-

1 Úvod do statistiky, náhodné jevy, pravděpodobnost náhodného jevu

- **Statistika =** vědní obor, který se zabývá zkoumáním jevů, které mají hromadný charakter; umožňuje rozvíjet lidské znalosti pomocí empirických dat
  - **Popisná** – zabývá se elementárními (jednoduchými) metodami sběru a zpracování hromadných údajů
  - **Matematická –** složitější metody sběru
  - **Ekonomická –** matematické i ekonomické úvahy; na pomezí společenské a metodologické vědy
- **ČSÚ =** hlavní orgán státní statistické služby; koordinuje sběr a zpracování statistických údajů na území ČR
  - Zabezpečuje získávání a zpracování údajů pro statistické účely a poskytuje statistické informace
  - Stanoví metodiku zjišťování včetně programu statistických zjišťování
  - Nezávislý na vládě a politických stranách

1.1 Základy teorie pravděpodobnosti

1.1.1 Teorie pravděpodobnosti

- Využívá se při zpracování a rozboru statistických údajů
- Zabývá se studiem zákonitostí, jimiž se řídí náhodné jevy
- Vznikla v polovině 17. století
- Podnětem pro vznik byly úvahy o hazardních hrách
- Největší zásluhu o rozvoj měla fyzika

1.1.2 Náhodné jevy

- **Náhoda** = komplex drobných příčin, které se od jednoho provedené pokusu k druhému významně liší; realizaci nazýváme *náhodným pokusem*
- **Náhodný pokus =** opakovatelný činnost, prováděná za stejných podmínek, která může v závislosti na náhodě vést k různým výsledkům (hod mincí, hod kostkou)
- **Náhodný jev =** výsledek náhodného pokusu a výchozím pojmem počtu pravděpodobností; při mnohonásobném opakování pokusu při zachování stejných podmínek se ve výskytu náhodné jevu objeví **jisté zákonitosti**
- **Jev jistý (U)** = nastane VŽDY při realizaci daných podmínek
- **Jev nemožná (V)** = jev, který nemůže nastat NIKDY
- **Jev náhodný** = za daných podmínek buď nastane nebo nenastane

1.1.3 Vztahy mezi náhodnými jevy

- **Průnik (součin) jevů A a B (A ∩ B)** – současné nastoupení jevu A i B (Vennovy diagramy)
- **Sjednocení (součet) jevů A a B (A U B)** – nastane alespoň jeden z jevů A a B
- **Rozdíl jevů A a B (A – B) –** náhodný jev, který se realizuje, nastane-li jev A a současně nenastane jev B
- **Jev A je částí jevu B (A C B) –** jestliže při každé realizaci jevu A nastává i jev B; jev A má za následek jev B
- **Rovnocennost jevů A a B (A = B) –** jevy jsou rovnocenné; pokud nastane A nastane B a naopak
- **Neslučitelnost jevů A a B (A ∩ B = V)** – výskyt jednoho jevu vylučuje možnost výskytu druhého jevu; průnik je jev nemožný

-

- **Opačný jev A ($\bar{A}$)** – jev, který spočívá v nenastoupení jevu A

1.1.4 Pravděpodobnost

- Číslo, které vyjadřuje míru možné realizace náhodného jevu A
- **A → P (A)**
- **Axiomatická definice pravděpodobnosti** - matematicky přesná definice, která tvoří základ celé současné teorie pravidelnosti (nejobecnější definice, zahrnuje klasickou i statistickou definici)
    - **Axiom** – základní předpoklad, ověřen zkušeností a už se nedokazuje

1.1.5 Definice pravděpodobnosti

- **Klasická definice –** předpokládá, že prostor elementárních jevů je konečný a že všechny tyto jevy jsou stejné možné (stejně pravděpodobné)
    - $P\ (A) = \frac{m}{n}$; m – počet příznivých pokusů pro opakování jevu A; n – počet všech pokusů
    - Jaká je pravděpodobnost, že v dodávce 10 krabic TV bude 5 vyřazeno z důvodu poškozených obalů? 5/100 = 0,05 = 5%
- **Statistická definice –** používá se tehdy, jestliže není splněn předpoklad stejného množství všech jevů; založena na stabilitě relativních četností, která vzniká po provedení pokusů
    - $P\ (A) = \lim_{n\to\infty}\left(\frac{m}{n}\right)$; m – počet pokusů, ve který nastal jev A; n – celkový počet nezávislých pokusů; m/n – relativní četnosti
    - S rostoucím počtem pokusů se relativní četnost stabilizuje a přibližuje se k určitému konstantnímu číslu
- Klasická a statistická definice nejsou v rozporu
- Rozdíl je v tom, že při použití klasické definice určujeme pravděpodobnost náhodného jevu **před provedením pokusu** a při použití statistické **vycházíme z již provedených pokusů**

1.1.6 Vlastnosti pravděpodobnosti

- **P libovolného jevu A je číslo z intervalu <0; 1>** : 0 ≤ P(A) ≤ 1
- **P nemožného jevu je rovna nule a jistého jedné: P(V) = 0; P(U) = 1**
- Dva rovnocenné jevy jsou i stejně pravděpodobné: P(A) = P(B)
- Jestliže jev A je částí jevu B: P(A)≤ P(B)
- **Mezi P jevu A a opačného jevu $\bar{A}$ existuje vztah: P($\bar{A}$) = 1 – P(A)**
- Pravděpodobnost sjednocení neslučitelných jevů je rovna součtu pravděpodobností těchto jevů.
- **Věta o sčítání pravděpodobnosti.**
    - Pomocí této věty vyjádříme pravděpodobnost sjednocení jevů A a B
    - Pro dva libovolné slučitelné jevy platí, že P jejich sjednocení je rovna součtu P jevů zmenšeného o P jejich průniku
        - **P(A U B) = P(A) + P(B) – P(A ∩ B)**
    - Pro dva neslučitelné jevy platí, že průnik je nemožný
- **Věta o násobení pravděpodobnosti.**
    - Pomocí této věty vyjádříme p průniku jevů
    - Závislost jevů je charakterizovaná pomocí podmíněné pravděpodobnosti:
        - $P(A/B) = \frac{P(A\cap B)}{P(B)};\ P(B/A) = \frac{P(B\cap A)}{P(A)}$
    - Pravděpodobnost náhodného jevu A je ovlivněna podmínkou, že nastal nějaký náhodný jev B, který má nenulovou P (A je závislý na B)

-

        - **P (A ∩ B) = P(A) x P(B/A) = P(B) x P(A/B)**
    - Pokud nastoupení jevu A neovlivňuje pravděpodobnost nastoupení jevu B, jevy A a B jsou nezávislé.
        - **P (A ∩ B) = P(A) x P(B)**

2 Náhodné veličiny, základní diskrétní a spojitá rozdělení

2.1 Modely rozdělení náhodných veličin

2.1.1 Náhodná veličina

- Proměnná, která nabývá konkrétních hodnot nebo hodnot z určitého intervalu; kvantitativní charakteristika náhodného pokusu
- Značíme velkými písmeny X, Y, jejich hodnoty malými písmeny X: x1, x2, …
- **diskrétní** – nabývá od sebe vzájemně oddělené hodnoty, *celočíselné* (počet telefonních hovorů, počet selat ve vrhu, …)
- **spojitá** – nabývá hodnot z *konečného či nekonečného intervalu* (výška rostliny, hmotnost zvířete, …)

2.1.2 Rozdělení náhodné veličiny

- Pravidlo, které každé hodnotě přiřazuje P, že náhodná veličina nabude této hodnoty
- Různé formy popisu rozdělení NV:
    - **Řada rozdělení (diskrétní NV)**
        - Umožňuje popsat rozdělení diskrétních NV
        - **Nejjednodušší** forma popisu, zapisujeme do tabulky; grafem je polygon

x_i	x_1	x_2	…	x_{n-1}	xn
p_i	p_1	p_2	…	p_{n-1}	pn

        - *příklad přednáška č.2 – slide 6-8*
    - **Distribuční funkce (diskrétní i spojité NV) – F(x)**
        - Umožňuje popsat rozdělení spojitých a diskrétních NV
        - P (X < x) značí P jevu, který nastane, když náhodná veličina X nabude hodnoty menší než x
        - Fce, která každému reálnému číslu přiřazuje p, že náhodná veličina nabude hodnoty menší než toto číslo -> **F (3) = P (x ≤ 3)**
            - **Vlastnosti:** interval <0;1>, neklesající fce, fce spojitá zleva
    - **Hustota pravděpodobností (spojitá NV)**
        - Pro spojitou náhodnou veličinu X existuje nezáporná fce f(x) -> derivace distribuční funkce

