Etapy stat.šetření

1. Plán šetření
2. Sběr dat
3. Popis a technické zpracování, tzv. **deskriptivní statistika**
4. Rozbor, závěry, interpretace, tzv. **induktivní statistika**

Statistická indukce – 4. etapa

Teorie odhadů

- Odhad průměru základního souboru
- Odhad pravděpodobnosti

Testování statistických hypotéz

- Srovnání dvou průměrů
- Srovnání pravděpodobností

Hodnocení závislosti

- Závislost kvantitativních veličin
- Závislost kvalitativních veličin

ZOBECNĚNÍ VÝSLEDKŮ VÝBĚROVÉHO ŠETŘENÍ NA CELÝ ZÁKLADNÍ SOUBOR

! PRAVDĚPODOBNOST, NÁHODNÁ VELIČINA A JEJÍ ROZDĚLENÍ

Pravděpodobnost

- Všechny **STAT výroky** – pravděpodobnostní charakter .

- Jejich věrohodnost vyjadřujeme
  - pstí, že tento výrok platí → **SPOLEHLIVOST**
  - pstí, že daný výrok neplatí → **RIZIKO**

- Dva základní pojmy teorie pravděpodobnosti – **náhodný jev** a **pravděpodobnost.**

- **Pravděpodobnost náhodného jevu** – kvantitativní charakteristika, která je mírou častosti výskytu zkoumaného jevu.

Pravděpodobnost

PST někt.náhodných jevů umíme ***vypočítat****, např. výhra ve Sportce, karetní hry ; vznik teorie pravděpodobnosti vázán na* ***hazardní hry*** *→ později užití v jiných oblastech lidské činnosti*

- **Klasická definice** **P (A) = m/n**

  m = počet příznivých výsledků v experimentu

  n = počet všech možných výsledků

  Příklady: kostka, mince

  karty

  sportka

  narození chlapce, dívky

- **Komplex podmínek** – souhrn předpisů, za nichž se experiment provádí.

Pravděpodobnost

- u většiny jevů v lékařské praxi PST ***neznáme*** (např.výskyt onemocnění v populaci, pooperační komplikace, účinek léku)

  → **pouze odhadujeme –** pomocí **RELATIVNÍ ČETNOSTI,**

  kt. zjišťujeme opakováním ***pokusu*** n. ***pozorování + měříme:***

  m/n *m = počet příznivých výsledků v experimentu*

  n = počet všech možných výsledků

→ blíží se skutečné PSTI tím více, čím větší počet pokusů a pozorování provedeme

Pravděpodobnost

Pravděpodobnost reakce pacienta na určitou léčbu?

Pouze **odhad** pomocí **relativních četností**

n pacientů

x vyléčeno

Podíl x/n ***odhaduje pravděpodobnost vyléčení P***

Vlastnosti pravděpodobnosti

1. 0 < nebo = P (A) < nebo = 1
2. P (A) = 0 pro jevy nemožné
3. P (A) =1 pro jevy jisté

Pravidla pro počítání

a) **Pravidlo pro sčítání**

A, B disjunktní – vzájemně se vylučují

P (A nebo B) = P (A) + (B)

A, B nejsou disjunktní

P (A nebo B) = P (A) + P (B) – P (A i B)

Vlastnosti pravděpodobnosti

b) **Pravidlo pro násobení**

A, B nezávislé

P (A i B) = P (A) . P (B)

A, B závislé

P (A i B) = P (A) . P (B/A)

= P (B) . P (A/B) př. dvě nemoci u člověka

P (A/B), P (B/A) jsou tzv. **podmíněné pravděpodobnosti.**

Platí: P (A/B) = P (A i B)/ P (B)

P (B/A) = P (A i B)/P (A)

Pravidlo pro sčítání i násobení se dá rozšířit na více jevů.

