují s parametry. Víme ovšem, že se - se zvolenou pravděpodobností - pohybují v intervalu (spolehlivosti), jehož obecný vzorec je:

$$C.I = \bar{X} \pm z . \sigma_X$$

$\bar{X}$ = vypočítaný výběrový průměr (statistika)
z = z-skóre korespondující s požadovanou úrovní pravděpodobnosti (hladinou významnosti). Pro HV=95% je to 1,96.
σ_X = standardní/směrodatná chyba distribuce výběrových průměrů

Interval spolehlivosti pro 95% HV znamená:
Jestliže bychom z populace opakovaně činili výběry stejné velikosti, v 95% z nich výběrů by se populační průměr nacházel uvnitř intervalu spolehlivosti (s 95% pravděpodobnost interval spolehlivosti tento populační průměr zahrnuje).

INTERVAL SPOLEHLIVOSTI (pro průměr na HV = 95%)

$$C.I._{95\%} = \bar{X} \pm 1,96 . s/\sqrt{N}$$

s = standardní/směrodatná chyba

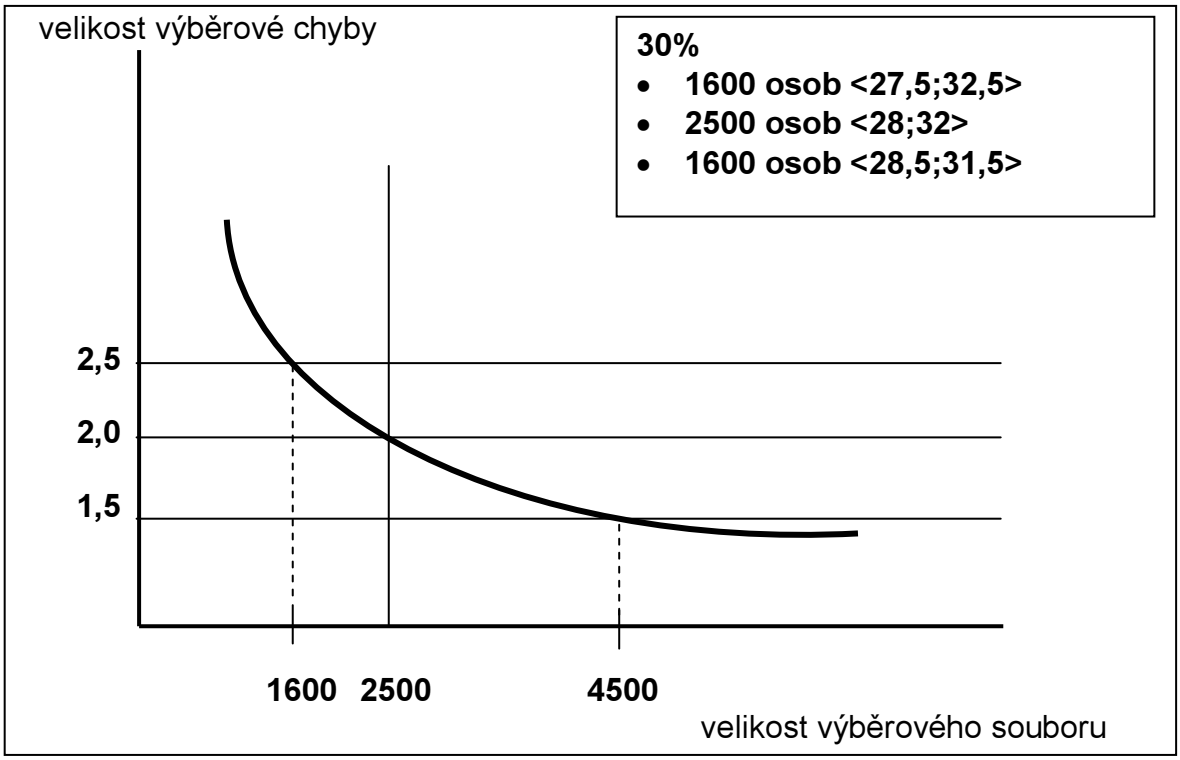
INTERVAL SPOLEHLIVOSTI (pro procento výskytu na HV = 95%)

$$C.I._{95\%} = p \pm 1,96 . \sqrt{p.(1-p) / N}$$

- p = pozorovaný podíl, kolem něhož je interval spolehlivosti konstruován
- N = velikost výběrového souboru

VELIKOST VÝBĚROVÉHO SOUBORU A VELIKOST VÝBĚROVÉ CHYBY

□ VÝBĚROVÁ CHYBA sice s velikostí vzorku klesá, ale po dosažení určité velikosti souboru je její zmenšování s dalším zvětšováním výběru nepodstatné. Proto není další růst početnosti výběrového souboru ekonomický.