2.2 Číselné charakteristiky náhodných veličin

2.2.1 Charakteristiky náhodných veličin

- Číselné charakteristiky umožňují jednoznačně a jednoduše popsat tvar rozdělené NV
- K základním charakteristikám patří *charakteristiky polohy* (určují střed rozdělení), *charakteristiky variability* (míra kolísání NV kolem střední hodnoty)

2.2.2 Charakteristiky polohy

- **Střední hodnota** náhodné veličiny – E(X) – pro diskrétní i spojitou fci

2.2.3 Charakteristiky variability

- **Rozptyl** – pro diskrétní i spojitou fci; D(X) ; na druhou; udává variabilitu ve čtvercích jednotek
- **Směrodatná odchylka** – odmocnina rozptylu; měří variabilitu v původních jednotkách veličiny

-

2.3 Základní typy rozdělení náhodných veličin

2.3.1 Rozdělení diskrétních NV

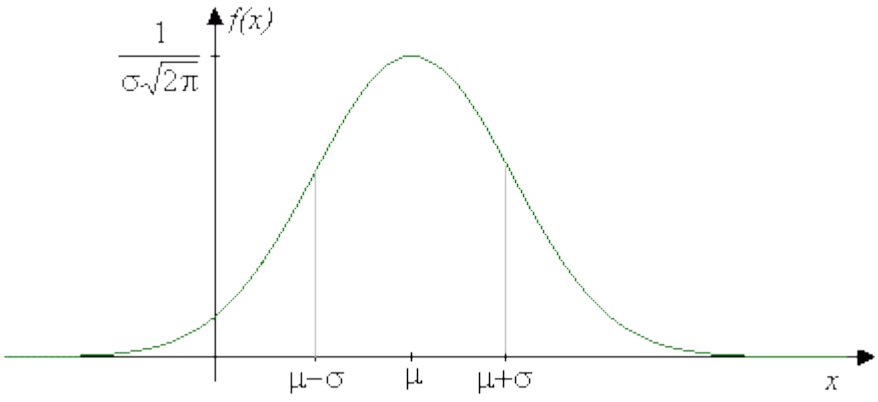
- **Alternativní rozdělení**
  - Nabývá pouze dvou hodnot: $X_1 = 1$ jestliže jev A nastane; $X_2 = 0$, jestliže jev A nenastane

x_i	0	1
p_i	1 - p	p

  - *Příklad přednáška č. 2 slide 18-19*
- **Binomické rozdělení**
  - Diskrétní veličina, kdy n<30 ; jev A může nastat s pravděpodobností p (p>0,1) a nenastane s 1-p; tzv. Bernoulliho schéma
  - Pravděpodobnost výskytu jevu A je stejná
  - Model s vracením
  - *Příklad přednáška č.2 slide 22-23*
- **Poissonovo rozdělení**
  - n>30; P výskytu sledovaného jevu v jednom pokuse je p ≤ 0,1
  - jedno z nejdůležitějších – řídí se jím počet výskytů náhodného jevu A během časového intervalu délky t (počet volání na telefonní ústřednu v určitém časovém intervalu, …)
  - výběr s vracením
  - *příklad přednáška č.2 slide 26-27*
- **Hypergeometrické rozdělení**
  - Vychází z opakování náhodného pokusu, kde pravděpodobnost nastoupení sledovaného jevu je závislá na výsledcích předcházejících pokusů
  - Výběr bez vracení
  - *Příklad přednáška č.2 slide 30*

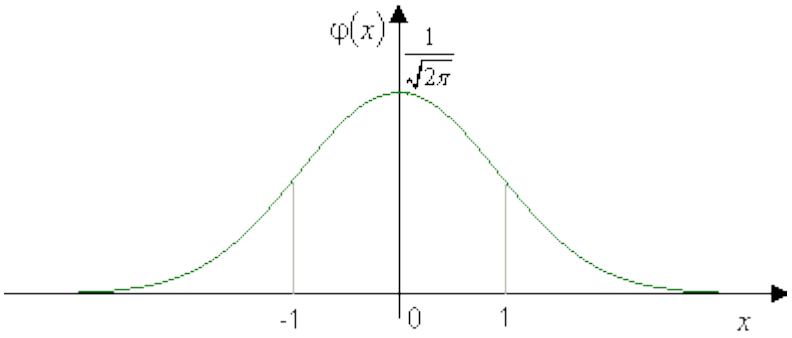
2.3.2 Rozdělení spojitých NV

- **Normální (Gaussovo)**
  - Nejčastěji se vyskytující ve všech vědních oborech (Gaussova křivka)
  - **Parametry N (μ; σ^2) –** střední hodnota; rozptyl
  - **Střední hodnota** je zároveň mediánem i modem (modus – nejčastější hodnota)
  - **Gaussova křivka** – graf hustoty pravděpodobnosti; symetrická kolem bodu x = μ, v němž dosahuje svého maxima



-

  - Jestliže μ = 0 a $\sigma^2 = 1$, hovoříme o **normovaném normálním rozdělení** – speciální typ normálního rozdělení se střední hodnotou rovnou nule a jednotkovým rozptylem N(0,1)



  - Pokud má náhodná veličina X normální rozdělení s parametry μ a σ2, můžeme tuto veličinu převést na veličinu U s normovaným normálním rozdělením, pomocí vztahu: $U = \frac{X-\mu}{\sigma}$; X = 1,2 násobek směrodatné odchylky

- **Rovnoměrné**
  - Řídí se jim takové NV, které mají stejnou možnost nabýt kterékoliv hodnoty z nějakého intervalu
- **Exponenciální**
  - U technických aplikací teorie pravděpodobnosti a při řešení fyzikálních problémů; parametr lambda

2.3.3 Speciální rozdělení

- Odvozená od normálního rozdělení
- **Pearsonovo χ2 rozdělení –** chí kvadrát rozdělení o n stupních volnosti; náhodné veličiny U, součet jejich čtverců (tzn. na druhou) je chí kvadrát
- **Studentovo t-rozdělení (základ t-testů) –** nezávislé veličiny t, kritické hodnoty tabelovány
- **Fisher – Snedecorovo rozdělení –** spojité; kritické hodnoty F rozdělení

3 Popisná statistika

3.1 Popisná statistika

3.1.1 Statistika

- Praktická činnost, spočívající ve sběru, zpracování a vyhodnocování statistických údajů (analýza)
- Disciplína, který se zabývá tím, jak získat informace z dat (numerických, kvalitativních)

3.1.2 Základní pojmy

- **Statistický soubor = základní jednotka;** množina jedinců, na které se dané statistické šetření provádí
  - **Základní –** obsahuje všechny jednotky
  - **Výběrový** – část základního souboru
- **Statistická jednotka =** prvek daného souboru
- **Rozsah souboru** = počet prvků statistického souboru
- **Statistický znak =** vlastnost stat. souboru
  - Různé klasifikace proměnných dle vlastností stat. znaku jehož jsou nositelem
    - **Nominální proměnná** – o jejích hodnotách můžeme říct za jsou stejné či různé (škola, fakulta, obor); lze u nich zjišťovat jen rozdělení četností

-

        - **Ordinální proměnná** – u hodnot můžeme určit pořadí (vzdělání)
        - **Kvantitativní proměnná** – můžeme přesně říct, o kolik je jedna hodnota vyšší než druhá (výška, hmotnost, věk)
- **Kvantitativní proměnné**
    - Diskrétní – nabývají celočíselných hodnot
    - Spojité – nabývají hodnot z intervalu
- **Kategoriální proměnné** (nominální, ordinální, kvantitativní diskrétní proměnné)
    - Alternativní – nabývají pouze dvou obměn (kategorií)
    - Množné – nabývají více než dvou obměn

3.1.3 Metody a formy statistického zkoumání

- **Statistické zjišťování** – vlastní sběr dat, slouží k získávání a shromažďování údajů o stat. jednotkách
- **Statistické zpracování** – roztřídění dat a shrnutí tak, aby vynikly charakteristické rysy zkoumaného souboru
- **Statistické vyhodnocování (rozbor)** – vychází z popisu zkoumaných jevů, na který navazuje podrobná analýza, jejímž cílem je určení stat. zákonitostí a vzájemných souvislostí v datech, rozbor získaných výsledků a formulace jevu

