Analýza dat pro Neurovědy

RNDr. Eva Janoušová

doc. RNDr. Ladislav Dušek, Dr.

Jaro 2012

Blok 3

Jak a kdy použít parametrické a neparametrické testy II.

Parametrické a neparametrické testy pro kvantitativní data – přehled

Typ srovnání	Parametrický test	Neparametrický test
1 skupina dat s referenční hodnotou – jednovýběrové testy:	Jednovýběrový t-test, jednovýběrový z-test	Wilcoxonův test
2 skupiny dat párově – párové testy:	Párový t-test	Wilcoxonův test, znaménkový test
2 skupiny dat nepárově – dvouvýběrové testy:	Dvouvýběrový t-test	Mannův-Whitneyův test, mediánový test
Více skupin nepárově:	ANOVA	Kruskalův- Wallisův test

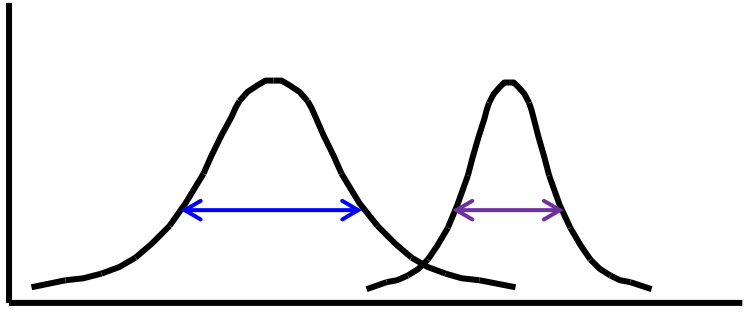
Osnova

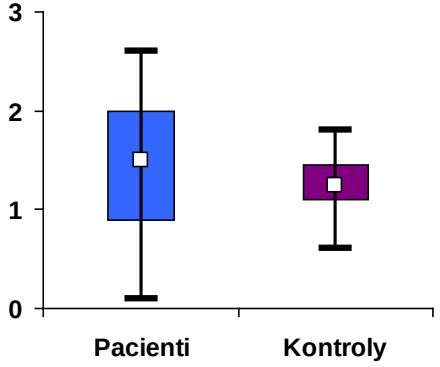
1. F-test
2. Analýza rozptylu (ANOVA) a její předpoklady
3. Problém násobného testování hypotéz a použití korekčních procedur
4. Kruskalův-Wallisův test
5. Analýza rozptylu jako lineární model

1. F-test

F-test

- **Srovnáváme rozptyly (variabilitu) dvou skupin dat, které jsou na sobě nezávislé** (mezi objekty neexistuje vazba).
- F-test patří mezi dvouvýběrové parametrické testy.
- Příklady: srovnání variability objemu hipokampu u pacientů s AD a kontrol.
- Použití: ověření předpokladu shodnosti (homogenity) rozptylů u dvouvýběrového t-testu.





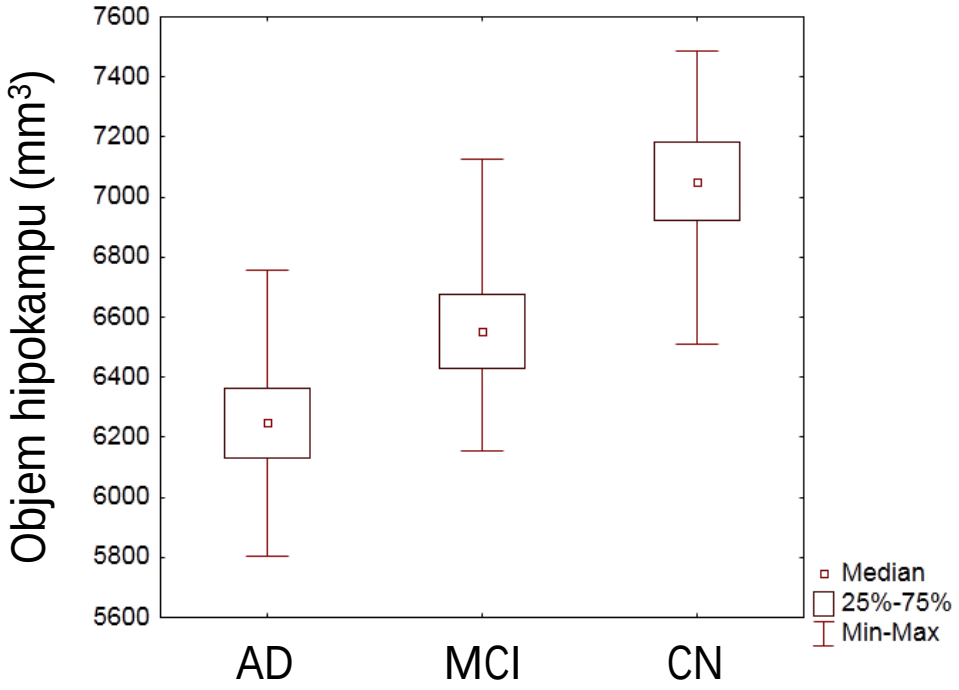
- Předpoklad: normalita dat v OBOU skupinách.
- Testová statistika: $F = \frac{s_1^2}{s_2^2}$, kde s_1^2 je rozptyl prvního výběru a s_2^2 je rozptyl druhého výběru

F-test

- **Příklad:** Chceme srovnat, zda se liší variabilita objemu putamenu podle pohlaví.
- Tzn. hypotézy budou mít tvar: $H_0 : s_M^2 = s_Z^2$ a $H_1 : s_M^2 \neq s_Z^2$
- **Postup:**
  1. Ověření normality hodnot v OBOU skupinách pomocí histogramu (tzn. vykreslíme histogram zvlášť pro muže a zvlášť pro ženy).
  2. Vykreslení krabicových grafů, které nám napoví, zda máme očekávat shodu nebo neshodu rozptylů.
  3. Aplikujeme statistický test (F-test je součástí dvouvýběrového t-testu v softwaru STATISTICA (tedy zvolíme t-test, independent, by groups)).
  4. Nulovou hypotézu zamítneme nebo nezamítneme:
     p=0,935 > 0,05 → nezamítáme nulovou hypotézu → Neprokázali jsme rozdíl ve variabilitě objemu putamenu podle pohlaví (na hladině významnosti α=0,05.)

2. Analýza rozptylu (ANOVA) a její předpoklady

Motivace



Jak můžeme ověřit, zda se liší objem hipokampu u pacientů s AD, pacientů s MCI a u zdravých kontrol?

A. Můžeme použít vhodný test pro dva výběry (např. dvouvýběrový t-test) a otestovat, jak se liší AD od MCI, AD od CN a MCI od CN – tedy provést 3 testy.

B. Můžeme použít vhodný test pro více než dvě srovnávané skupiny.

V čem je zásadní rozdíl mezi A a B?

Motivace – pokračování

- Problém s možností A je v **násobném testování hypotéz**:

  S narůstajícím počtem testovaných hypotéz nám roste také pravděpodobnost získání falešně pozitivního výsledku, tedy pravděpodobnost toho, že se při našem testování zmýlíme a ukážeme na statisticky významný rozdíl tam, kde ve skutečnosti žádný neexistuje (chyba I. druhu).

- Máme tři testy, v každém 95% pravděpodobnost, že neuděláme chybu I. druhu.
- Pro všechny tři testy to tedy znamená: 0,95 × 0,95 × 0,95 = 0,857.
- Pravděpodobnost, že neuděláme chybu I. druhu nám celkově klesla na 0,857.
- Pravděpodobnost, že uděláme chybu I. druhu nám celkově stoupla na 0,143.

Motivace – pokračování

- Lepší volbou je:

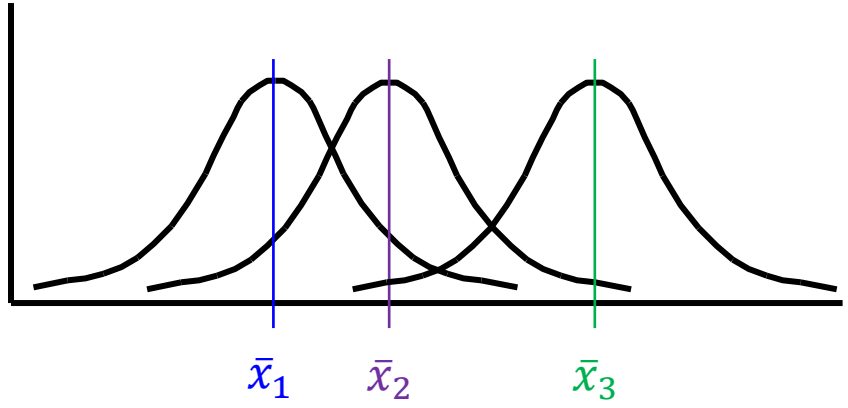
  B. Použít vhodný test pro více než dvě srovnávané skupiny.

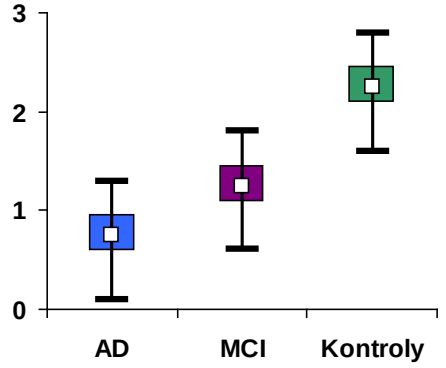
- **Analýza rozptylu (ANOVA = „ANalysis Of VAriance“) je statistickou metodou, která umožňuje testovat rozdíl v průměrech více než dvou skupin. Přitom se jedná o jeden test.**

- Více než dvě skupiny mohou být dány přirozeně (např. sledujeme rozdíl mezi věkovými kategoriemi) nebo uměle (např. sledujeme rozdíl v účinnosti několika typů léčby).

Analýza rozptylu (ANOVA) jednoduchého třídění

- **Srovnáváme tři a více skupin dat, které jsou na sobě nezávislé** (mezi objekty neexistuje vazba).
- Příklady: srovnání objemu hipokampu u pacientů s AD, pacientů s MCI a kontrol, srovnání kognitivního výkonu podle čtyř kategorií věku.