Podmíněné pravděpodobnosti jsou užitečné pro hodnocení **rizika** nemoci v populaci

Vlastnosti pravděpodobnosti

Např. Ze srovnání pravděpodobností

P (Ca), P (Ca/K), P(Ca/N)

Lze usuzovat na riziko kouření (K,N) na výskyt karcinomu plic (Ca)

Např. P (Ca/K)/P (Ca) udává, kolikrát je větší pravděpodobnost výskytu karcinomu plic u kuřáků než v celé populaci

P (Ca/K)/P (Ca/N) udává, kolikrát je větší pravděpodobnost výskytu karcinomu plic u kuřáků než u nekuřáků

Výběrový/základní soubor

Výběrový soubor

- reprez. náhodný výběr
- výběrové (empirické) rozdělení četností
- popis rozdělení: tabulka, graf
- stat. ukazatele = výběrové charakteristiky: **m, s, p** (ozn. latinkou)
- jsou to charakteristiky náhodných veličin, tzn. mění se výběr od výběru + je nutné počítat s chybami (výběrové, náhodné)

Základní soubor

- soubor, který nás zajímá
- teoretické rozdělení četností (matematický model)
- popis rozdělení: pravděpodobnostní rozdělení
- stat. ukazatele = parametry: **μ,σ, π** (ozn. řeckou abecedou)
- jsou to konstanty, zpravidla neznámé, pro $n \to \infty$ platí, že $m \to \mu$, $s \to \sigma$, $p \to \pi$

<u>**Statistická indukce**</u> = *usuzování z vlastností výběru na vlastnosti základního souboru*

Empirické a pravděpodobnostní rozdělení

→ každá veličina, kterou zkoumáme, je ovlivněna řadou nepatrných **náhodných vlivů**, což způsobuje její **variabilitu** – tzn. veličina nabývá u různých subjektů různých hodnot.

- měříme-li veličinu **ve výběrovém souboru**, pak rozložení hodnot této veličiny znázorňujeme na základě **empiricky** zjištěných četností

- každá veličina má své **pravděpodobnostní (teoretické) rozdělení**

  Pravděpodobnostní rozdělení (pravděpodobnostní křivka) vyjadřuje očekávání, jak často se budou jednotlivé hodnoty vyskytovat v nekonečně velkém souboru

Empirické a pravděpodobnostní rozdělení

- **Pravděpodobnostní (teoretické) rozdělení**

  v takovém rozložení jsou na ose x všechny hodnoty, kterých může veličina potenciálně nabývat a na ose y jsou zaneseny pravděpodobnosti, se kterými se dané hodnoty vyskytují

X

- v **empirickém rozdělení** (polygon četností) jsou popsány četnosti, se kterými se naměřené hodnoty vyskytovaly ve výběrovém souboru

Typy pravděpodobnostních rozdělení

- přírodním i společenským jevům vlastní *různorodost* a *mnohotvárnost* X v praxi lze vystačit s poměrně malým množstvím modelů.

Pozn.

- *s veličinou zacházíme jako s normálně rozdělenou, pokud nemáme dostatečné důvody pro vyvrácení této domněnky*
- *rozložení většiny veličin lze převést na normální rozdělení*

Diskrétní veličiny

- binomické rozdělení (jev – nejev)
- rovnoměrné rozdělení
- Poissnovo rozdělení (vzácné jevy)