3.1.4 Statistické zjišťování

- Získávání primárních údajů
    - **Přímé pozorování** – vážení, měření, sčítání (zjišťování počtu osob)
    - **Dotazování** – výkaznictví (formuláře, …), dotazník, rozhovor
- **ÚPLNÉ** – **základní soubor** – zjišťuje se hodnota příslušného znaku u všech statistických jednotek (např. sčítání lidu, cenzus)
    - **Výhody** – přesné výsledky o sledovaném souboru i o každé jeho statistické jednotce
    - **Nevýhody** – velké finanční náklady, pracnost, nespolehlivost zjišťovaných dat; těžko se kontroluje správnost údajů, protože se můžou objevit nepravdivé nebo záměrně zkreslené údaje
- **NEÚLNÉ – výběrový soubor –** omezení jen na část stat. souboru a zkoumání některých stat. Jednotek; výběrové zjišťování musí obsahovat podstatné a charakteristické rysy základního souboru
    - **Výhody –** rychlejší, nižší náklady
    - **Nevýhody –** umožňuje pouze odhady o průběhu sledovaných procesů v základním souboru; údaje jsou zatížené chybou – *výběrová chyba (chyba odhadu)*

3.1.5 Statistické zpracování

- Poskytuje základní přehled a kontrolu dat o hodnotách proměnných
    - **Třídění**
        - Jednorozměrné rozdělení četností (jeden třídící znak) – barva vlasů
        - Vícerozměrné rozdělení četností (více znaků) – barva vlasů a pohlaví
    - **Tabelování**
        - Výsledky třídění se zapisují do tabulky rozdělení četností
    - **Grafické znázorňování**

3.1.6 Rozdělení četností - třídění

- **Absolutní četnost n_i –** počet opakování hodnoty znaku v původní řadě dat; skutečný počet jednotek; součet je roven rozsahu

-

- **Relativní četnost f_i –** jaká část vyšetřovaného souboru má hodnotu znaku x_i; porovnání různých četností mezi sebou, vyjádřen v %
- **Kumulativní absolutní četnost N_i**
  - Vzniká postupným načítáním
- **Kumulativní relativní četnost F_i**
- **Intervalové rozdělení četností**
  - Určité zjednodušení, které zachycuje četnost intervalu a ne skutečně zjištěných hodnot; místo hodnot četností intervaly
  - **Počet intervalů (k):** pro jeho stanovení neexistuje jednotný přepis, odpovídá odmocnině z n
  - **Šířka (délka) intervalů (h):** tvar intervalového rozdělení lze ovlivnit délkou intervalů; malá šířka – podstatné vlastnosti nezřetelné; široký – tvar rozdělená ve hrubých rysech
    - $\boldsymbol{h = \frac{R}{k}}$ **;** x – hodnota znaku, R – variační rozpětí ($x_{max} – x_{min}$)
    - Musí být jednoznačně určeno, kam kterou jednotku zařadit

3.1.7 Rozdělení četností – tabelování

x_i	četnost n_i	relativní četnost f_i	kumulativní četnost N_i	kumulativní procenta F_i
1	12	40,0	12,0	40,0
2	11	36,7	23,0	76,7
3	5	16,7	28,0	93,3
4	2	6,7	30,0	100,0
Celkem	30	100,0		

- **Intervalové rozdělení četností**

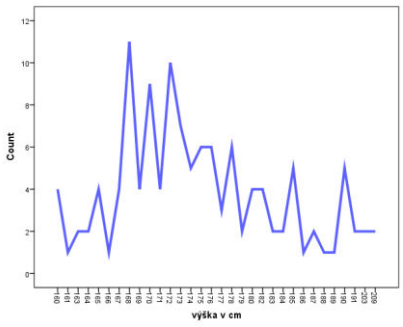
interval	n_i	procenta	N_i	F_i
2 - 5	3	10,0	3	10,0
6 - 9	3	10,0	6	20,0
10 - 13	7	23,3	13	43,3
14 - 17	14	46,7	27	90,0
18 - 21	3	10,0	30	100,0
celkem	**30**	**100,0**		

3.1.8 Rozdělení četností – grafická analýza

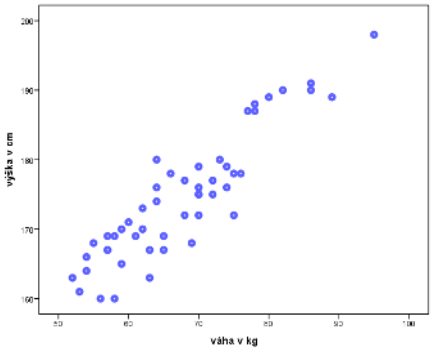
- Grafické zobrazení dává rychlou a přehlednou představu o tendencích a charakteristických rysech analyzovaných proměnných
- Grafy umožňují rychlou kontrolu sledovaných údajů; účinný prostředek pro prezentaci stat. výsledků
- **Spojnicové grafy (polygon četností)** – prosté rozdělení četností

-

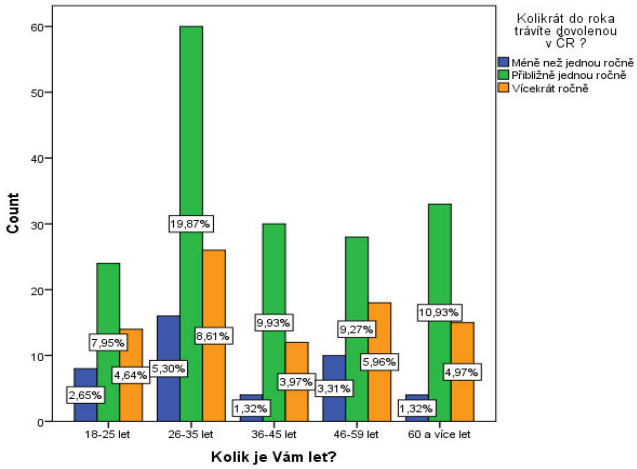
-



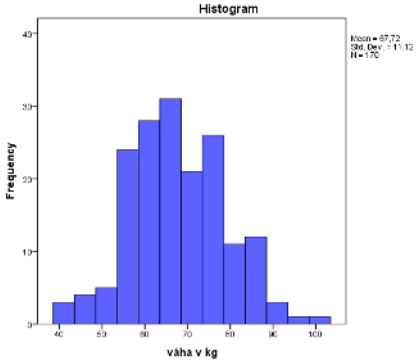
  - **pro diskrétní znak neužívat spojnicové grafy** (nelze spojit 0 a 1 dítě)
- **bodové grafy** – zobrazení a porovnání číselných hodnot (srovnání velkých počtů datových bodů bez ohledu na čas)



- **výsečové (koláčové) grafy –** zobrazují velikost položek jedné datové řady úměrně k součtu položek; lze zobrazit pouze jednu datovou jednotku (kolik vám je let?)
- **sloupcové grafy** – data četností; možno porovnat různé hodnoty kategorií proměnné



- **histogram** – pro intervalové rozdělení četností jedné proměnné



-

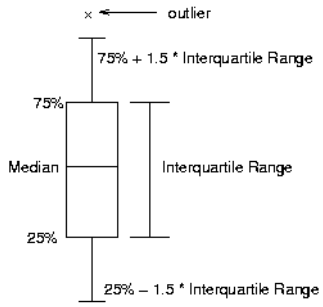
- na rozhraní mezi tabulkami a grafy mezi zobrazením strojových dat a souhrnných údajů jsou grafy, které poskytují informace o rozdělení četností, a také o základních číselných charakteristikách sledovaných proměnných (minimum, maximum, medián, …) -> tyto grafy řadíme do části statistiky, kterou nazýváme **průzkumový (explorační) analýza dat**

3.1.9 Percentily

- Hodnoty oddělující určité procento pozorování; necitlivé k extrémním hodnotám
- **P – procentní kvantil** – hodnota x_p, která odděluje *p* procent nejnižších hodnot rozdělení od *(1-p)* procent nejvyšších hodnot -> odděluje soubor na několik částí
- **Nejčastěji používané kvantily**:
  - **Kvartily** – dělí uspořádaný datový soubor na 4 stejně obsazené části
    - $\mathbf{x_{0,25}}$ – spodní kvartil (Q_1); $\mathbf{x_{0,5}}$ – medián; $\mathbf{x_{0,75}}$ – horní kvartil (Q_3)
  - **decily** – dělí datový soubor na desetiny
  - **percentily** – dělí datový soubor na setiny

3.1.10 Kvantily

- Charakteristiky **polohy rozdělení**, lze použít i k charakterizování variability
- **Interkvartilové rozpětí**
  - $IQR = x_{0,75} - x_{0,25}$
- **Kvartilová odchylka**
  - $$s_k = \frac{x_{0,75} - x_{0,25}}{2}$$
- **Boxplot** – krabicový graf; data zobrazeny pomocí kvartilů; srovnání rozložení četností v několika skupinách