- Předpoklady: **normalita dat ve VŠECH skupinách**, **shodnost (homogenita) rozptylů VŠECH srovnávaných skupin**, nezávislost jednotlivých pozorování.
- Testová statistika: $F = \frac{S_A / df_A}{S_e / df_e}$ - vysvětlení později

Analýza rozptylu (ANOVA) jednoduchého třídění

- **Příklad:** Chceme srovnat, zda se liší objem hipokampu podle typu onemocnění (tzn. u pacientů s AD, pacientů s MCI a zdravých kontrol).
- Tzn. hypotézy budou mít tvar: $H_0 : m_{AD} = m_{MCI} = m_{CN}$

  $H_1 :$ nejméně jedno m_i je odlišné od ostatních

- **Postup:**
  1. Popisná sumarizace objemu hipokampu podle typu onemocnění.
  2. Ověření normality hodnot ve VŠECH skupinách.
  3. Ověření shodnosti rozptylů VŠECH skupin.
  4. Aplikujeme statistický test.
  5. Nulovou hypotézu zamítneme nebo nezamítneme:
     p<0,001 < 0,05 → zamítáme nulovou hypotézu → Rozdíl v objemu hipokampu podle typu onemocnění je statisticky významný (na hladině významnosti α=0,05.)

Ověření normality dat

- **Graficky:**
  - histogram
  - krabicový graf (box-plot)
  - Q-Q graf

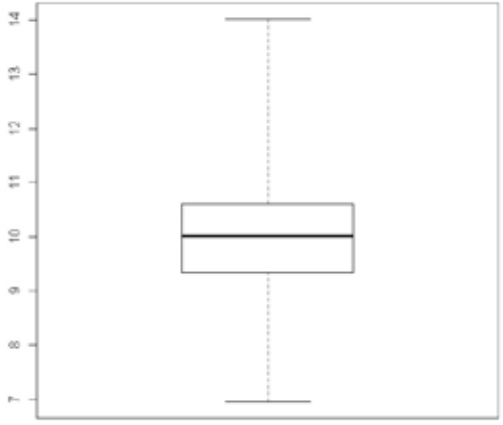
- **Testy normality:**
  - Shapirův-Wilkův test
  - Kolmorovův-Smirnovův test

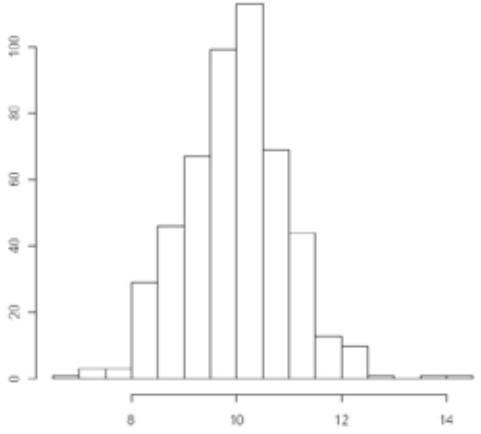
- Testy nejsou vždy nejlepším nástrojem! Vždy je důležité se podívat i očima!

- Pokud o sledované veličině prokazatelně víme, že v cílové populaci nabývá normální rozdělení (např. výška lidské postavy), ale v daném souboru normální rozdělení nepotvrdíme, **pak s naším náhodným výběrem není něco v pořádku** – např. není reprezentativní.

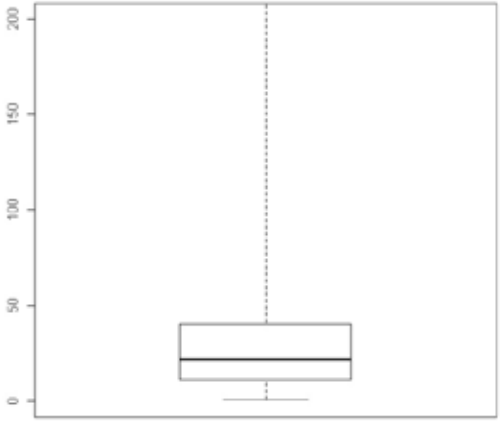
Ověření normality graficky – krabicový graf a histogram

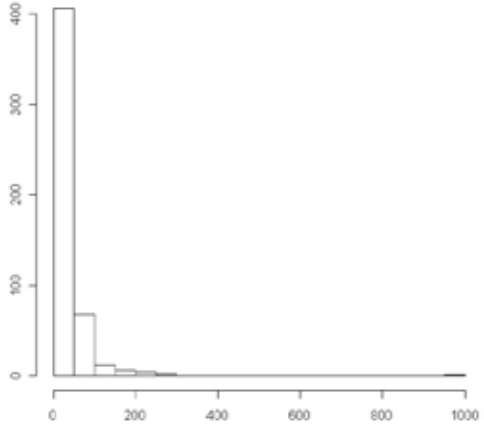
- Normální rozdělení





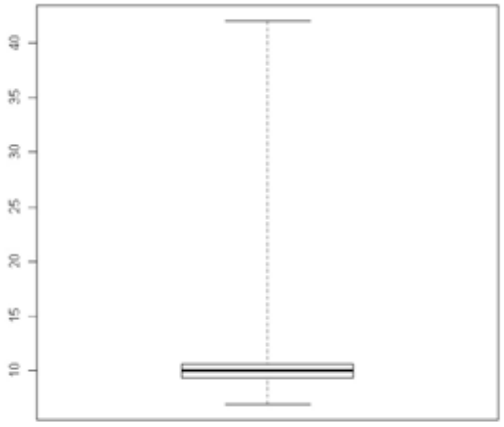
- Log-normální rozdělení

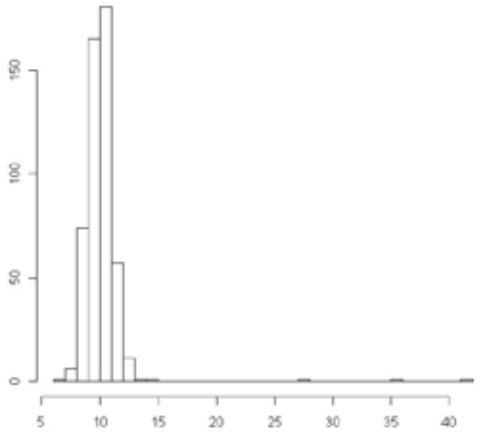




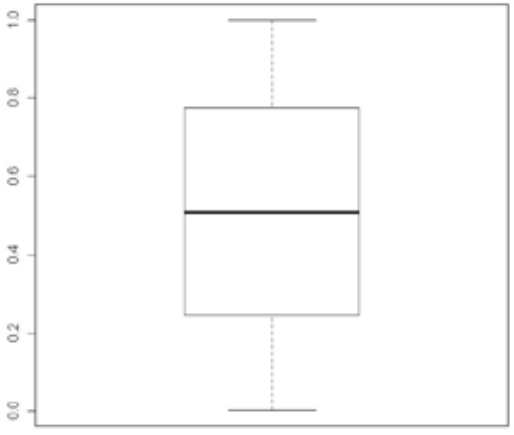
Ověření normality graficky – krabicový graf a histogram

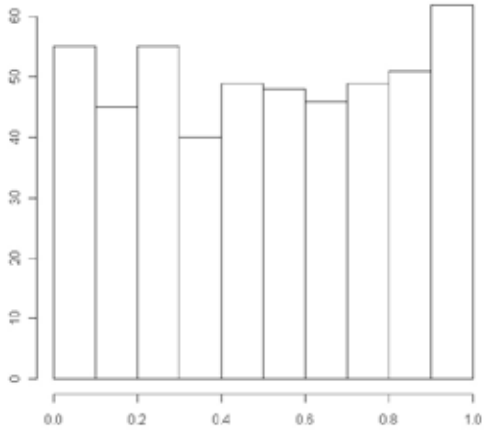
- Normální rozdělení s odlehlými hodnotami





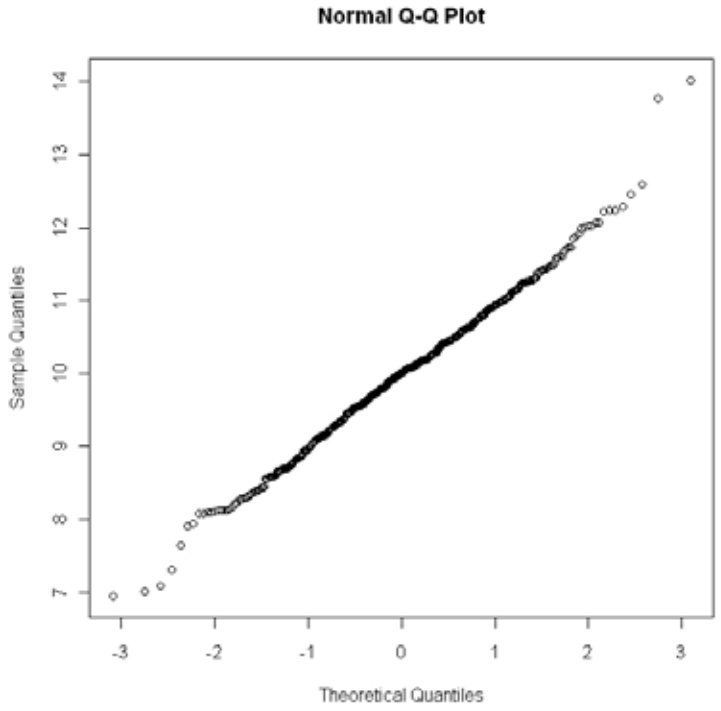
- Rovnoměrně spojité rozdělení





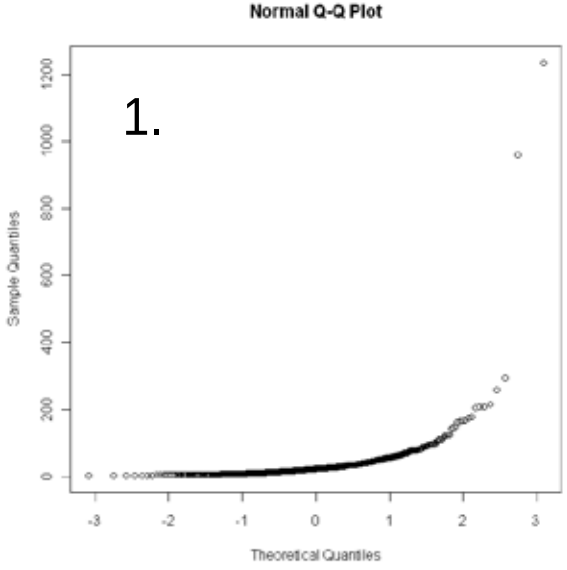
Ověření normality graficky – Q-Q graf

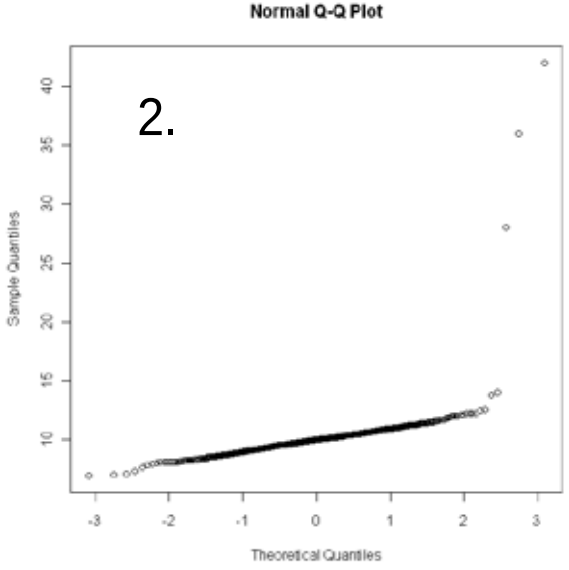
- Q-Q graf proti sobě zobrazuje kvantily pozorovaných hodnot a kvantily teoretického rozdělení pravděpodobnosti (zde normálního rozdělení).
- V případě shody leží všechny body na přímce.
- Normální rozdělení:

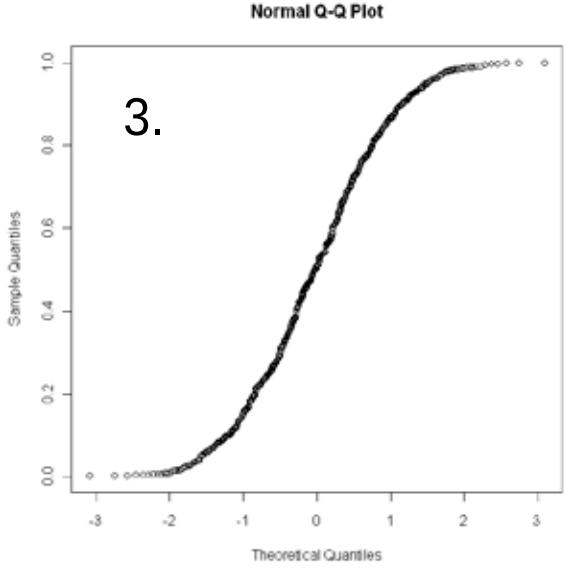


Ověření normality graficky – Q-Q graf

1. Log-normální rozdělení
2. Normální rozdělení s odlehlými hodnotami
3. Rovnoměrně spojité rozdělení







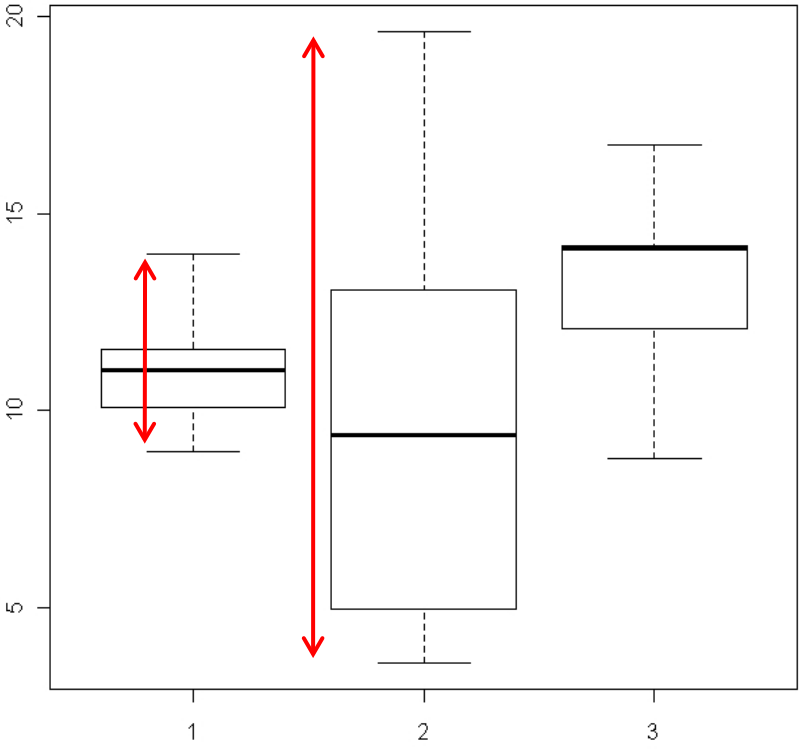
Ověření normality pomocí testů

- **Shapirův-Wilkův test** – v podstatě se jedná o proložení seřazených hodnot regresní přímkou vzhledem k očekávaným hodnotám normálního rozdělení. Má tedy přímý vztah k Q-Q plotu – vyhodnocuje, jak moc se Q-Q plot liší od ideální přímky. Doporučován pro menší vzorky, může být „moc“ přísný pro velké vzorky.

- **Kolmogorovův-Smirnovovův test** – založen na srovnání výběrové distribuční funkce s teoretickou distribuční funkcí odpovídající normálnímu rozdělení. K-S test hodnotí maximální vzdálenost mezi těmito dvěma distribučními funkcemi. V praxi se používá korekce dle Lillieforse.

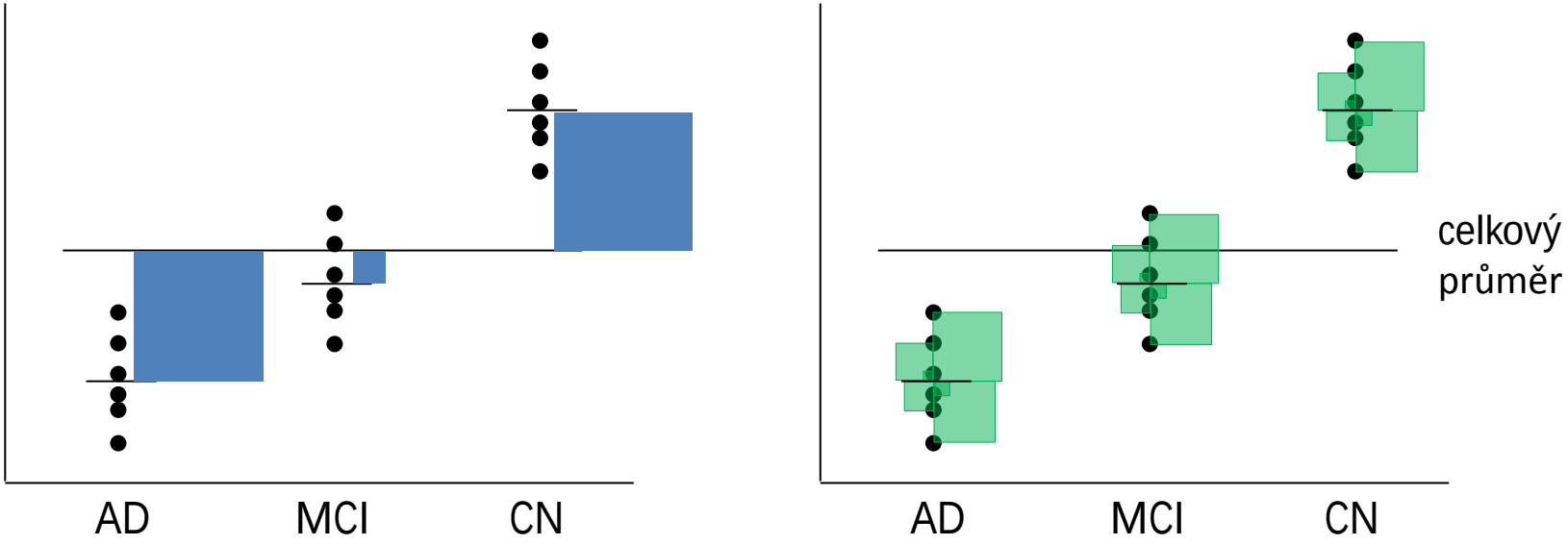
Ověření shody (homogenity) rozptylů

- **Grafické ověření** – krabicový graf, histogram.
- **Leveneův test** – často používaný.
- **Bartlettův test**



Analýza rozptylu (ANOVA) – princip

- Srovnání variability (rozptylu) mezi výběry s variabilitou uvnitř výběrů.

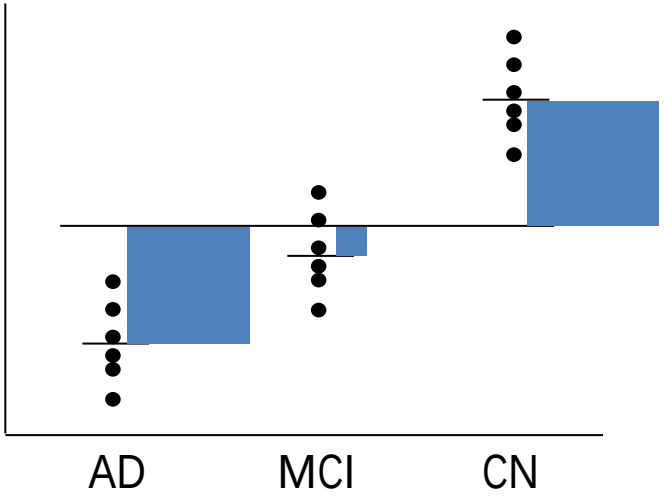


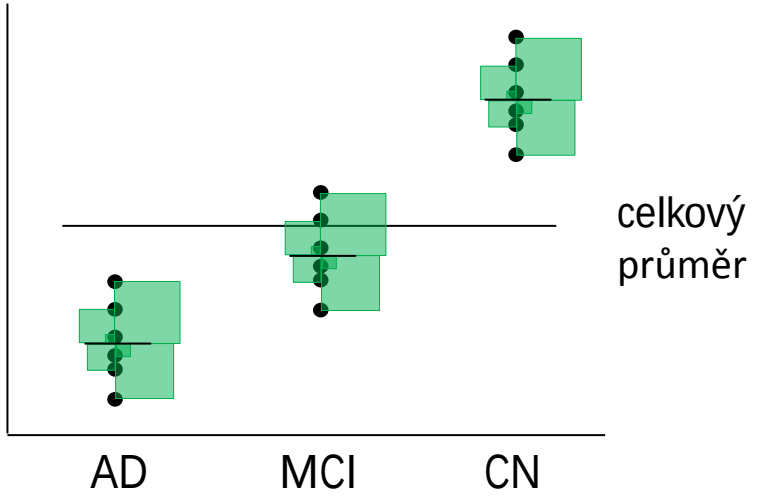
- Tabulka analýzy rozptylu jednoduchého třídění (One-Way ANOVA):

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	F statistika	p-hodnota
Mezi skupinami	S_A	$df_A = k - 1$	$MS_A = S_A/df_A$	$F = \frac{S_A / df_A}{S_e / df_e}$	p
Uvnitř skupin (reziduální var.)	S_e	$df_e = n - k$	$MS_e = S_e/df_e$		
Celkem	S_T	$df_T = n - 1$			