Spojité veličiny

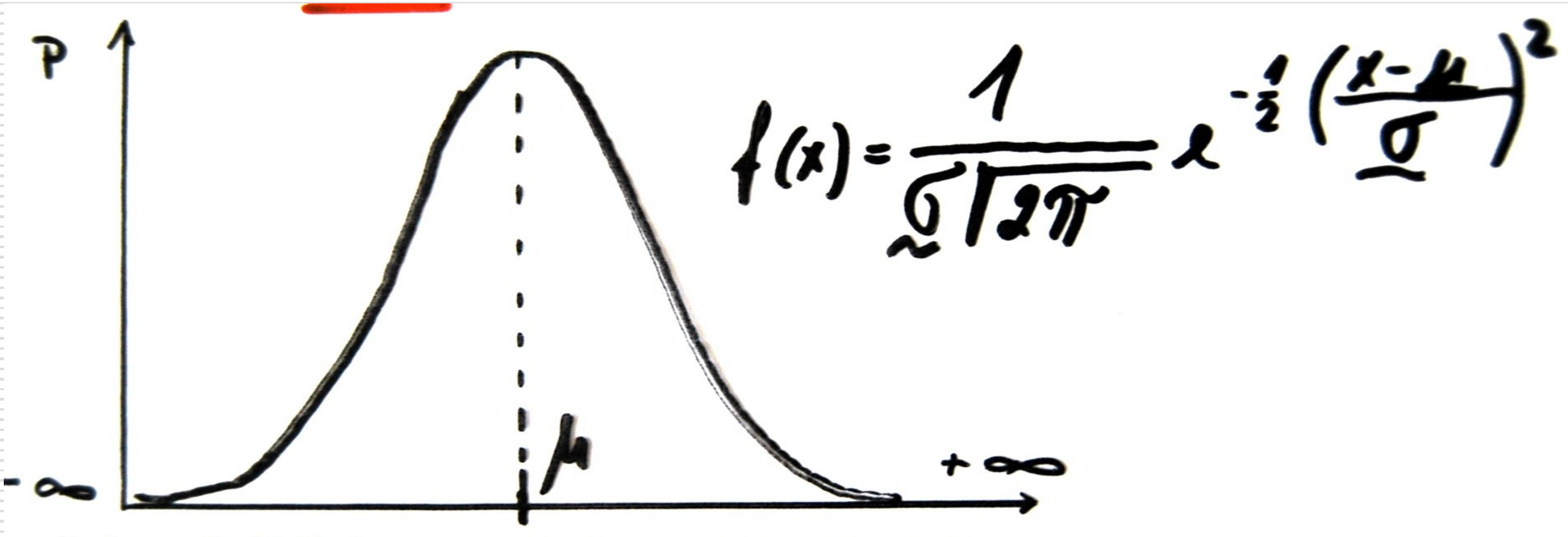
- normální rozdělení
- Studentovo t-rozdělení
- Snedecorovo F-rozdělení
- Chí-kvadrát rozdělení

Normální rozdělení

Spojitý znak → ***normální rozdělení*** – je-li vytvářen nahromaděním velkého počtu nepatrných, nezávislých příčin nahodilého charakteru

Normální rozdělení

- matematický model rozdělení četností **náhodné veličiny**



- frekvenční křivka normálního rozdělení je jednoznačně určena dvěma parametry: **μ, σ**