3.2 Statistické vyhodnocování

3.2.1 Základní statistické charakteristiky

- statistickými charakteristikami nazýváme **číselné hodnoty**, které podávají základní informaci o vlastnostech stat. souboru z hlediska odhalení variability, stupně symetrie a špičatosti, normalitě rozdělení a v neposlední řadě nalezení vybočujících a podezřelých prvků ve výběru
- charakteristiky:
  - polohy (úrovně)
  - variability
  - šikmosti a špičatosti

3.2.2 Charakteristiky polohy

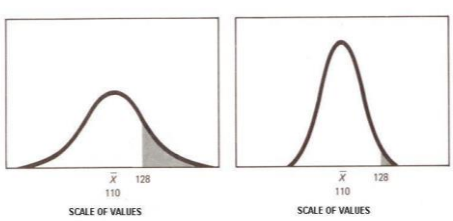
- Poskytuje základní informaci o daném rozdělení
- Míra polohy je taková hodnota NV, kolem které se soustřeďují všechny ostatní hodnoty NV

-

- Dvě základní skupiny:
  - Průměry
    - **Aritmetický průměr** $\bar{x}$
      - Pro všechny tyto charakteristiky je společné, že jsou určovány na základě všech naměřených hodnot znaku
      - ve stejných jednotkách jako daná proměnná; součet odchylek roven 0; citlivý na extrémní hodnoty
      - prostý (nejdůležitější – všechny hodnoty), vážený (pokud počítáme z rozdělených četností)
    - **geometrický průměr**
      - používáme u časových řad; n-tá odmocnina ze součinu n hodnot
    - **harmonický průměr**
      - v indexní analýze; průměr převrácených hodnot
    - **chronologický průměr**
      - elementární průměr využívaný při problematice časových řad
  - Ostatní střední hodnoty
    - **Medián** $\tilde{x}$
      - Data musí být seřazeny podle velikosti; prostřední hodnota (lichý uprostřed, sudý průměr z dvou prostředních čísel)
      - U soboru s lichým rozsahem: $\tilde{x} = x_{\frac{n+1}{2}}$ – jednotky uspořádáme podle velikosti
      - u souboru se sudým rozsahem: $\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$ – prostý aritmetický průměr dvou sousedních prostředních hodnot
    - **modus** $\hat{x}$
      - hodnota znaku s největší četností

3.2.3 Charakteristiky variability (proměnlivosti)

- Variabilita = vzdálenost hodnot od střední hodnoty
- **Absolutní –** absolutní rozdíl hodnot znaků od střední hodnoty nebo od sebe navzájem
  - **Variační rozpětí R**
    - Délka intervalu, ve kterém se hodnoty nacházejí
    - **R = x_{max} – x_{min},** nevysvětluje variabilitu uvnitř intervalu
    - Nejjednodušší míra absolutní variability
  - **Rozptyl S^2**
    - Rozptýlení hodnot
    - Měří současně variabilitu hodnot kolem $\bar{x}$ a variabilitu ve smyslu vzájemných odchylek jednotlivých hodnot znaku; ukazuje odchylka od střední hodnoty i mezi sebou
    - Odmocněním získáme **směrodatnou odchylku**
  - Dvě rozdělení se stejnou střední hodnotou a odlišným rozptylem.

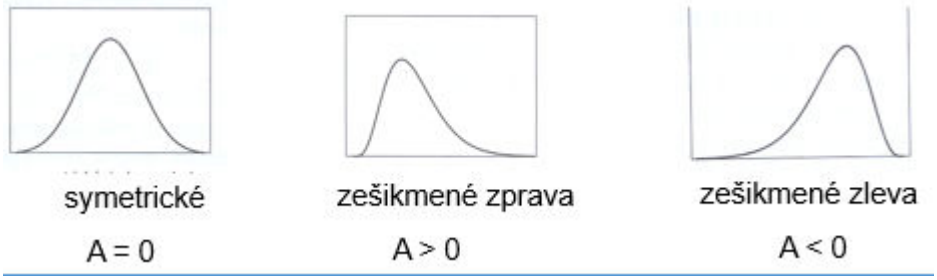


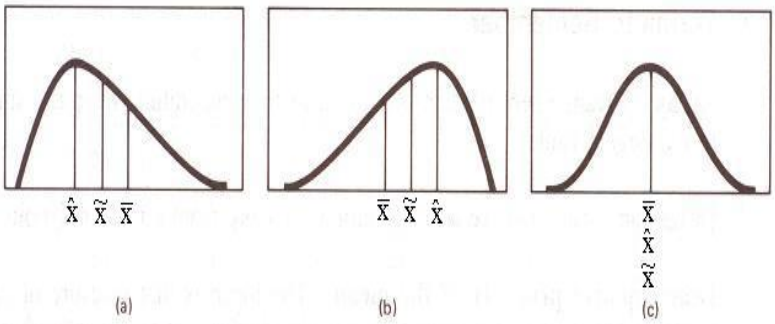
-

- **Relativní –** absolutní míra variability v poměru ke střední hodnotě sledovaného znaku
  - **Variační koeficient V**
    - Relativní míra variability
    - Poměr směrodatné odchylky a průměru
    - Vyjadřuje se v %
    - Čím nižší variabilita, tím je soubor uspořádanější, kompaktnější a data jsou si blíž

3.2.4 Charakteristika šikmosti

- Založeny na srovnání koncentrace malých hodnot sledovaného znaku a s koncentrací velkých hodnot tohoto znaku
- **Koeficient šikmosti (A)** vyjadřuje symetrii uspořádání dat kolem aritmetického průměru





- **Koeficient špičatosti E** – určuje koncetraci hodnost souboru kolem průměru
  - **Záporná hodnota** – podnormální špičatost
  - **Kladná** – nadnormální špičatost

3.2.5 Průzkumová analýza dat

- Všechny výše uvedené metody je v různých obměnách možné použít pro průzkumovou analýzu, jednorozměrných i více… souborů dat
- Zaměřena na splnění předpokladů:
  - Prvky ve výběru jsou vzájemně nezávislé
  - Výběr je homogenní (všechny x pocházejí ze stejného rozdělení p s konstantním rozptylem)

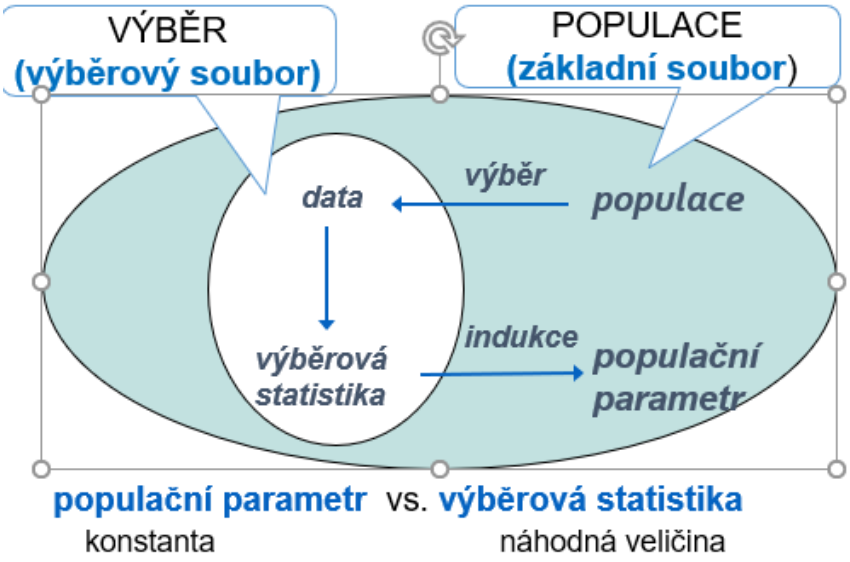
-

  - Data mají normální rozdělení
- v SPSS explore, descriptives, …F

4 Úvod do teorie odhadu

4.1 Statistická indukce teorie odhadu

4.1.1 Statistická indukce



-

- **Postup:**
  - Pracujeme převážně s výběrem
  - Uděláme indukci (odhad) na základě výběrových dat, jak se chová celý populační parametr
  - Zjistím výběrové statistiky a na základě nich uděláme indukci
  - Populační parametr je konstanta a výběr je NV (to, s čím pracujeme)
- **Základní označení:**

Název charakteristiky	Základní soubor	Výběrový soubor
Rozsah souboru	**N**	**n**
Absolutní četnost	$\mathbf{N_i}$	$\mathbf{n_i}$
Průměr	**μ**	$\bar{x}$
Rozptyl	$\mathbf{\sigma^2}$	$\mathbf{s^2}$
Směrodatná odchylka	**σ**	**s**
Relativní četnost	**π**	**p** ($\mathbf{f_i}$)