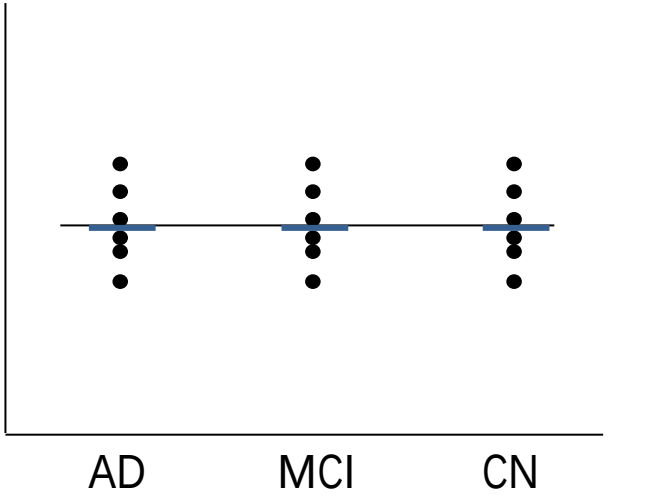
ANOVA – 2 ukázkové situace

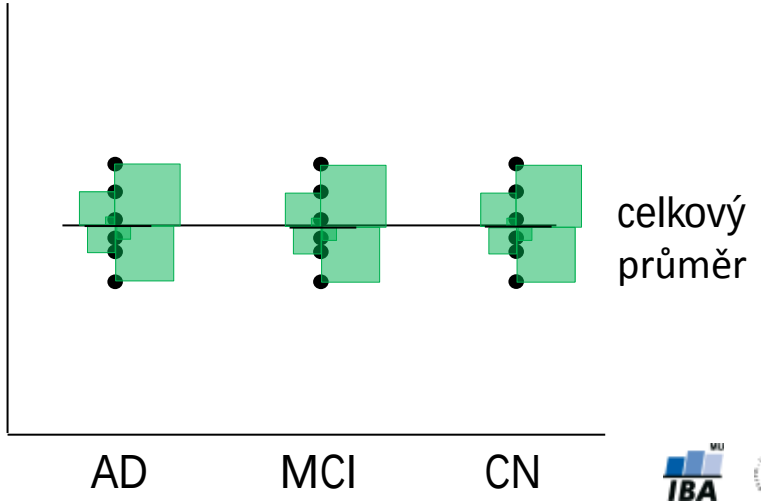
- Rozdíl ve všech třech skupinách:





- Žádný rozdíl mezi skupinami:





Výsledky ANOVA testu

- Tabulka analýzy rozptylu jednoduchého třídění:

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	F statistika	p-hodnota
Mezi skupinami	S_A = 71 422 222	$df_A = k - 1$ = 2	$MS_A = S_A/df_A$ = 35 711 111	$F = \frac{S_A / df_A}{S_e / df_e} = 1103{,}6$	0,00
Uvnitř skupin (reziduální var.)	S_e = 26 857 142	$df_e = n - k$ = 830	$MS_e = S_e/df_e$ = 32 358		
Celkem	S_T = 98 279 364	$df_T = n - 1$ = 832			

- Výsledek ze softwaru STATISTICA:

Variable	Analysis of Variance (Data_neuro_vycistena2) Marked effects are significant at p < ,05000							
	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
Hippocampus volume (mm3)	71422222	2	35711111	26857142	830	32358,00	1103,625	0,00

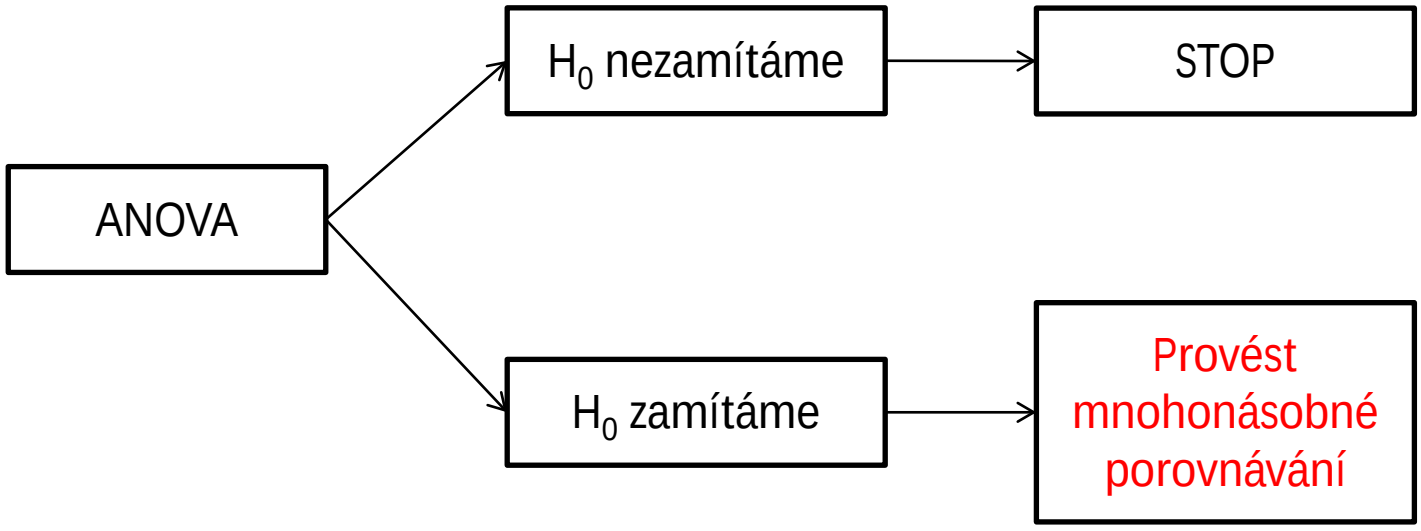
Analýza rozptylu (ANOVA) jednoduchého třídění

- **Příklad:** Chceme srovnat, zda se liší objem hipokampu podle typu onemocnění (tzn. u pacientů s AD, pacientů s MCI a zdravých kontrol).
- Tzn. hypotézy budou mít tvar: $H_0 : m_{AD} = m_{MCI} = m_{CN}$

  $H_1 :$ nejméně jedno m_i je odlišné od ostatních

- **Postup:**
  1. Popisná sumarizace objemu hipokampu podle typu onemocnění.
  2. Ověření normality hodnot ve VŠECH skupinách.
  3. Ověření shodnosti rozptylů VŠECH skupin.
  4. Aplikujeme statistický test.
  5. Nulovou hypotézu zamítneme nebo nezamítneme:
     p<0,001 < 0,05 → zamítáme nulovou hypotézu → Rozdíl v objemu hipokampu podle typu onemocnění je statisticky významný (na hladině významnosti α=0,05.)

Další kroky analýzy



3. Problém násobného testování hypotéz a použití korekčních procedur

Korekce na násobné srovnání výběrů

- Zamítneme-li analýzou rozptylu nulovou hypotézu o celkové rovnosti středních hodnot, má smysl se ptát, jaké skupiny se od sebe nejvíce liší.
- Toto srovnání lze provést pomocí testů pro dva výběry, ale je nutné korigovat výslednou hladinu významnosti testu, abychom se vyhnuli chybě I. druhu.
- Nejjednodušší metoda: **Boferroniho procedura** - korekce hladiny významnosti: $\alpha^* = \alpha/m$, kde m je počet provedených testů. Ekvivalentně lze vynásobit p-hodnotu počtem provedených testů. Nevýhodou je, že je konzervativní pro velké m, tedy počet provedených testů.
- Pro analýzu rozptylu: **Tukeyho** a **Schefféhо post hoc testy**.
- Může se stát, že při použití různých korekcí nám mohou vyjít výsledky různě (např. při použití Schefféhо testu nám vyjde statisticky významný rozdíl mezi skupinou AD a MCI a při použití Tukeyho testu nám rozdíl statisticky významný nevyjde).

Poznámka

- Může nastat situace, kdy zamítneme H_0 u ANOVY, ale metodami mnohonásobného porovnávání nenajdeme významný rozdíl u žádné dvojice středních hodnot. K tomu dochází zvláště tehdy, když p-hodnota pro ANOVU je jen o málo nižší než zvolená hladina významnosti.

- Důvod: post-hoc testy (tzn. metody mnohonásobného porovnávání) mají obecně menší sílu než ANOVA, proto nemusí odhalit žádný rozdíl.

Korekce na násobné srovnání – jiná situace

- Problém násobného testování („Multiple Testing Problem“) nastává, i když je provedeno **větší množství testů na různých proměnných** v rámci jednoho hodnocení dat.
- Příklad: zjišťování, zda se liší objem šedé hmoty u dvou skupin subjektů v každém voxelu obrazu.

- Korekce:
  - **Bonferroniho korekce** – kontroluje pravděpodobnost, s jakou dostaneme falešně pozitivní výsledek (kontroluje chybu I. druhu); konzervativní pro velký počet provedených testů.
  - **False discovery rate (FDR)** – kontroluje podíl falešně pozitivních výsledků mezi všemi statisticky významnými výsledky (např. pokud je FDR 0,05 a počet všech statisticky významných výsledků bude 1000, tak můžeme očekávat, že 50 výsledků bude falešně pozitivních).

Úkol 1.

- **Zadání:** Zjistěte, zda se liší objem pallida podle typu onemocnění (nezapomeňte ověřit předpoklady).

- **Řešení:**

Variable	Analysis of Variance (Data_neuro_vycistena2) Marked effects are significant at p < ,05000							
	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
Pallidum_volume (mm3)	229575,6	2	114787,8	34702692	830	41810,47	2,745432	0,064804

4. Kruskalův-Wallisův test

Co dělat, když nejsou splněny předpoklady u ANOVy?

1. **Zkusit data transformovat** – např. logaritmická transformace by měla pomoci s normalizací rozdělení a stabilizací rozptylu u log-normálních dat.

2. **Použít neparametrické testy** – např. Kruskalův-Wallisův test nevyžaduje předpoklad normality, pracuje stejně jako neparametrický Mannův-Whitneyův test.

Kruskalův-Wallisův test

- Neparametrická alternativa analýzy rozptylu (ANOVy).
- Testuje se, zda jsou srovnatelné distribuční funkce (obdobně jako u Mannova-Whitneyova testu).
- Hypotézy mají tvar: $H_0 : F_1(x) = F_2(x) = ... = F_k(x)$

  $H_1 :$ nejméně jedna F_i je odlišná od ostatních

- Princip Kruskalova-Wallisova testu (podobný jako u Mannova-Whitneyova testu):
  1. Všechny hodnoty ze všech výběrů dohromady uspořádáme vzestupně podle velikosti → každé hodnotě přiřadíme pořadí.
  2. Spočítáme součet pořadí hodnot u každého výběru.
  3. Na základě těchto dvou součtů vypočteme testovou statistiku.
- Tzn. za platnosti nulové hypotézy jsou spojená data dobře promíchaná a průměrná pořadí v jednotlivých souborech jsou podobná.
- Odlehlé hodnoty nejsou problém, protože pracujeme s pořadími.