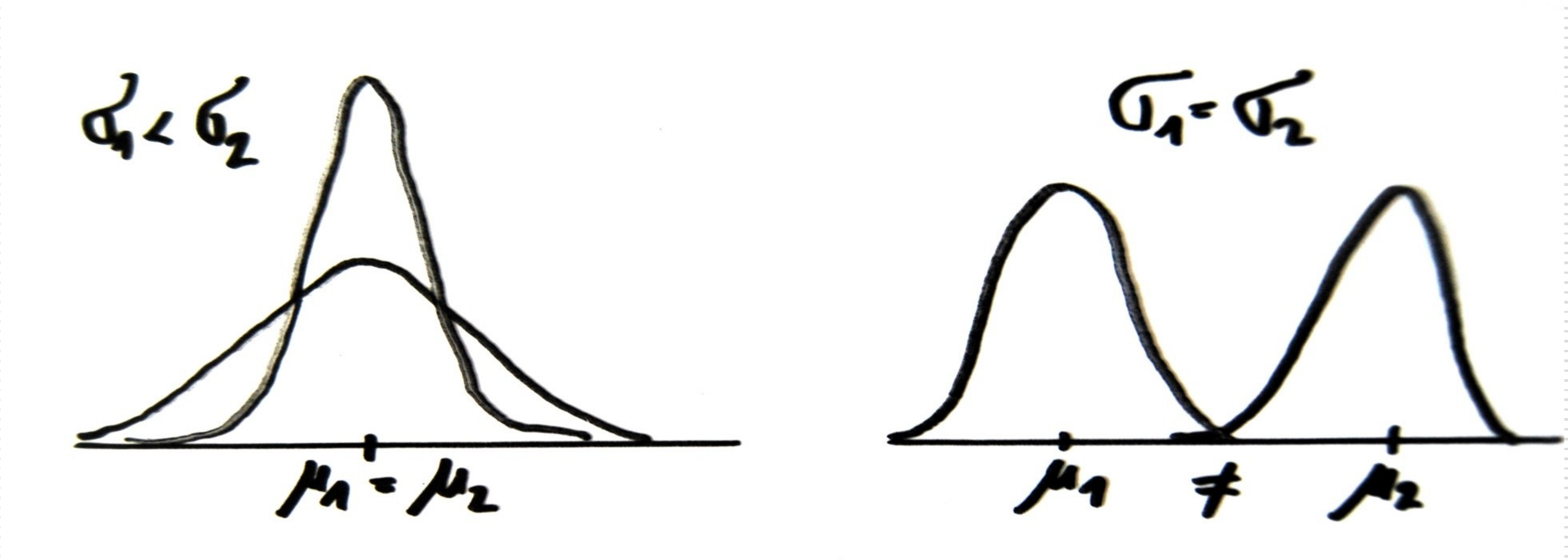
  μ určuje polohu křivky (analogie **m**) **mí**

  σ určuje tvar křivky (analogie **s**) **sigma**

Normální rozdělení

μ určuje polohu křivky (analogie **m**) **mí**

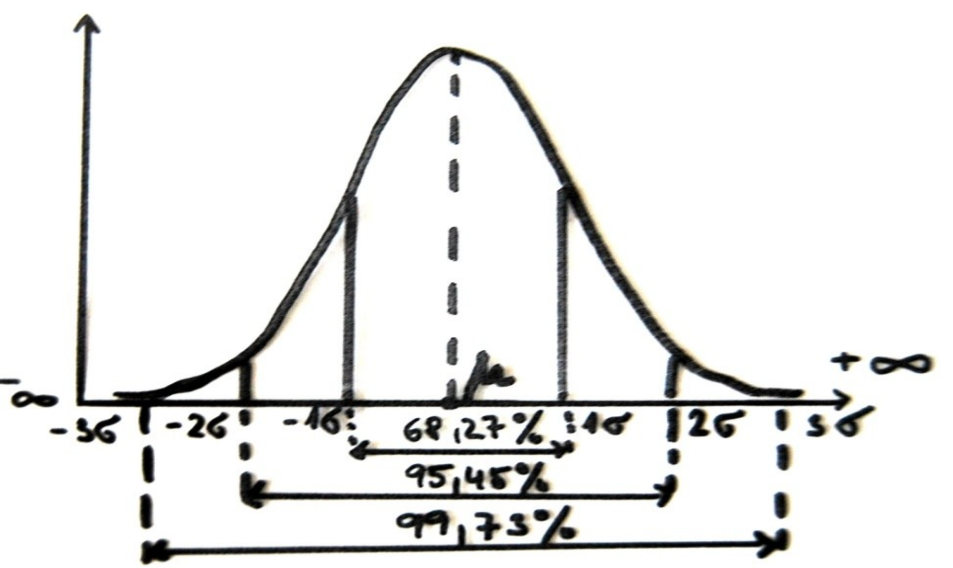
σ určuje tvar křivky (analogie **s**) **sigma**