- **Statistická indukce (usuzování)** – na základě zkoumání náhodného výběru (reprezentativní zástupce populace – ZS) činíme závěry o ZS
- **Metody**:
  - **Teorie odhadu** – odhad populačních charakteristik; určení typu rozdělení sledovaného znaku, respektive některých charakteristik
  - **Testování hypotéz** – ověřování statistických hypotéz

-

4.1.2 Způsoby výběru (základní druhy výběrových zjišťování)

- **Anketa** – oslovuje jen určitou část stat. jednotek; provádí se rozesláním dotazníků či osobním kontaktem
- **Metoda základního masivu** – používá se, pokud se soubor skládá z několika velkých jednotek a z velkého počtu malých jednotek; nedovoluje zobecňovat výsledky na celý soubor, protože zkoumaný jev vykazuje jiné zákonitosti a tendence
    - **Záměrný výběr** – o zahrnutí jednotek do VS rozhodují různá logická hlediska a subjektivní názor vybírajícího; nelze pořizovat objektivní odhady, provádí znalec problematiky
    - **Náhodný výběr** – nejpoužívanější; soubor se rozdělí na výběrové jednotky a každé se přiřadí p jejího zahrnutí do VS; o zahrnutí výběrové jednotky rozhoduje *náhoda*
        - Dělíme dle pravděpodobnosti výběru:
            - **Prostý náhodný výběr** – výběr se stejnými p; přímý výběr – ze ZS se vybírají přímo stat. jednotky, nikoliv jejich skupiny (=každý prvek má stejnou p, že bude vybrán)
            - **Výběr s nestejnými pravděpodobnostmi** – výběrové jednotky mají přiřazeny různé p, je nutné předem znát určité doplňkové informace o sledovaných jednotkách
        - Dělíme dle toho, zda jsou zahrnuty nebo nejsou do výběru:
            - **s vracením** – vybereme jednotku a zase vrátíme do souboru; stále stejná p že jednotky vybereme; nesnižuje rozsah souboru
            - **bez vracení** – vybranou jednotku už nevracíme do souboru; snižuje se rozsah souboru a zvyšuje se p výběru jiné jednotky
- Techniky náhodného výběru:
    - **Losování** – důkladné promíchání všech jednotek
    - **Tabulka náhodných čísel** – sestaveny pomocí speciálních algoritmů
    - **Systematický výběr** – jednotky seřazeny do posloupnosti, pořadí nesouvisí se zjišťovanou skutečností
- Odhadujeme rozsah, četnost, aritmetický průměr, rozptyl, směrodatnou odchylku, variační koeficient; informace obdržíme ze ZS

4.2 Teorie odhadu

- Odhad neznámé populační charakteristiky
- **Bodový** – neznámou charakteristiku ZS nahradíme hodnotou určité výběrové charakteristiky; jedná se o odhad jedním číslem
- **Intervalový** – poskytuje číselný interval, ve kterém lze očekávat odhadovanou charakteristiku s předem zvolenou pravděpodobností
- **Předpoklady pro konstrukci**:
    - Výběry ze ZS s normálním rozdělením
    - Dostatečný rozsah (n>30)

4.2.1 Bodový odhad

- Odhadovaná statistika T je bodový odhad parametru Θ
- Musí splňovat určité vlastnosti:
    - **Nestrannost** – při opakovaných výběrech kolísá odhad tak, že v průměru se odhady z výběru rovnají populační hodnotě; odhad nesmí vést k systematickému nadhodnocování či podhodnocování odhadované charakteristiky

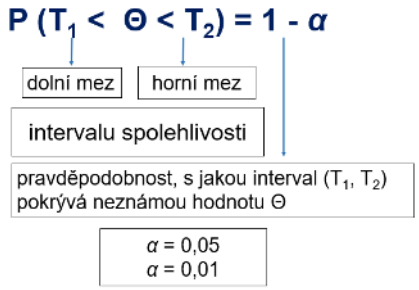
-

  - **Konzistence** – s rostoucím rozsahem výběru se zvyšuje pravděpodobnost, že se odhad bude co nejvíce blížit skutečné hodnotě odhadované populační charakteristiky
  - **Vydatnost** – rozptyl odhadu při opakovaných výběrech je malý
  - **Postačující** – statistika T je postačující, pokud obsahuje všechny informace o populační charakteristice; neexistuje-li žádná další statistika, který by obsahovala o odhadované charakteristice ZS nějakou další informaci
- **Bodový odhad průměru** $\hat{\mu} = \bar{x}$
  - výběrový průměr je nestranným odhadem populačního průměru
- **bodový odhad populačního rozptylu** $\hat{\sigma}^2 = s^2$
  - nestranným odhadem populačního rozptylu je výběrový rozptyl
- **bodový odhad populační relativní četnosti π = f_i**
  - bodový odhad je postačující pro velké VS, pro menší je lepší použít intervalový odhad

4.2.2 Intervalový odhad

- **Přesnost odhadu** – maximální chyba, které se při odhadu s danou spolehlivostí dopustíme
- Spočívá ve stanovení intervalu, ve kterém se neznámá charakteristika vyskytuje s předem známou pravděpodobností
- Při intervalovém odhadu se vychází z konstrukce tří možných typů intervalu:
  - Oboustranný interval spolehlivosti
  - Jednostranný interval spolehlivosti
    - Levostranný – omezen zdola
    - Pravostranný – omezen shora
- α je hladina významnosti

4.2.3 Oboustranný interval spolehlivosti



-

4.2.4 Jednostranné intervaly spolehlivosti

- $P(-\infty < \theta < T_2) = 1 - \alpha$ **– pravostranný interval spolehlivosti**
- $P(T_1 < \theta < \infty) = 1 - \alpha$ **– levostranný interval spolehlivosti**

4.2.5 Interval spolehlivosti pro populační průměr μ

- Předpoklad – výběr z velké populace s normálním rozdělením
- **Oboustranný interval spolehlivosti**

$$P(\bar{x} - \Delta < \mu < \bar{x} + \Delta) = 1 - \alpha$$

Δ .. přípustná chyba odpadu -> charakterizuje přesnost odhadu, tj. udává maximální chybu, které se můžeme při odhadu střední hodnoty populačního průměru μ při předem zvolené hladině významnosti **α** dopustit.

-

- Konstrukce intervalu spolehlivosti se odvíjí od znalosti či neznalosti populačního rozptylu

-

- **Odhad při známém populačním rozptylu σ^2**
  - $\Delta = u_\alpha \frac{\sigma}{\sqrt{n}}$; u – kritická hodnota normálního rozdělení pro alfa; σ – populační směrodatná odchylka; n – rozsah výběrového souboru
- **Odhad při neznámém populačním rozptylu σ^2**
  - $\Delta = t_\alpha(f) \frac{s}{\sqrt{n}}$; t – kritická hodnota Studentova t-rozdělení pro hladinu významnosti alfa a počet stupňů volnosti *f = n-1;* s – výběrová směrodatná odchylka; n – rozsah souboru

a) Odhadněte pomocí bodového i intervalového odhadu (α = 0,05), jaká je průměrná rychlost projíždějících vozidel.

$$\mathbf{n = 30;\ \bar{x} = 56{,}43;\ s = 7{,}286;\ t_{0{,}05(29)} = 2{,}045}$$

$$\Delta = t_{\alpha(f)} \frac{s}{\sqrt{n}} = 2{,}045 \quad \frac{7{,}286}{\sqrt{30}} = 2{,}72$$

$$P(\bar{x} - \Delta < \mu < \bar{x} + \Delta) = 1 - \alpha$$

$$P(56{,}43 - 2{,}72 < \mu < 56{,}43 + 2{,}72) = 1 - 0{,}05$$

$$P(53{,}71 < \mu < 59{,}15) = 0{,}95$$

b) U kolika automobilů by musela být rychlost měřena, aby bylo možné na 5% hladině významnosti odhadnout populační průměr s maximální přípustnou chybou 1 km?

$$\Delta = t_{\alpha(f)} \frac{s}{\sqrt{n}} \qquad 1 = 2{,}045 \frac{7{,}286}{\sqrt{n}}$$

$$\mathbf{n = 220}$$

Descriptives

			Statistic
rychlost auta v km/hodina	Mean		56,43
	95% Confidence Interval for Mean	Lower Bound	53,71
		Upper Bound	59,15
	5% Trimmed Mean		56,50
	Median		56,00
	Variance		53,082
	Std. Deviation		7,286
	Minimum		40

Příklad v SPSS

Před vlastním výpočtem nezbytné provést průzkumovou (exploratční) analýzu dat !!!!!!