Kruskalův-Wallisův test

- **Příklad:** Chceme srovnat, zda se liší MMSE skóre podle typu onemocnění.
- Tzn. hypotézy budou mít tvar: $H_0 : F_{AD}(x) = F_{MCI}(x) = F_{CN}(x)$

  $H_1 :$ nejméně jedna F_i je odlišná od ostatních

- **Postup:**
  1. Popisná sumarizace MMSE skóre podle typu onemocnění.
  2. Vykreslení histogramů MMSE skóre pro jednotlivé skupiny subjektů, abychom viděli, že není splněn předpoklad normálního rozdělení → proto použijeme neparametrický test.
  3. Aplikujeme statistický test.
  4. Nulovou hypotézu zamítneme nebo nezamítneme:
     p<0,001 < 0,05 → zamítáme nulovou hypotézu → **MMSE skóre je u pacientů s AD, MCI a u kontrol statisticky významně odlišné.**

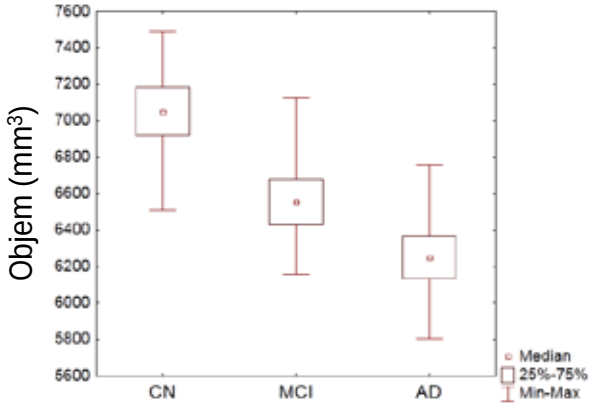
Úkol 2.

- **Zadání:** Zjistěte, zda se liší objem pěti mozkových struktur podle typu onemocnění (použijte Kruskalův-Wallisův test).

Výsledky srovnání objemů mozkových podle typu onemocnění

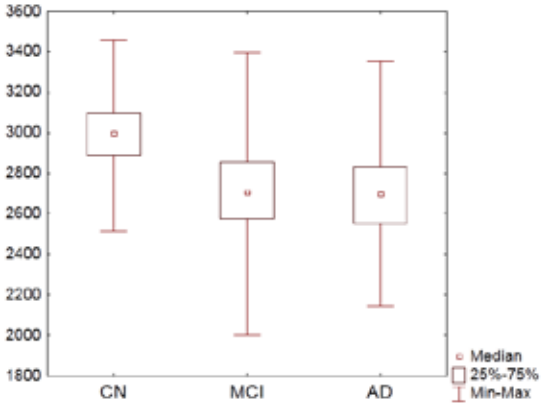
Hipokampus (p < 0,001*)

* Statisticky významný rozdíl: ADxMCI, ADxCN, MCIxCN

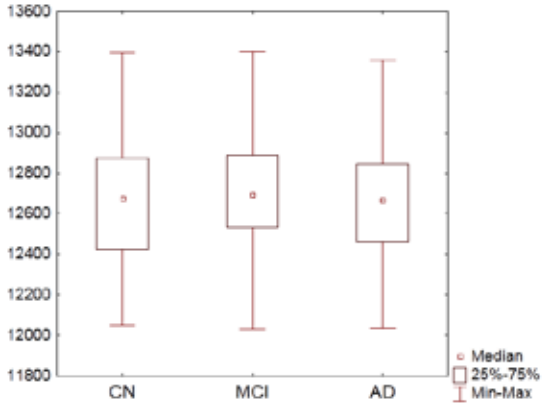


Amygdala (p < 0,001*)

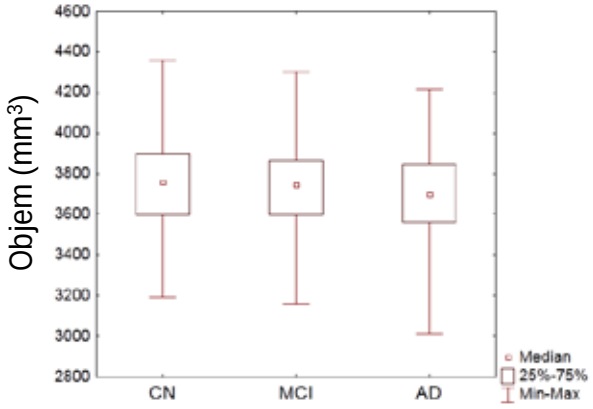
* Statisticky významný rozdíl: ADxCN, MCIxCN



Thalamus (p = 0,214)

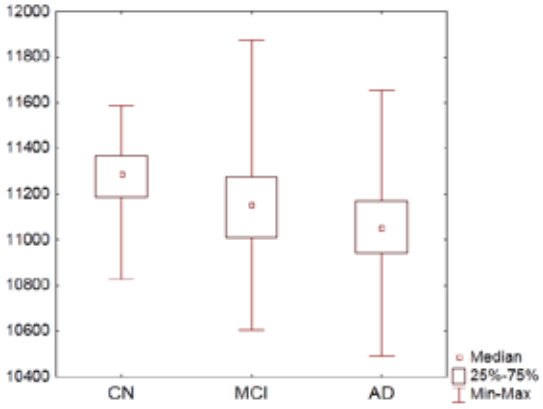


Pallidum (p = 0,078)



Putamen (p < 0,001*)

* Statisticky významný rozdíl: ADxMCI, ADxCN, MCIxCN



Úkol 3.

- **Zadání:** Zjistěte, zda se liší váha podle typu onemocnění. Pokud nejsou splněny předpoklady, zkuste váhu logaritmovat. Proveďte i popisnou sumarizaci váhy podle typu onemocnění včetně výpočtu intervalů spolehlivosti.

- **Řešení:**

	N	Geometrický průměr	Dolní mez IS	Horní mez IS	Medián	Minimum	Maximum
CN	230	76,9	75,3	78,5	76,0	52,0	135,0
MCI	406	75,4	74,1	76,7	75,5	52,0	140,0
AD	197	70,3	68,6	71,9	70,0	44,0	106,0

$p<0,001$*

*Statisticky významný rozdíl: ADxMCI, ADxCN

5. Analýza rozptylu jako lineární model

Analýza rozptylu jako lineární model

- Analýza rozptylu pro jednu vysvětlující proměnnou (jednoduché třídění) lze zapsat jako lineární model:

$$Y_{ij} = \mu_i + e_{ij} = \mu + \alpha_i + e_{ij}$$

Reziduum

Populační průměr

***i*-tý efekt faktoru A**

- Nulovou hypotézu pak lze vyjádřit jako: $H_0 : a_1 = a_2 = \mathrm{K} = a_k$

- **Rozšířením tohoto zápisu můžeme definovat další modely ANOVA**: více faktorů, hodnocení interakcí, opakovaná měření na jednom subjektu.

Analýza rozptylu dvojného třídění

- Uvažujeme dvě vysvětlující proměnné zároveň.
- Zápis modelu:

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

Reziduum

j-tý efekt faktoru B

Populační průměr

i-tý efekt faktoru A

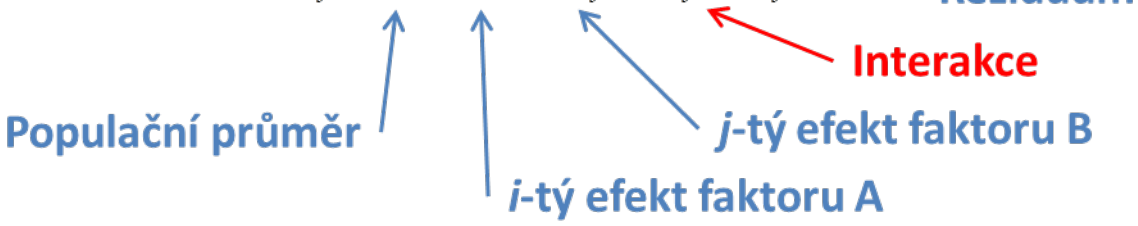
- Nulové hypotézy pak máme dvě: $H_{01} : a_1 = a_2 = \mathrm{K} = a_k$, $H_{02} : b_1 = b_2 = \mathrm{K} = b_r$

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	*F* statistika	*p*-hodnota
Faktor A	S_A	$df_A = k - 1$	$MS_A = S_A / df_A$	F_A	p
Faktor B	S_B	$df_A = r - 1$	$MS_B = S_B / df_B$	F_B	p
Rezidua	S_e	$df_e = (k - 1)(r - 1)$	$MS_e = S_e / df_e$		
Celkem	S_T	$df_T = n - 1 = kr - 1$			

Analýza rozptylu dvojného třídění s interakcí

- Uvažujeme dvě vysvětlující proměnné a zároveň i jejich společné působení.
- Zápis modelu:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ij}$$



- Nulové hypotézy pak máme tři:

$$H_{01} : g_{11} = g_{12} = \mathrm{K} = g_{kr} \qquad H_{02} : a_1 = a_2 = \mathrm{K} = a_k \qquad H_{03} : b_1 = b_2 = \mathrm{K} = b_r$$

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	*F* statistika	*p*-hodnota
Faktor A	S_A	$df_A = k - 1$	$MS_A = S_A / df_A$	F_A	p
Faktor B	S_B	$df_A = r - 1$	$MS_B = S_B / df_B$	F_B	p
Interakce A×B	S_{AB}	$df_{AB} = (k-1)(r-1)$	$MS_{AB} = S_{AB} / df_{AB}$	F_{AB}	p
Rezidua	S_e	$df_e = n - kr$	$MS_e = S_e / df_e$		
Celkem	S_T	$df_T = n - 1$			

Blok 4

Jak analyzovat kategoriální a binární data.

Typy dat - opakování

- **Kvalitativní (kategoriální) data:**
  - Binární data

  - Nominální data

  - Ordinální data

- **Kvantitativní data:**
  - Intervalová data

  - Poměrová data

Osnova

1. Binomické rozdělení
2. Poissonovo rozdělení

Motivace

- Kromě spojitých dat se setkáváme také s daty kategoriálními, jejichž nejjednodušším případem jsou data binární.
- Binární data jsou popsána binomickým rozložením.
- Od chování binomického rozložení je odvozena:
  - popisná statistika binárních dat (procento výskytu jevu)
  - interval spolehlivosti pro binární data
  - binomické testy pro srovnání procentuálního výskytů jevů v různých skupinách.