□ Stanovíme-li si přípustnou výběrovou chybu (nakolik se, s jistou zvolenou pravděpodobností, mohou výsledky zjištěné ve výběrovém souboru odchylovat od skutečnosti v základním souboru), můžeme určit potřebnou velikost výběrového souboru. A to s přihlédnutím k homogenitě základního souboru z hlediska vlastností, které nás zajímají (usuzujeme na ni z velikosti rozptylu).

VELIKOST VÝBĚROVÉHO SOUBORU A VÝBĚROVÁ CHYBA (NA HV=95%)

výběrová chyba v %	interval spolehlivosti	velikost výběru (sample size)	výběrová chyba v %	interval spolehlivosti	velikost výběru (sample size)
1,0	±1,0	10000	6,0	±6,0	277
1,5	±1,5	4500	6,5	±6,5	237
2,0	±2,0	2500	7,0	±7,0	204
2,5	±2,5	1600	7,5	±7,5	178
3,0	±3,0	1100	8,0	±8,0	156
3,5	±3,5	816	8,5	±8,5	138
4,0	±4,0	625	9,0	±9,0	123
4,5	±4,5	494	9,5	±9,5	110
5,0	±5,0	400	10,0	±10,0	100
5,5	±5,5	330			

- Výběrová chyba (Sampling Error). Pro 95% hladinu významnosti (Confidence Level) de facto dvě standardní chyby
- Interval spolehlivosti (Confidence Interval) vymezují Confidence Limits) - jeho hraniční hodnoty.
- V případě tabelovaných hodnot se předpokládá heterogenní soubor (50%:50%) - platí pro alternativní neboli binomické proměnné. Vezmeme-li v úvahu heterogenitu souboru, je výpočet intervalu spolehlivosti složitější

VÝBĚROVÁ CHYBA V ZÁVISLOSTI NA HOMOGENITĚ VÝBĚROVÉHO SOUBORU

- V každém sloupci výběrové chyby pro příslušná procenta sledované vlastnosti ve výběrovém souboru
- V každém řádku výběrové chyby pro danou velikost výběrového souboru

velikost výběru	1% nebo 99%	5% nebo 95%	10% nebo 90%	15% nebo 85%	20% nebo 80%	25% nebo 75%	30% nebo 70%	35% nebo 65%	40% nebo 60%	45% nebo 55%	50%
25	4,0	8,7	12,0	14,3	16,0	17,3	18,3	19,1	19,6	19,8	20,0
50	2,8	6,2	8,5	10,1	11,4	12,3	13,0	13,5	13,9	14,1	14,2
75	2,3	5,0	6,9	8,2	9,2	10,0	10,5	11,0	11,3	11,4	11,5
100	2,0	4,4	6,0	7,1	8,0	8,7	9,2	9,5	9,8	9,9	10,0
150	1,6	3,6	4,9	5,9	6,6	7,1	7,5	7,8	8,0	8,1	8,2
200	1,4	3,1	4,3	5,1	5,7	6,1	6,5	6,8	7,0	7,0	7,1
250	1,2	2,7	3,8	4,5	5,0	5,5	5,8	6,0	6,2	6,2	6,3
300	1,1	2,5	3,5	4,1	4,6	5,0	5,3	5,5	5,7	5,8	5,8
400	0,99	2,2	3,0	3,6	4,0	4,3	4,6	4,8	4,9	5,0	5,0
500	0,89	2,0	2,7	3,2	3,6	3,9	4,1	4,3	4,4	4,5	4,5
600	0,81	1,8	2,5	2,9	3,3	3,6	3,8	3,9	4,0	4,1	4,1
800	0,69	1,5	2,1	2,5	2,8	3,0	3,2	3,3	3,4	3,5	3,5
1000	0,63	1,4	1,9	2,3	2,6	2,8	2,9	3,1	3,1	3,2	3,2
2000	0,44	0,96	1,3	1,6	1,8	1,9	2,0	2,1	2,2	2,2	2,2
5000	0,28	0,62	0,85	1,0	1,1	1,2	1,3	1,4	1,4	1,4	1,4

Pramen: A Broadcast Research Primer. National Association of Broadcasters, Washington, DC 1976, p. 19.