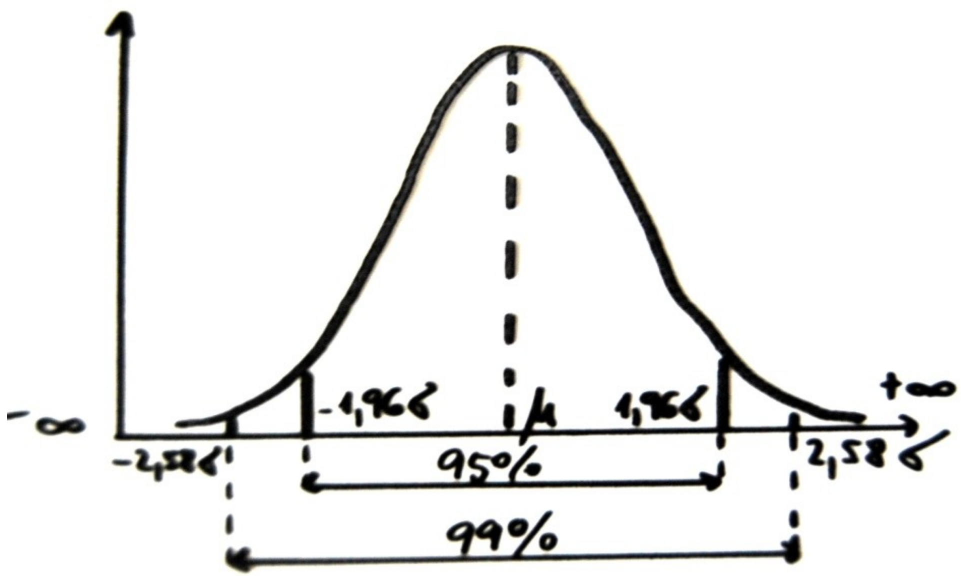


Vlastnosti normálního rozložení

- Frekvenční křivky normálního rozložení mají pro různé veličiny různý tvar (**σ**) a polohu (**μ**)

- Pro všechny ale platí, že intervaly, ve kterých se odhadovaná proměnná nachází s <u>pravděpodobností 95 nebo 99%, lze vyjádřit jako odchylky od **μ** v násobcích **σ** </u>:





Odhady parametrů (1)

1)Bodové odhady

$$\mu \doteq m$$
$$\pi \doteq p$$
$$\sigma \doteq s$$

<u>Požadavky na bodové odhady:</u>

a) **Konzistence** – s rostoucím VS se výběrová charakteristika více blíží k parametru

b) **Nestrannost** – odhady parametru provedené na základě různých VS kolísají kolem hodnoty neznámého parametru na obě strany

c) **Minimální rozptyl** – uvedené kolísání musí být co nejmenší

<u>Nevýhody bodových odhadů:</u>

- neznáme jejich **spolehlivost a přesnost**

Odhady parametrů (2)

2/ Intervalové odhady

- Neznámý parametr odhadujeme intervalem vytvořeným kolem tzv. nejlepšího nestranného bodového odhadu

- **Interval spolehlivosti** (konfidenční interval)
- **Spolehlivost** si určujeme sami – buď 95% nebo 99%

jde o pravděpodobnost, že odhadovaný parametr se nachází v daném intervalu

95% CI (-;-)
99% CI (-;-)

Odhady parametrů (3)

- doplněk spolehlivosti vyjadřuje **riziko odhadu** – tj. riziko, že odhadovaný parametr leží mimo interval

při spolehlivosti 95% → **riziko odhadu 5%**

při spolehlivosti 99% → **riziko odhadu 1%**

Odhad průměru základního souboru (parametru μ) [mý]

1. <u>Nejlepší bodový odhad</u> parametru **μ** je výběrový průměr **m**
2. V souborech, kde **n>30**, se výběrový průměr chová jako náhodná veličina, která má <u>normální rozdělení</u>
3. V souborech, kde **n<30**, používáme model <u>Studentova rozdělení</u> (konstanty 1,96, příp. 2,58 se nahrazují jinými – viz. skripta,str. 25)
4. Každý výběrový průměr je zatížen chybou – jde o tzv. **standardní chybu** průměru Se_m, kterou odhadujeme ze vztahu: (střední chyba)