Binomické rozdělení

Binomické rozdělení

- Diskrétní rozdělení, které **popisuje počet výskytů sledované události** (ve formě nastala/nenastala) **v sérii *n* nezávislých pokusech**, kdy v každém pokusu **je stejná pravděpodobnost výskytu** této události.
- Značení: Bi(n,π)
- Parametry:

  n ... počet nezávislých pokusů

  r ... počet, kolikrát nastala sledovaná událost (r = 0...n)

  p = r/n ... pravděpodobnost nastání sledované události (p ~π)
- Pravděpodobnost, že sledovaná událost nastane r-krát, lze vypočítat:

$$P(X = r) = \binom{n}{r} p^r (1 - p)^{n-r} = \frac{n!}{r!(n - r)!} \times p^r \times (1 - p)^{n-r}$$

- **Střední hodnota:** $EX = n \cdot p$
- **Rozptyl:** $DX = n \cdot p \cdot (1 - p)$
- Příklady: výskyt nežádoucích účinků léku u léčených pacientů, počet zemřelých pacientů mezi léčenými pacienty, počet pacientů s výsledkem neuropsychologického testu pod normou

Binomické rozdělení – příklad

- **Př. Pravděpodobnost narození chlapce je 0,5. Jaká je pravděpodobnost toho, že mezi čtyřmi dětmi v rodině je 0, 1,... až 4 chlapců. Vypočítejte i jaký je nejpravděpodobnější počet chlapců v této rodině.**

- Řešení: n = 4 (4 děti v rodině)
  r = 0, 1, 2, 3, 4 chlapců

$$P(X = r) = \binom{n}{r} p^r (1 - p)^{n-r} = \frac{n!}{r!(n-r)!} \times p^r \times (1 - p)^{n-r}$$

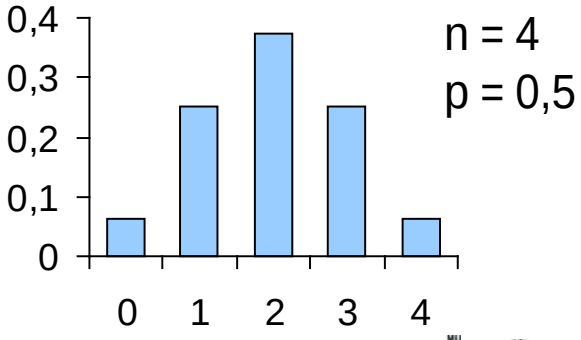
$$P(X = 0) = \frac{4!}{0!\,4!} \times 0{,}5^0 \times (1 - 0{,}5)^4 = 0{,}0625$$

$$P(X = 1) = \frac{4!}{1!\,3!} \times 0{,}5^1 \times (1 - 0{,}5)^3 = 0{,}2500$$

$$P(X = 2) = \frac{4!}{2!\,2!} \times 0{,}5^2 \times (1 - 0{,}5)^2 = 0{,}3750$$

$$P(X = 3) = 0{,}2500$$

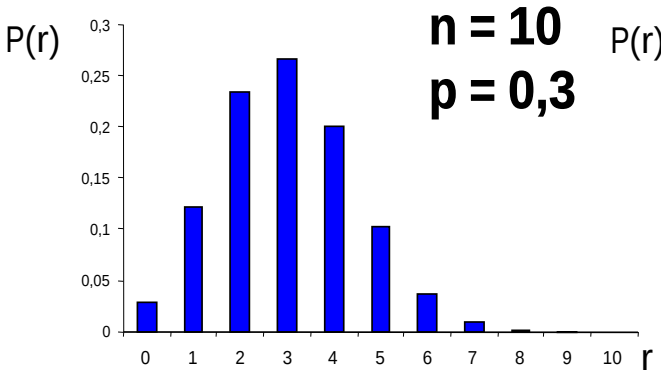
$$P(X = 4) = 0{,}0625$$

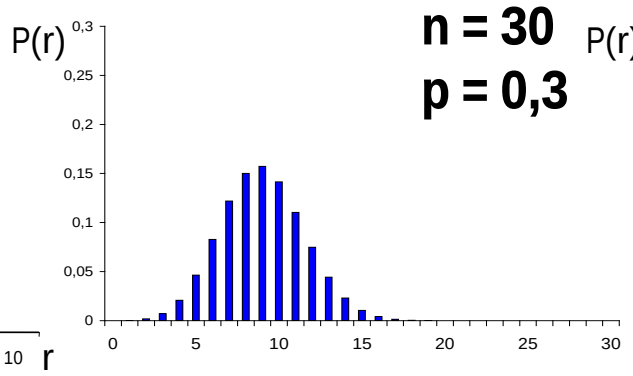


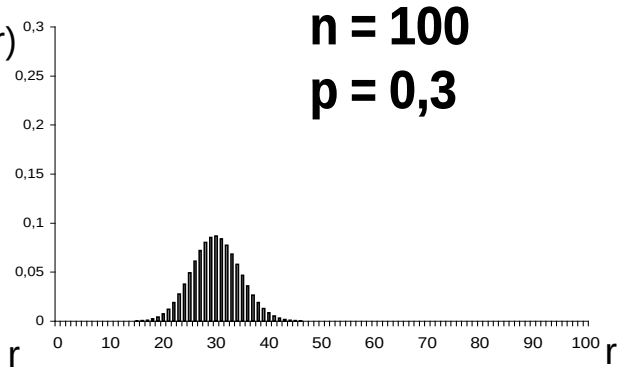
Nejpravděpodobnější počet chlapců – střední hodnota: $E(X) = n \cdot p = 4 \cdot 0{,}5 = 2$

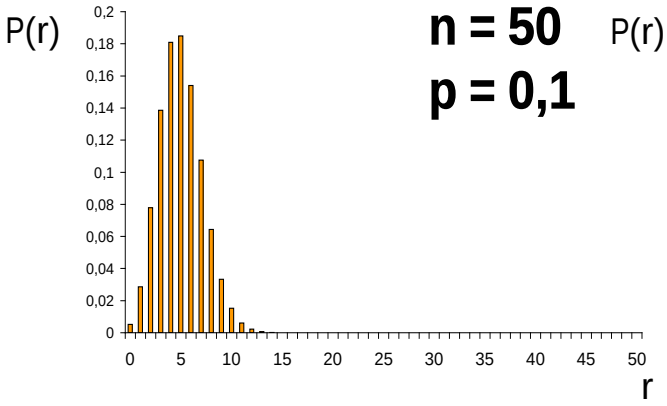
Binomické rozdělení – tvar pro různé *n* a *p*

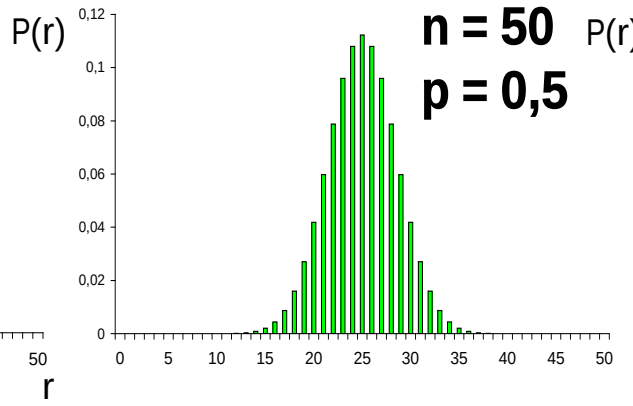
- Čím vícekrát opakujeme experiment, tím menší relativní podíl připadá na jednotlivé hodnoty X, neboť všechny dohromady musí dát součet 1 (100%).
- Rozdělení s *p*=0,5 je symetrické kolem středu osy *x*, menší či větší *p* posouvá střed rozdělení směrem k limitním hodnotám (tedy hodnotám 0 či *n*).

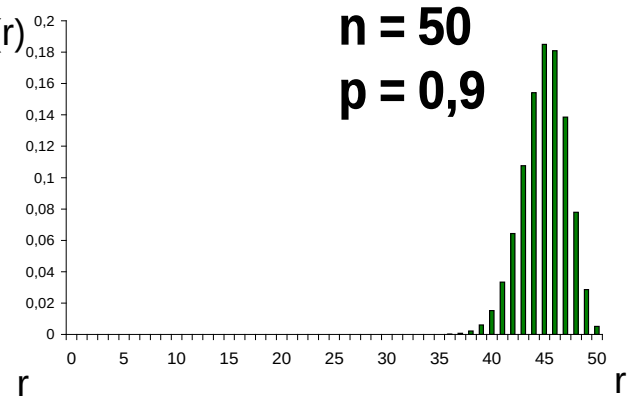






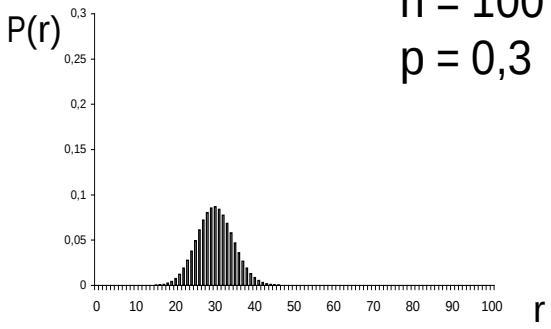






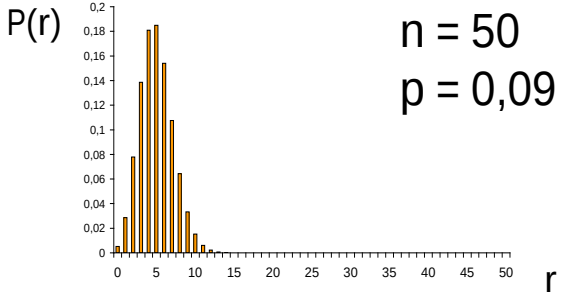
Binomické rozložení – speciální případy

- Pokud n=1, jde o tzv. alternativní rozdělení a daná událost buď nenastane nebo nastane jednou.
- Pokud náhodný experiment opakujeme mnohokrát (n je velké), rozdělení se začne podobat spojitému rozdělení → aproximace na normální rozdělení.



n = 100
p = 0,3

- Aproximace normálním rozdělením však nebude platit pro velmi nízké a velmi vysoké hodnoty *p* → u nízkých hodnot *p* aproximace na Poissonovo rozdělení (pro $n > 30$ a $p < 0{,}1$).



n = 50
p = 0,09

Binomické rozdělení - interval spolehlivosti - příklad

- **Př. Sledování výskytu nežádoucích účinků u *n* = 100 pacientů se schizofrenií léčených daným přípravkem. Nežádoucí účinky se vyskytly u 60 jedinců. Odhadněte pravděpodobnost výskytu nežádoucích účinků a tento odhad doplňte o 95% interval spolehlivosti.**

- Vzorečky:

$$p \approx p; \quad p = r/n$$ (bodový odhad parametru π)

$$p - Z_{1-\alpha/2} \times \sqrt{\frac{p(1-p)}{n-1}} \leq p \leq p + Z_{1-\alpha/2} \times \sqrt{\frac{p(1-p)}{n-1}}$$ (interval spolehlivosti pro π)

- Řešení:

$$p = 60/100 = 0{,}6$$

$$0{,}6 - 1{,}96 \times \sqrt{\frac{0{,}6 \times (1-0{,}6)}{100-1}} \leq p \leq 0{,}6 + 1{,}96 \times \sqrt{\frac{0{,}6 \times (1-0{,}6)}{100-1}}$$

$$0{,}6 - 1{,}96 \times 0{,}049 \leq p \leq 0{,}6 + 1{,}96 \times 0{,}049$$

$$0{,}503 \leq p \leq 0{,}697$$

- Pravděpodobnost výskytu nežádoucích účinků je 0,6 (0,503; 0,697).