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
rychlost auta v km/hodina	,148	30	,092	,974	30	,656

$$P(\bar{x} - \Delta < \mu < \bar{x} + \Delta) = 1 - \alpha$$

$$P(53{,}71 < \mu < 59{,}15) = 95\%$$

Δ = ...

- Odvodíme z výsledku
- Přípustná chyba 2,72
- Rozdíl: 59,15 – 53,71 = 5,44
- 5,44 : 2 = 2,72

4.2.6 Intervalový odhad rozptylu

- Předpoklad – výběr z populace s normálním rozdělením, parametr μ je neznámý
- **Oboustranný interval spolehlivosti rozptylu**

$$P\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(f)} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(f)}\right) = 1 - \alpha$$

$\chi^2_{\frac{\alpha}{2}}(f)$, $\chi^2_{1-\frac{\alpha}{2}}(f)$ … kritické hodnoty χ^2 rozdělení s *f* = n -1 stupni volnosti

-

- **Oboustranný interval spolehlivosti**

4.2.7 Intervalový odhad relativní četnosti

- π … relativní četnost jednotek s vlastností A v populaci
- p … relativní četnost jednotek s vlastností A ve výběru
- výběry s opakováním -> binomické rozdělení
- výběry bez opakování -> hypergeometrické rozdělení
- velký výběrový soubor – aproximace normálním rozdělením
- předpoklad – velký rozsah výběru

$$P\left(s\sqrt{\frac{n-1}{\chi^2_{\frac{\alpha}{2}}(f)}} < \sigma < s\sqrt{\frac{n-1}{\chi^2_{1-\frac{\alpha}{2}}(f)}} \right) = 1-\alpha$$

symbolika viz odhad rozptylu

(f) – počet stupňů volnosti

- **Interval spolehlivosti pro populační relativní četnost π**

četnost π

$$P(p-u_\alpha\sqrt{\frac{p(1-p)}{n}} < \pi < p+u_\alpha\sqrt{\frac{p(1-p)}{n}}) = 1-\alpha$$

-

4.2.8 Neparametrický odhad mediánu

- Citlivost aritmetického průměru na vybočující hodnoty, důležité u malých souborů
- Medián = robustní charakteristika polohy
- $\tilde{x}$ **– výběrový medián; M – populační medián**
- **Intervalový odhad populačnímu mediánu M**
  - Předpoklad – spojitost znaku X
  - data uspořádaná od min po max; podle rozsahu *n* se nalezne takové přirozené číslo *k*, pro které platí:
    - $P\,(x_k \leq M \leq x_{n-k+1}) \geq 1-\alpha$
    - k najdu v tabulkách podle n

5 Testování statistických hypotéz

5.1 Úvod do testování statistických hypotéz

5.1.1 Testování statistických hypotéz

- Druhý základní úkol statistické indukce; postup, při kterém ověřujeme, zda předem vyslovená hypotéza platí pod vlivem provedených pozorování
- **Hypotéza** – předpoklad, domněnka
- **Statistická hypotéza** – určitá podtřída vědeckých hypotéz
  - Tvrzení o parametrech rozdělení
    - Rovnost dvou parametrů, dvou a více; shodou parametrů VS a ZS; shodou empirického a teoretického rozdělení
  - Tvrzení o tvaru rozdělení
    - Na základě náhodného výběru posuzujeme, zda stat. hypotéza je pravdivá či ne
  - Úkolem je konstrukce adekvátních matematických metod, podle kterých budeme posuzovat platnost či neplatnost zkoumané statistické hypotézy

-

- Testovaná stat. hypotéza = **nulová hypotéza H_0**
  - Předpoklad, který jsme vyslovili o určité charakteristice
  - Tvrdí, že neexistuje rozdíl mezi dvěma nebo větším počtem rozdělení
  - **H_0: $\Theta = \Theta_0$**
  - **Nulová hypotéza říká:**
    - Parametr ZS se statisticky významně neliší od hypotetické jednotky
    - Parametr je roven určité pevné hodnotě
    - mezi sledovanými objekty neexistuje rozdíl
  - **alternativní hypotéza H1:**
    - popírá platnost nulové hypotézy
    - přijímáme v případě zamítnutí nulové hypotézy
    - obvykle se vyjadřuje jako „existence diference“ nebo „existence závislosti“ mezi proměnnými
    - nemusí jít vždy o přesný opak nulové hypotézy
      - **oboustranná alternativa**
        - $H_1: \Theta \neq \Theta_0$
      - **jednostranná alternativa**
        - $H_1: \Theta > \Theta_0$ – pravostranná alternativa
        - $H_1: \Theta < \Theta_0$ – levostranná alternativa
- **Test** – postup, kterým na základě náhodného výběru ověřujeme, zda hypotéza platí či nikoli
  - **Podle znalosti parametru rozlišujeme testy**:
    - **Parametrické** – vyžadují znalost typu rozdělení a parametrů ZS (hypotézy se týkají hodnot parametrů rozdělení)
    - **Neparametrické** – neznáme typ rozdělení ani parametry rozdělení populační charakteristiky
  - **Podle počtu výběrových souborů rozlišujeme testy**:
    - Jednovýběrové
    - Dvouvýběrové
    - Vícevýběrové
  - **Podle stanovení alternativní hypotézy**
    - Oboustranné - $H_1: \Theta \neq \Theta_0$
    - Jednostranné
      - Pravostranné $H_1: \Theta > \Theta_0$
      - Levostranné $H_1: \Theta < \Theta_0$
  - Pro test nulové hypotézy proti alternativní hypotézy používáme speciální náhodnou veličinu, kterou nazýváme **testovací kritérium T (testová statistika)**
    - Charakterizuje stupeň nesouladu mezi tvrzením nulové hypotézy a pozorováním (nadměrnými daty)

5.1.2 Obor možných hodnot testovacího kritéria

- T rozdělujeme na dvě množiny:
  - **Kritický obor K** (obor zamítnutí H0)
  - **Obor přijetí P**
- Základní princip testování stat. hypotéz:
  - Padne-li vypočtená hodnota testovacího kritéria do kritického oboru, zamítáme H0 a přijímáme H1
  - Padne-li vypočtená hodnota testovacího kritéria do oboru přijetí, H0 nezamítáme
- **Rozhodovací pravidlo (oboustranný test)**

-

- Tα – kritická hodnota pro hladinu významnosti α

$|T| > T_\alpha$ **zamítneme H_0** na hladině významnosti α ve prospěch H_1

$|T| < T_\alpha$ **H_0 nelze** na hladině významnosti α **zamítnout**

- **Rozhodovací pravidlo – alternativní způsob (při využití SPSS)**
  - Založeno na p-hodnotě (sig)
  - **P < α –** zamítneme H0 na hladině významnosti α
  - **P > α –** nezamítneme H0 na hladině významnosti α
- Můžeme se dopustit dvou chybných závěrů:
  - **Chyba 1. druhu (α) –** spočívá v zamítnutí H0, ačkoliv je pravdivá
    - Čím menší α, tím menší chyba
  - **Chyba 2. druhu (β)** – spočívá v přijetí H0, i když je nesprávná
    - Její doplněk do 1, tzn. 1-β vyjadřuje *p* správného zamítnutí testované hypotézy
    - Doplněk nazýváme **síla testu** – tzn. *p*, že se nedopustíme chyby 2. druhu
  - P = 1-α
  -

		Závěr testu	
		H_0 platí	H_0 neplatí
Skutečnost	H_0 pravdivá	správné rozhodnutí 1 - α	**chyba 1 druhu α**
	H_0 nepravdivá	**chyba 2. druhu β**	správné rozhodnutí 1 - β

5.1.3 Obecný postup při testování hypotéz

1. Formulace H0 a H1
2. Volba hladiny významnosti α
3. Volba vhodné testové statistiky
4. Výpočet testového kritéria
5. Vymezení kritického oboru pro platnost nulové hypotézy
6. Rozhodnutí (p> α – zamítáme H0, p < α – nezamítáme H0)
7. Intepretace

5.2 Parametrické testy – jednovýběrové

5.2.1 Jednovýběrové testy

- Na základně jednoho souboru rozhodujeme, zda neznámý populační parametr je nebo není roven určité předpokládané číselné hodnotě, či zda je neznámý parametr větší(menší) než předpokládaná číselná hodnota
- **Test hypotézy o populačním rozptylu**
  - Předpoklad – výběr z normálního rozdělení N (μ, σ^2), kde jsou tyto parametry neznámé
  - **H_0: $\sigma^2 = \sigma^2_0$** (předpokládaná hypotetická hodnota rozptylu)
  - Rozdíl mezi odhadovaným rozptylem a předpokládanou hodnou může být:
    - Nevýznamný
    - Statisticky významný (nenáhodný)