Závěr:

$$SE_m = \frac{s}{\sqrt{n-1}}$$

95% CI $\quad m \pm 1{,}96 \cdot \frac{s}{\sqrt{n-1}}$

99% CI $\quad m \pm 2{,}58 \cdot \frac{s}{\sqrt{n-1}}$

Vlastnosti odhadu

1) **Spolehlivost** – volí se předem, jde o stanovení pravděpodobnosti, obvykle 0,95 nebo 0,99
2) **Přesnost** – je dána délkou intervalu, <u>čím kratší je interval, tím je vyšší přesnost odhadu</u>

$$m \pm \boxed{1{,}96 \cdot \frac{s}{\sqrt{n-1}}}$$

Obě vlastnosti spolu souvisí

<u>Přesnost</u> odhadu lze ovlivnit:

a) snížením či zvýšením **P** spolehlivosti
b) snížením či zvýšením **n** (velikost souboru)
c) snížením či zvýšením **s** (homogenita souboru)

Příklad 1:

Odhadněte průměrnou vitální kapacitu plic mužů 40-50 letých na podkladě výběrového šetření 200 mužů, u kterých jsme zjistili:

$m = 4{,}82$

$s = 0{,}67$

$n = 200$

Příklad 2:

Skupina A:

Odhadněte průměrnou hladinu hemoglobinu v populaci zdravých mužů z náhodného výběru 100 jedinců s průměrnou hodnotou m = 152,4 g/l a směrodatnou odchylkou s = 18,2 g/l se spolehlivostí: a) 95% b)99%

Skupina B:

Odhadněte průměrnou hladinu hemoglobinu v populaci zdravých mužů z náhodného výběru 35 jedinců s průměrnou hodnotou m = 152,4 g/l a směrodatnou odchylkou s = 18,2 g /l se spolehlivostí: a) 95% b) 99%

Skupina C:

Odhadněte průměrnou hladinu hemoglobinu v populaci zdravých mužů z náhodného výběru 100 jedinců s průměrnou hodnotou m = 152,4 g/l a směrodatnou odchylkou s = 14,8 g/l se spolehlivostí: a) 95% b) 99%

Odhad pravděpodobnosti ZS (parametru π) [pí]

1. Nejlepší bodový odhad je **relativní četnost**

$$p = \frac{k}{n} \rightarrow \pi$$

   n = počet pozorování

   k = počet pozorování, u nichž nastal sledovaný jev

2. Pro pravděpodobnosti sice platí binomické rozdělení, ale pokud platí

$$n \cdot p \cdot (1-p) > 9$$

   můžeme vycházet z normálního rozdělení

3. **Standardní chybu SE** odhadujeme ze vztahu:

$$SE = \sqrt{\frac{p(1-p)}{n}} \qquad SE = \sqrt{\frac{p(100-p)}{n}} \quad \%$$

číslo

Pro 95 % CI $\pi = p \pm 1{,}96\sqrt{\frac{p(1-p)}{n}}$

99 % CI $\pi = p \pm 2{,}58\sqrt{\frac{p(1-p)}{n}}$

Příklad:

Odhadněte pravděpodobnost výskytu zrakové vady u studentů LF na základě výběrového šetření u 200 studentů

$n = 200$

$k = 80$

$p = 0{,}40$ (40%)

Příklad:

Ve výběru 100 šestiměsíčních zdravých dětí náhodně vybraných z brněnské populace byl sledován hemoglobin v g%.

$n = 100$ $m = 13{,}10$ $s = 1{,}90$

1) Určete interval, ve kterém se pohybuje hemoglobin u 95% vyšetřených dětí.
2) Odhadněte průměrné množství hemoglobinu v základním souboru se spolehlivostí 0,95. Jaká je přesnost tohoto odhadu?
3) U kolika dětí musíme provést šetření, aby přesnost odhadu průměru byla při spolehlivosti 0,95 nejméně $\pm$ 0,2.