Binomické rozdělení – interval spolehlivosti

- **Ovlivnění šířky intervalu spolehlivosti (IS):** $p \pm Z_{1-\alpha/2} \times \sqrt{\frac{p(1-p)}{n-1}}$
  - hodnotou *p* – IS bude nejširší pro *p* = 0,5
  - hodnotou *n* – IS širší při malém *n* než při velkém
  - hodnotou α – IS širší pro malé α (hladinu spolehlivosti) – tzn. 99% IS bude širší než 95% IS

- **Interval spolehlivosti bez aproximace na normální rozdělení** (pokud hodnoty *p* jsou velmi nízké nebo velmi vysoké):

Dolní hranice IS:

$$D = \frac{r}{r + (n - r + 1) \times F_{\alpha/2}^{(n_1;n_2)}}$$

... kde:

$$n_1 = 2(n - r + 1); \quad n_2 = 2r$$

Horní hranice IS:

$$H = \frac{(r+1) \times F_{\alpha/2}^{(n_1';n_2')}}{n - r + (r+1) \times F_{\alpha/2}^{(n_1';n_2')}}$$

... kde:

$$n_1' = 2(r+1) = n_2 + 2$$

$$n_2' = 2(n - r) = n_1 - 2$$

Statistické testování binomických dat

1. **Liší se odhad *p* od předpokládané (referenční) hodnoty π?**
   (Např. liší se procento pacientů s nežádoucími účinky léčby od předpokládaného procenta?)
   → **jednovýběrový binomický test** (tzn. test pro podíl u jednoho výběru)

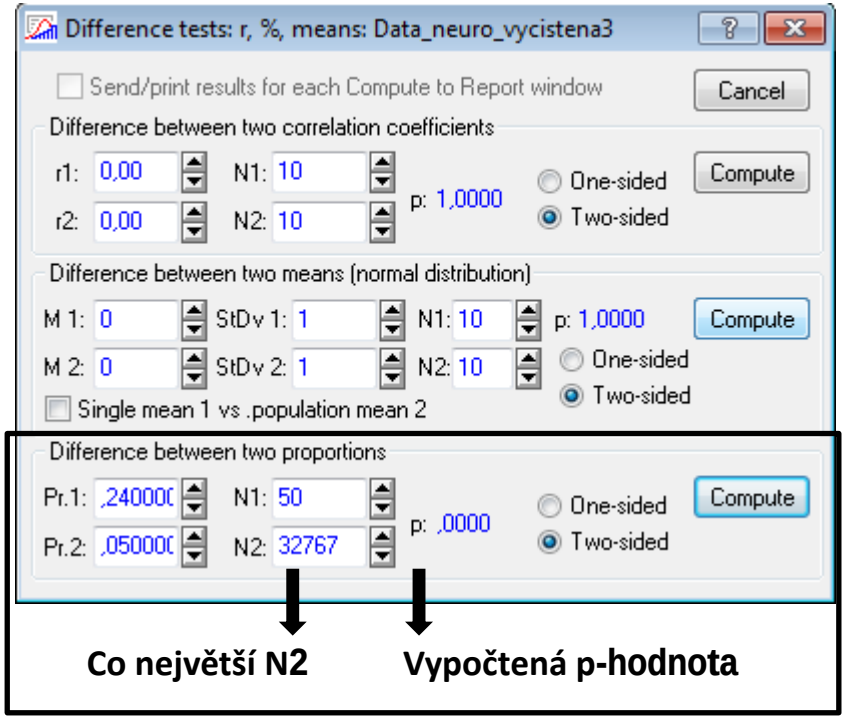
2. **Liší se *p* ve dvou souborech?**
   (Např. liší se podíl pacientů s nežádoucími účinky léčby podle typu léčby?)
   → **dvouvýběrový binomický test** (tzn. test pro podíl u dvou výběrů)

Jednovýběrový binomický test

- **Příklad:** Mezi 50 pacienty s Alzheimerovou chorobou je 12 pacientů s MMSE skóre nižším než daná hranice. Ověřte, zda podíl pacientů s nižším skóre je stejný jako v běžné populaci.
- Tzn. hypotézy budou mít tvar: $H_0 : p = p$ a $H_1 : p \neq p$

- **Řešení:**
- π = 0,05 (v populaci – hranice skóre jsou dělána tak, aby 5% populace bylo nižší než hranice)
- p = 12/50 = 0,24

- **Závěr:**
  Podíl pacientů s nižším MMSE skóre je statisticky významně odlišný od podílu v běžné populaci.

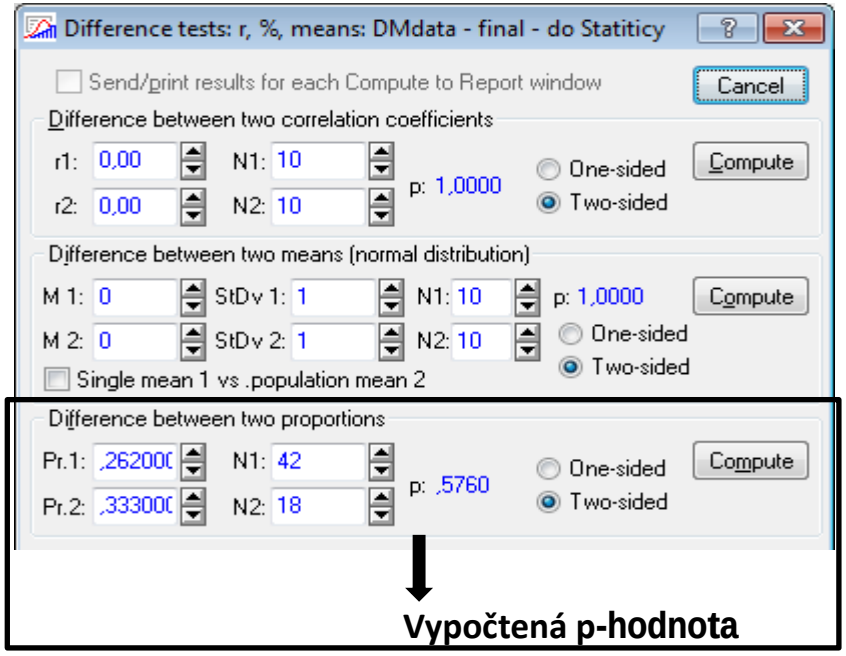


Dvouvýběrový binomický test

- **Příklad:** Mezi 42 pacienty s Alzheimerovou chorobou (AD) je 11 pacientů s MMSE skóre nižším než daná hranice. Mezi 18 pacienty s mírnou kognitivní poruchou (MCI) je 6 pacientů s MMSE skóre nižším než daná hranice. Ověřte, zda se podíly pacientů s nižším skóre u pacientů s AD a MCI liší.
- Tzn. hypotézy budou mít tvar: $H_0 : p_1 = p_2$ a $H_1 : p_1 \neq p_2$

- **Řešení:**
- $p_1 = 11/42 = 0{,}262$
- $p_2 = 6/18 = 0{,}333$

- **Závěr:**
  Neprokázali jsme, že by se podíl subjektů s nižším MMSE skóre lišil u pacientů s AD a MCI.



Poissonovo rozdělení

Poissonovo rozdělení

- Diskrétní rozdělení, které **popisuje počet výskytů sledované události na danou jednotku** (času, plochy, objemu), když se tyto události vyskytují vzájemně nezávisle s konstantní intenzitou (parametr λ).
- Značení: Po(λ)
- Jedná se o zobecnění binomického rozdělení pro $n \to \infty$ a $p \to 0$

  (aproximace je funkční již při n > 30, p < 0,1): $\mathrm{Bi}(n, p) \to \mathrm{Po}(n \times p)$

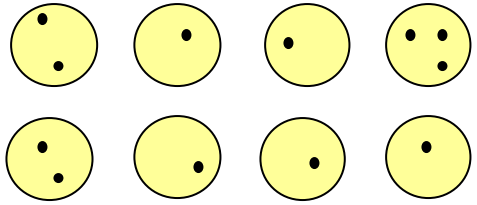
  Pravděpodobnost, že sledovaná událost nastane r-krát, lze vypočítat:

$$P(X = r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

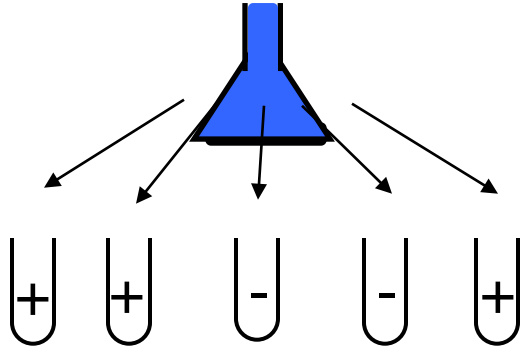
- **Střední hodnota:** EX = λ (λ vyjadřuje střední počet jevů na jednu experimentální jednotku)
- **Rozptyl:** DX = λ
- Příklady: počet krvinek v poli mikroskopu, počet pooperačních komplikací během určitého časového intervalu po výkonu, počet pacientů, kteří přišli do ordinace během jedné hodiny, počet částic, které vyzáří zářič za danou časovou jednotku

Poissonovo rozdělení – příklady

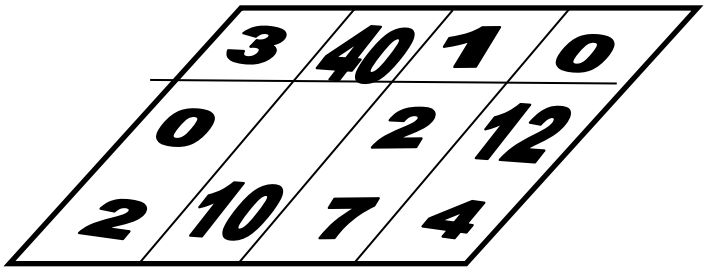
Výskyt jevu na experimentální jednotku
(mutace bakterií na inkubačních miskách)



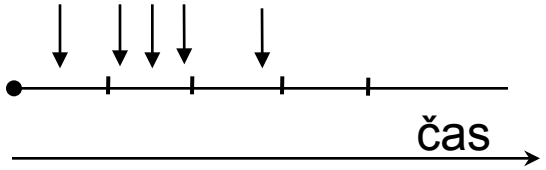
Orientační stanovení jevu
(např. produkce plynu bakteriemi)