LEKCE 04b
ZÁKLADY TESTOVÁNÍ HYPOTÉZ

STATISTICKÉ HYPOTÉZY

neboli formální výroky o: neznámých parametrech základního souboru, o tvaru rozložení četností, o statistických vztazích mezi soubory či proměnnými v něm….
TESTOVÁNÍ směřuje k zobecnění dat výběrového souboru na základní soubor.
Jinak řečeno: Statistické hypotézy jsou domněnkami o populaci, jejichž pravdivost ověřujeme (testujeme) pomocí výběrových souborů z této populace.

OBVYKLE SE TESTUJE ZDA

- Zkoumaný výběrový soubor pochází ze základního souboru s určitým rozdělením (zda je výběr reprezentativní).
  - Jak se odchyluje věkový průměr ve výběrovém souboru od známého věkového průměru populace.
  - Jak se odchyluje struktura volebních preferencí ve výběrovém souboru od známé struktury těchto preferencí v populaci.
- Dva výběry pocházejí ze (stejného) základního souboru s určitým rozdělením.
  Liší se průměrné mzdy žen a mužů tak, že to nemůže být vysvětleno náhodou?
- Zda je možno považovat studovaný soubor za náhodně uspořádaný (zda mezi proměnnými neexistuje žádný vztah).
  Například distribuci proměnné lze považovat za náhodně uspořádánu, jestliže jsou všechny její kategorie stejně početné.
- Jak se hodnota odchyluje od určitého standardu
  - Jak se odchyluje průměrná pracovní doba od zákonem stanovené délky pracovní doby.
  - Jak se odchyluje vzdělanostní struktura čtenářů časopisu RESPEKT od vzdělanostní struktury populace.
  - Jak se odchyluje průměrné IQ ve skupině delikventů od 100 bodů.

NULOVÁ HYPOTÉZA

Obvykle se testuje NULOVÁ HYPOTÉZA (H_0) jako specifický model statistické hypotézy. NULOVÁ HYPOTÉZA PŘEDPOKLÁDÁ STAV „NEEXISTENCE“ (ROZDÍLU) ČI STAV SHODY.
PŘÍKLADY NULOVÝCH HYPOTÉZ

- Rozložení hodnot znaku se neliší od nějakého teoretického rozložení (například normálního nebo náhodného).
- Rozložení četností hodnot proměnné (vlastností jednotky), např. příjmu, věku, míry anomie, spokojenosti v životě, …, ve výběrovém souboru odpovídá rozložení proměnné v populaci (neliší se od něho).
- Mezi dvěma parametry, např. mezi průměrným příjmem mužů a žen, není v základním souboru rozdíl (usuzujeme na to ze zjištěných statistik ve výběrovém souboru).
- Mezi empirickým a náhodným rozložením hodnot v kontingenční tabulce není rozdílu (empirické rozložení je náhodné, neexistuje vztah nejen mezi 2 proměnnými, které tabulku tvoří, ale ani mezi jejich variantami).

Hypotéza se zamítá:
Hypotézy lze zásadně prohlásit za falešné (tedy zamítnout jejich platnost), nikoliv však dokazovat jejich platnost. Hypotéza nemůže být přímo dokázána, nýbrž může být jen zamítnuta jí odporující (nulová) hypotéza.

DVA VÝSLEDKY TESTOVÁNÍ H_0

- **NEMÁME DŮVOD ZAMÍTNOUT MODEL NULOVÉ HYPOTÉZY A PROTO JI PŘIJÍMÁME.**
  Příklady:
  - Rozdíl mezi dvěma populačními průměry neexistuje.
  - Rozdíl mezi dvěma populačními průměry, existuje, ale je tak malý, že ho nemůžeme určit.

    Například rozdíl 10 Kč u průměrných ročních příjmů mužů a žen, nebo 1 bod u průměrného IQ dvou skupin ap.