-

  - Test je založen na ověření, zda výběrový rozptyl a předpokládaný rozptyl se iší statisticky významně nebo náhodně
    1. **$H_0: \sigma^2 = \sigma^2_0$**
       $H_1: \sigma^2 \neq \sigma^2_0$

- **Test hypotézy o populačním průměru**
  - Předpoklad – výběr ze souboru s normálním rozdělením
  - **$H_0: \mu = \mu_0$** (předpokládaná hypotetická hodnota)
  - Test při známém populačním rozptylu σ^2
    1. **$H_0: \mu = \mu_0$**
       $H_1: \mu \neq \mu_0$
- **Test hypotézy o relativní četnosti**
  - Předpoklad – dostatečně velký rozsah výběrového souboru (n>50)
  - **$H_0: \pi = \pi_0$**

1. Testujeme $H_0 : \mu = \mu_0$ proti $H_1 : \mu \neq \mu_0$ za předpokladu, že neznáme populačním rozptyl σ^2.
2. $\alpha = 0{,}05$
3. **One-Sample Test**

	Test Value = 55					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
rychlost	1,078	29	,290	1,433	-1,29	4,15

5. p = 0,290
6. $p < \alpha \Rightarrow H_0$
7. Řidiči dodržují předepsanou rychlost.

6 Parametrické testy – dvouvýběrové

- Umožňují porovnávat neznámé hodnoty parametru mezi dvěma základními soubory
- **Test hypotézy o populačních rozptylech**
- **Test hypotézy o populačních průměrech**
  - Závislé výběry
    - Jedna výběrová skupina, dvě šetření; stejné šetření v jiném čase
  - Nezávislé výběry
    - Dva výběry, jedno šetření; žádná vazba mezi výběry
- **Hypotézy o relativních četnostech**

6.1.1 Test H_0 o shodě dvou rozptylů

- Předpoklad – dva nezávislé výběry z populace s normálním rozdělením, parametry jsou neznámé
- **Test o shodě dvou rozptylů -> F-test** (SPSS: test homogenity rozptylů -> Leveneův test)
- **$H_0: \sigma_1^2 = \sigma_2^2$**

6.1.2 Test H_0 o shodě dvou průměrů

- Předpoklad – výběry z populace s normálním rozdělením
- **T-testy**
  - Nezávislé výběry
    - Dvouvýběrový t-test nebo Welchův test
  - Závislé výběry
    - párový t-test (**$H_0: \mu_1 = \mu_2$**)
- nejdřív shoda rozptylů (**$H_0: \sigma_1^2 = \sigma_2^2$**), pak shoda průměrů
- pokud jsou shodné tak dvouvýběrový test, pokud ne, tak Welchův test
- **$H_0: \mu_1 = \mu_2$**

7 Parametrické testy – vícevýběrové

7.1.1 Vícevýběrové testy

- Umožňují porovnávat neznámé hodnoty parametru mezi více než dvěma základními soubory
- **Test hypotéz o populačních rozptylech**
- **Test hypotézy o populačních průměrech – ANOVA**

-

7.1.2 Test H0 o shodě rozptylů

- Předpoklad – nezávislé výběry z populace s normálním rozdělením, kde jsou parametry neznámé
- **Cochranův test, Barlettův test** (SPSS: test homogenity rozptylů -> Leveneův test)
- $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 ..... = \sigma_k^2$

7.1.3 Test H0 o populačních průměrech

- Anova – analýza rozptylu
  - Původně vypracovaná pro potřeby zemědělského výzkumnictví ke statistickému hodnocení polních pokusů
  - Obecný statistický postup, využívaný v přírodních i společenských vědách, při zpracování kvantitativních výsledků experimentů

7.1.4 Analýza rozptylu

- Podle počtu sledovaných faktorů můžeme analýzu rozptylu rozdělit do 3 základních skupin:
  - **Analýza rozptylu jednoduchého třídění**
    - Vliv 1 faktoru na kvantitativní znak
  - **Analýza rozptylu dvojného třídění**
    - Vliv 2 faktoru na kvantitativní znak
  - **Vícefaktorové modely analýzy rozptylu**
- Pro zpracování modelů analýzy rozptylu je důležité, zda je při všech kombinacích faktorů realizován stejný počet měření či nikoliv
  - Rozsahy jednotlivých tříd jsou různé:
    - n_i **(*i* = 1, 2, …, *m*) –** nevyvážený neortogonální model
  - rozsahy v jednotlivých třídách jsou shodné:
    - $n_1 = n_2 = ... = n_m$ - vyvážený ortogonální model
- Obecně umožňuje vyhodnotit průkaznost rozdílu mezi průměry nezávislých výběrových souborů
- Soubor metod, pomocí kterých lze sledovat vliv jednoho nebo více faktorů na populační průměr
- **Podmínky analýzy:**
  - Data pocházejí z normálního rozdělení
  - Nezávislost náhodných chyb
  - ZS, ze kterých jsou pořízeny VS mají shodné rozptyly – **nejdůležitější podmínka**

7.1.5 ANOVA – jednoduchého třídění

- Sledovaný znak je ovlivňován pouze jediným faktorem, který sledujeme na několika jeho úrovních
- Úrovně faktoru:
  - Určité množství kvantitativního faktoru
  - Určitá varianta kvalitativního faktoru
- Naměřené hodnoty uspořádáme podle jednoho třídícího kritéria, tzn. podle úrovní sledovaného faktoru, do tolika tříd, na kolika úrovních tento faktor sledujeme

-

1. H_0: $\mu_1 = \mu_2 = ... = \mu_m$
   H_1: alespoň jedna dvojice průměrů se liší
2. Volba hladiny významnosti α
3. Volba testové statistiky
4. Výpočet statistiky $F = \frac{s_1^2}{s_r^2}$
5. Kritická hodnota $F_{\alpha(m-1;m(n-1))}$
6. Rozhodnutí
7. Odpověď $F > F_{\alpha(m-1;m(n-1))} \rightarrow$ **zamítáme** H_0

- pokud dojde k zamítnutí H0, činíme závěr, že alespoň jeden průměr se významně liší

- Je tedy třeba provést podrobnější vyhodnocení výsledků -> Post Hoc analýza (Scheffeho test, Duncanův test, …)
-

7.1.6 Analýza rozptylu

- <u>Předpoklady:</u>
  - Normalita rozdělení náhodné veličiny
  - Statistická nezávislost náhodných chyb
  - Shoda rozptylů náhodných chyb
- H_0: $\mu_1 = \mu_2 = ... = \mu_m$
- <u>Podstata metody:</u>
  - Rozklad celkové pozorované variability na dvě části
  - $S = S_1 + S_r$
    - S… celkový součet čtverců
    - S_1 … součet čtverců mezi řádky (charakterizuje vliv faktoru na sledovaný znak)
    - Sr … reziduální součet čtverců (reziduální = rozdíl; charakterizuje působení náhodných příčin)
  - Pro vyjádření vzorců se používá tzv. **tečkový způsob zápisu**
- V spss:
  - Ověřit normalitu dat
  - Analýza rozptylu (Leveneho test)
  - Anova
  - Scheffeho metoda (pokud zamítneme H0)

-

7.1.7 Testování předpokladů shody rozptylů

Leveneův test

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$

H_1: alespoň jedna dvojice rozptylů se liší

Test of Homogeneity of Variances

pH level

Levene Statistic	df1	df2	Sig.
,189	2	15	,830

$p > \alpha \rightarrow H_0$ nezamítáme

Podrobnější vyhodnocení výsledků analýzy rozptylu

Multiple Comparisons

Dependent Variable: Kyselost

Scheffe

(I) Země	(J) Země	Mean Difference (I-J)	Std. Error	Sig.
Aljaška	Florida	,22667	,18299	**,482**
	Texas	-,39000	,18299	**,138**
Florida	Aljaška	-,22667	,18299	,482
	Texas	-,61667*	,18299	**,015**
Texas	Aljaška	,39000	,18299	,138
	Florida	,61667*	,18299	,015

*. The mean difference is significant at the 0.05 level.

Statisticky významný rozdíl v kyselosti deště byl zjištěn mezi **Floridu** a **Texasem** ($p<\alpha$).

1. H_0: $\mu_1 = \mu_2 = \mu_3$
2. H_1: alespoň jedna dvojice průměru se liší
2. $\alpha = 0,05$ 3. **ANOVA**

Kyselost

	Sum of Squares		df		Mean Square		F	Sig.
Between Groups	S_1	1,168	f_1	2	s_1^2	,584	5,811	,014
Within Groups	S_r	1,507	f_r	15	s_r^2	,100		
Total	S	2,674	f	17				

4. 5.