Výskyt jevu v prostoru
(počet buněk v sčítacím poli preparátu)

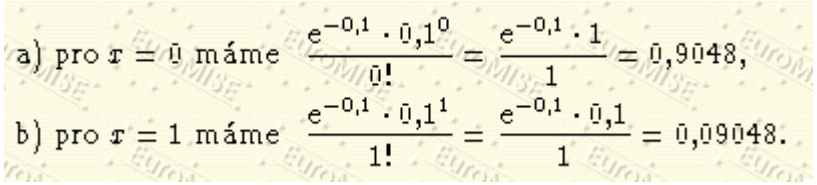


Výskyt jevu v čase
(vyzáření částice v určitých časových intervalech)



Poissonovo rozdělení – příklad

- **Příklad:** Předpokládejme, že v určité populaci krys se vyskytuje albín s pravděpodobností π=0,001, ostatní krysy jsou normálně pigmentované. Ve vzorku 100 krys náhodně vybraných z této populace určete pravděpodobnost, že vzorek a) neobsahuje albína, b) obsahuje právě jednoho albína.
- *Řešení:* Pravděpodobnost výskytu albína je π=0,001. Předpokládaný počet albínů ve výběru o rozsahu *n* je λ=n*π (průměr binomické náhodné veličiny), tj. v našem příkladu λ=n*π=100*0,001=0,1. Počet albínů označme *x*. Potom:

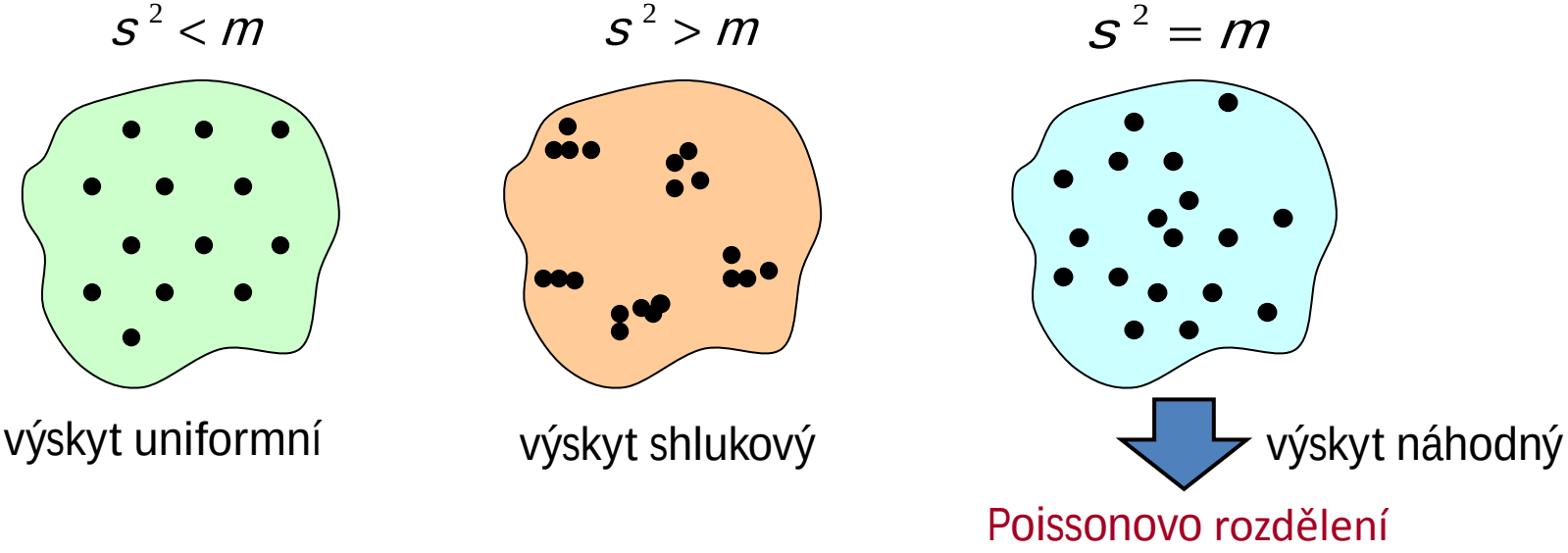


- Jak je vidět, pravděpodobnost, že ve vzorku 100 krys nebude žádný albín, je desetkrát vyšší než pravděpodobnost, že ve vzorku bude právě jeden albín. Pravděpodobnosti výskytu dvou a více albínů jsou již velmi malé.

Převzato z Zvárová, J. (2001) *Základy statistiky pro biomedicínské obory*. Praha: Karolinum.

Poissonovo rozdělení – předpoklady

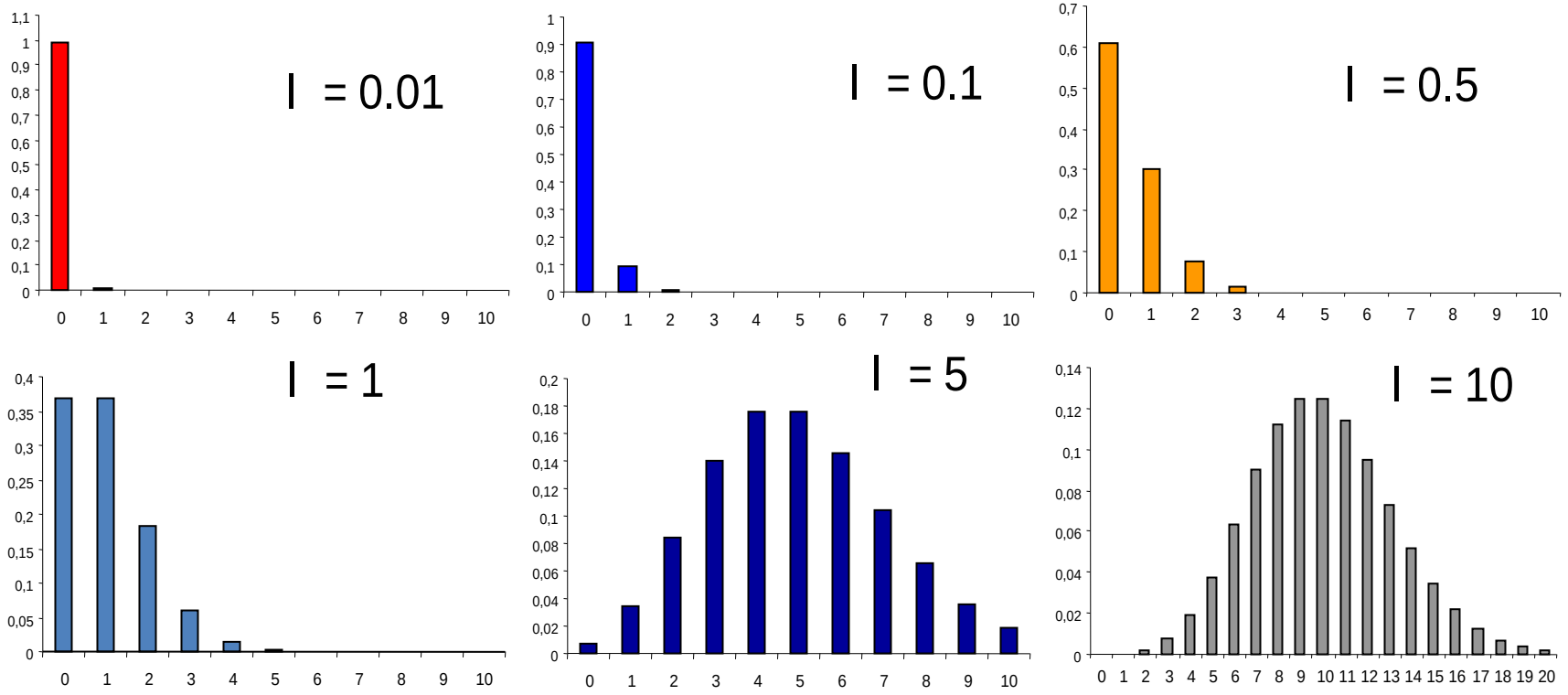
- výskyt jevu je zcela náhodný (tedy náhodný v čase nebo prostoru podle typu situace)



- výskyt jevu v konkrétní experimentální jednotce nijak nezávisí na tom, co se stalo v jiných jednotkách
- není možné, aby 2 nebo více jevů nastaly současně, přesně ve stejném místě prostoru nebo ve stejném časovém okamžiku
- pro každý dílčí časový okamžik, prostorou jednotku apod. je pravděpodobnost výskytu stejná

Poissonovo rozdělení – tvar pro různé λ

- Čím větší je λ, tím více se tvar Poissonova rozdělení blíží normálnímu rozdělení.



Poissonovo rozdělení – intervaly spolehlivosti - příklad

- **Př. Za 10 hodin vyzářil zářič 1500 částic. Spočtěte průměrný počet vyzářených částic za hodinu a tento odhad průměrného počtu částic doplňte o 95% interval spolehlivosti.**

- Vzorečky:

  $\lambda \approx \bar{x}$ (bodový odhad parametru λ)

  $$\bar{x} - Z_{1-\alpha/2} \times \sqrt{\frac{\bar{x}}{n}} \leq \lambda \leq \bar{x} + Z_{1-\alpha/2} \times \sqrt{\frac{\bar{x}}{n}}$$ (interval spolehlivosti pro λ)

- Řešení:

$$\bar{x} = 1500 / 10 = 150$$

$$150 - 1{,}96 \times \sqrt{\frac{150}{10}} \leq \lambda \leq 150 + 1{,}96 \times \sqrt{\frac{150}{10}}$$

$$150 - 1{,}96 \times 3{,}873 \leq \lambda \leq 150 + 1{,}96 \times 3{,}873$$

$$142 \leq \lambda \leq 158$$

- Průměrný počet částic vyzářených za hodinu je 150 (142;158).

Poissonovo rozdělení – interval spolehlivosti

- **Ovlivnění šířky intervalu spolehlivosti (IS):** $\bar{x} \pm Z_{1-a/2} \times \sqrt{\frac{\bar{x}}{n}}$
  - hodnotou λ – IS širší při velkém λ
  - hodnotou *n* – IS širší při malém *n* než při velkém
  - hodnotou α – IS širší pro malé α (hladinu spolehlivosti) – tzn. 99% IS bude širší než 95% IS

- **Interval spolehlivosti bez aproximace na normální rozdělení**:

Dolní hranice IS:

$$D = \frac{c^2_{a/2}(n_1)}{2}$$

... kde:

$n_1 = 2r$

Horní hranice IS:

$$H = \frac{c^2_{1-a/2}(n_2)}{2}$$

... kde:

$n_2 = n_1 + 2 = 2r + 2$