- **DATA NEODPOVÍDAJÍ H_0 (jejich existence je při platnosti H_0 vysoce nepravděpodobná) A PROTO JI ZAMÍTÁME. JEJÍ ZAMÍTNUTÍ VŠAK VĚTŠINOU NESTAČÍ PRO PŘIJETÍ ALTERNATIVNÍ HYPOTÉZY.**

  Příklady alternativních hypotéz:
  1. **Nulová hypotéza:** *Rozložení příjmů ve výběrovém a základním souboru jsou shodné (odmítnutí H_0 de facto znamená prokázání, že výběr není náhodný respektive rozdíl může být způsoben výběrovou chybou).*
     Alternativní hypotézy (directional hypotheses):
     - *Výběrový soubor má v průměru nižší příjmy než základní.*
     - *Výběrový soubor má v průměru vyšší příjmy než základní.*

     Alternativní hypotéza (non-directional hypothesis):
     - *Výběrový soubor má v průměru vyšší nebo nižší příjmy než soubor základní.*

  2. **Nulová hypotéza:** *Mezi vzděláním a výší příjmu není žádný vztah (v základním souboru)***.**
     Alternativní hypotézy (directional hypotheses):
     - *Čím vyšší vzdělání, tím vyšší příjem.*
     - *Čím vyšší vzdělání, tím nižší příjem.*

     Alternativní hypotéza (non-directional hypothesis):
     - *Se změnou vzdělání se mění i výše příjmu.*

HLADINA VÝZNAMNOSTI (significance level)

Nazývá se tak pravděpodobnost, že náhodná odchylka (daná výběrovou chybou) překročí určitou danou hodnotu, nazývanou hranice významnosti či KRITICKÁ HODNOTA. Představuje velikost rizika chyby, jež připustíme.

Zjištěné (empirické) odchylky, vyskytující se s pravděpodobností MENŠÍ NEŽ JE ZVOLENÁ HLADINA VÝZNAMNOSTI (HV), se nazývají STATISTICKY VÝZNAMNÉ (signifikantní) na této zvolené hladině.

TESTOVACÍ KRITÉRIUM

Každému testovacímu kritériu PŘÍSLUŠÍ TEORETICKÉ ROZDĚLENÍ (normální rozložení, t neboli Studentovo rozložení, F rozložení, ...).

Tabelovány bývají jeho KRITICKÉ HODNOTY.
Hodnoty, jež příslušná náhodná veličina překročí s určitou danou pravděpodobností, tj. na určité hladině významnosti (vyčteno z teoretického rozložení testovacího kritéria např. existuje jen 5% pravděpodobnost výskytu hodnot větších než kritická).

Základem TESTOVÁNÍ je porovnávání vypočítané (empirické) hodnoty testovacího kritéria (hodnota t, hodnota F, hodnota chí-kvadrát, …) , s tabelovanými kritickými hodnotami.

POSTUP TESTOVÁNÍ

- Zvolíme vhodné TESTOVACÍ KRITÉRIUM, jehož teoretické rozložení (standardizované normální rozložení, Studentovo rozložení, rozložení chí-kvadrátu…).
- Vypočítáme z dat výběrového souboru jeho empirickou hodnotu (z-skóre jemuž odpovídá standardizované normální rozložení, t hodnotu jíž odpovídá Studentovo rozložení, chí-kvadrát jemuž odpovídá rozložení chí-kvadrát …).
- Porovnáme vypočítanou statistiku s jejím teoretickým rozložením - s její KRITICKOU HODNOTOU (T*).

- Je-li vypočítaná hodnota testovacího kritéria menší než hodnota kritická ($T < T^*$), je to případ, který je při platnosti H_0 natolik pravděpodobný, že existující odchylka může být považována za náhodu. H_0 nezamítáme a tvrdíme, že ROZDÍL NENÍ STATISTICKY VÝZNAMNÝ.
- Je-li vypočítaná hodnota testovacího kritéria větší než kritická ($T \geq T^*$), je to případ, který je při platnosti H_0 tak málo pravděpodobný, že je takřka nemožný. H_0 zamítáme a tvrdíme, že ROZDÍL JE STATISTICKY VÝZNAMNÝ.