6. $p < \alpha \rightarrow H_0$ zamítáme
7. Mezi alespoň dvěma zeměmi je statisticky významný rozdíl v kyselosti deště

8 Testy dobré shody

8.1 Testy dobré shody

- Ověření shody rozdělení s pravděpodobnostním modelem.
- Umožňují srovnání **empirického** (výběrového) rozdělení s jistým rozdělením teroetickým.
- Rozdělení populace z něhož byl výběrový soubor pořízen, je určitého konkrétního typu.
- **Hledáme teoretické rozdělení, jehož volba je založena na věcných úvahách o sledovaném jevu, popřípadě na základě odahu typu teoretického rozdělení z grafického vyobrazení výběrového rozdělení četností.**
  - χ^2 – test dobré shody
  - Kolmogorovův-Smirnovův test
  - Davidův test normality
  - SPSS: test normality -> Shapiro-Wilkův test

8.1.1 χ^2 – test dobré shody

- nejčastěji lze použít pro ověřování hypotéz:
  - H0 předpokládá, že výběr **pochází z rozdělení určitého typu** jehož parametry jsou dány (např. normální)
  - H0 předpokládá, že výběr pochází z populace, v níž jsou četnosti jednotlivých variant roztříděny dle nějakého znaku do *k* skupin (tříd) a podíly variant v populaci jsou rovny číslům *p0,1 … p0,k* (**libovolné pravděpodobnostní rozdělení** při neznámých parametrech rozdělení)
- **H_0: výběr pochází z populace s … rozdělením**
- **H_1: výběr nepochází z populace … rozdělením**

-

H_0: výběr pochází z populace s . . . . . rozdělením

H_1: non H_0

Testová statistika $\chi^2 = \sum_{j=1}^{k} \frac{(n_{ej} - n_{oj})^2}{n_{oj}}$

n_{ej}….*empirické (skutečné) četnosti v intervalu*

n_{oj}…. teoretické četností v j-té skupině (j = 1, 2, …k)

Kritická hodnota testové statistiky $\chi^2_{\alpha(k-c-1)}$

k… počet tříd – intervalů

c… počet parametrů ověřované distribuční funkce

Rozhodnutí

$\chi^2 > \chi^2_{\alpha(k-c-1)} \Rightarrow H_0$

zamítnutí nulové hypotézy ve prospěch alternativní hypotézy na hladině významnosti α.

Odpověď

Náhodný výběr není ze základního souboru s daným rozdělením pravděpodobností.

- **hlavní podmínkou použití testu**
  - Dostatečný rozsah výběrového soubor (n > 50)
  - Teoretické četnosti > 5

8.1.2 Kolmogorov - Smirnovův test

- Testování shody empirického rozdělení s pravděpodobnostním modelem
- Používá se k ověření hypotézy, zda pořízený výběr pochází z rozdělení se spojitou distribuční fcí F(x), která je plně specifikovaná
- Známe typ i příslušné parametry rozdělení
- Vychází z původních napozorovaných hodnot a nikoli z údajů setříděných do skupin; nedochází ke ztrátě informace obsažené ve výběru; test je založen na porovnávání kumulativních četností
- **H0: výběr pochází z populace s ….. | H1: výběr nepochází ….**

8.1.3 Davidův test normality

- Velmi malá síla testu, používá se jen pro rychlou informaci o normalitě rozdělení
- **H0: náhodný výběr pochází z normálního rozdělení**

9 Neparametrické testy

9.1 Neparametrické testy

9.1.1 Testy

- **Parametrické**
  - Testy hypotéz o parametrech rozdělení
  - Předpokládáme normalitu rozdělení
  - Dostatečný počet pozorování
- **Neparametrické**
  - Není nutná znalost tvaru rozdělení zkoumané veličiny
  - Použitelnost pro znaky kvantitativní a kvalitativní (ordinální data)
  - Výpočetní jednoduchost
  - Zejména pro výběry o malém rozsahu
  - Menší síla (menší schopnost zamítat nesprávnou nulovou hypotézu)

9.1.2 Neparametrické testy

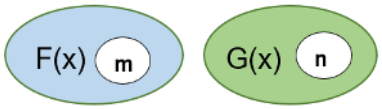
- Nejdůležitější podtřídu tvoří pořadové testy – pracuje s pořadovými čísly hodnot, očíslovaných a seřazených podle velikosti
  - **Nejmenší hodnota má pořadové číslo 1, největší *n***
  - Základní podmínkou pořadových testů je, aby byly spojité
- **Dvouvýběrový Wilcoxonův test (obdoba t-testu)**

-

- **Wilcoxonův test** (obdoba párového testu)
- **Znaménkový test** (obdoba párového testu)
- **Kruskal-Wallisův test** (Anova)
- **Némenyiho metoda** (obdoba T-metody)
- **Dunnova metoda** (obdoba S-metody)

9.1.3 Dvouvýběrový Wilcoxonův test, Mann-Whitneyův U test

- Ekvivalentní testy, které jsou neparametrickou obdobou t-testu pro dva nezávislé výběry
- Testujeme hypotézu, že dva nezávislé výběry **X a Y** o rozsazích **m a n** pocházejí ze stejného základního souboru (z populací se stejným mediánem)
- **H0: F(x) = G (x)**
- **H1: F(x) ≠ G(x)**
- **H_0: $\mu_1=\mu_2$**



- Postup:

  - Oba výběry spojíme do jednoho souboru (sdružený výběr) a hodnoty uspořádáme podle veliksoti
  - Přiřadíme jim pořadová čísla od 1 do N (m+n) – stejným hodnotám přiřadíme tzv. průměrné pořadí
  - Vypočteme veličiny

Byly sledovány výkony běžců ze dvou klubů, přičemž v každém klubu jsou používány částečně odlišné tréninkové metody. Posuďte, zda se výsledky běžců klubu A a B liší. (α=0,05)

	Čas potřebný na uběhnutí 100 metrů (sekund)						
klub A	11,1	10,5	10,0	9,7	12,0	10,7	10,9
klub B	10,3	11,0	10,1	9,7	9,5	10,8	

Ranks

	klub	N	Mean Rank	Sum of Ranks
čas	klub A	7	8,07	56,50
	klub B	6	5,75	34,50
	Total	13		

SUM of Rank → součet těch pořadí

Test Statistics[a]

	čas
Mann-Whitney U	13,500
Wilcoxon W	34,500
Asymp. Sig. (2-tailed)	,283

a. Grouping Variable: klub

Alternativní hypotéza → nejsou stejné
Nulová hypotéza → Výsledky sportovního klubu A a B se od sebe statisticky významně neliší, výsledky jsou stejné
Název testu (symbol), testová statistika (když nevím jaký je to test)

9.1.4 Znaménkový test

- Porovnání dvou závislých výběrů, obdoba párového t-testu
- Lehký vpočet, malá síla
- Ověřujeme hypotézu, zda závislé výběry se významně liší svou polohou

9.1.5 Wilcoxonův test

- Porovnání dvou závislých výběrů, obdoba párového t-testu
- Slouží k ověřování hypotézy o shodě úrovně ve dvou souborech, z nichž byly pořízeny párové výběry
- Postup:
  - Pro každou dvojici závislých pozorování vypočteme diferenci (di = xi – yi)
  - Absolutním hodnotám diferencí přiřadíme pořadová čísla (vynecháme nulové)
  - Sečteme zvlášť pořadová čísla kladných a záporných diferencí
  - Testové kritérium je menší z hodnot W+ a W- **(W = min(W+, W-)**
  - W<Walfa zamítá se H0

-

- **H_0: $\mu_1=\mu_2$**

9.1.6 Kruskal-Wallisův

- Neparametrická obdoba analýzy rozptylu jednoduché třídění
  - Hodnoty seřadíme do jedné rostoucí posloupnosti a každé hodnotě přiřadíme pořadové číslo, při shodě přiřazujeme průměrné pořadí
- **H_0: $\mu_1 = \mu_2 = \ldots\ldots = \mu_n$**

9.1.7 Neparametrické metody mnohonásobného porovnávání

- Jestliže v závěru Kruskal-Wallisova testu zamítáme H0, pokračujeme porovnáním každé dvojice výběru (obdoba Scheffeho)
  - **Némenyiho metody (T-metoda)**
    - H0: i-tý a j-tý výběr pochází ze stejného rozdělení
    - pro vyvážené modely
    - postup při této metoda spočívá v porovnání všech diferencí s kritickou hodnotou
  - **Dunnova metoda**
    - Pro vyvážené modely
    - H0: diference mezi výběry není průkazná