Jednovýběrový Wilcoxonův test a jeho asymptotická varianta
(neparametrická obdoba jednovýběrového t-testu)

Frank Wilcoxon (1892 – 1965): Americký statistik a chemik

Nechť $X_1, ..., X_n$ je náhodný výběr ze spojitého rozložení s hustotou $\varphi(x)$, která je symetrická kolem mediánu $x_{0,50}$, tj.
$\varphi(x_{0,50} + x) = \varphi(x_{0,50} - x)$. Nechť c je reálná konstanta.
Testujeme hypotézu H_0: $x_{0,50} = c$
proti oboustranné alternativě H_1: $x_{0,50} \neq c$ nebo
proti levostranné alternativě H_1: $x_{0,50} < c$ nebo
proti pravostranné alternativě H_1: $x_{0,50} > c$.

Postup provedení testu:

a) Utvoříme rozdíly $D_i = X_i - c$, $i = 1, ..., n$. (Jsou-li některé rozdíly nulové, pak za n bereme jen počet nenulových hodnot.)

b) Absolutní hodnoty $|D_i|$ uspořádáme vzestupně podle velikosti a spočteme pořadí R_i.

c) Zavedeme statistiky

$S_W^+ = \sum_{D_i>0} R_i^+$, což je součet pořadí přes kladné hodnoty D_i,

$S_W^- = \sum_{D_i<0} R_i^-$, což je součet pořadí přes záporné hodnoty D_i.

Přitom platí, že součet $S_W^+ + S_W^- = n(n+1)/2$.

Je-li H_0 pravdivá, pak $E(S_W^+) = n(n+1)/4$ a $D(S_W^+) = n(n+1)(2n+1)/24$.

d) Testová statistika = $\min(S_W^+, S_W^-)$ pro oboustrannou alternativu,

= S_W^+ pro levostrannou alternativu,

= S_W^- pro pravostrannou alternativu.

e) H_0 zamítáme na hladině významnosti α, když testová statistika je menší nebo rovna tabelované kritické hodnotě.

Asymptotická varianta jednovýběrového Wilcoxonova testu:

Pro $n \geq 30$ lze využít asymptotické normality statistiky S_W^+.

Platí-li H_0, pak $U_0 = \dfrac{S_W^+ - E\left(S_W^+\right)}{\sqrt{D\left(S_W^+\right)}} = \dfrac{S_W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \approx N(0,1)$.

Kritický obor:

pro oboustrannou alternativu $W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$,

pro levostrannou alternativu $W = \left(-\infty, -u_{1-\alpha}\right\rangle$,

pro pravostrannou alternativu $W = \left\langle u_{1-\alpha}, \infty\right)$

H_0 zamítáme na asymptotické hladině významnosti α, když $U_0 \in W$.

Příklad: U 12 náhodně vybraných zemí bylo zjištěno procento populace starší 60 let:

4,9 6,0 6,9 17,6 4,5 12,3 5,7 5,3 9,6 13,5 15,7 7,7.

Na hladině významnosti 0,05 testujte hypotézu, že medián procenta populace starší 60 let je 12 proti oboustranné alternativě.

Řešení:

Testujeme hypotézu H_0: $x_{0,50} = 12$ proti oboustranné alternativě H_1: $x_{0,50} \neq 12$.

Vypočteme rozdíly pozorovaných hodnot od čísla 12: -7,1 -6,0 -5,1 5,6 -7,5 0,3 -6,3 -6,7 -2,4 1,5 3,7 -4,3.

Absolutní hodnoty těchto rozdílů uspořádáme vzestupně podle velikosti. Kladné rozdíly přitom označíme červeně:

usp. $\lvert x_i - 12 \rvert$	0,3	1,5	2,4	3,7	4,3	5,1	5,6	6	6,3	6,7	7,1	7,5
pořadí	1	2	3	4	5	6	7	8	9	10	11	12

$S_W^+ = 1 + 2 + 4 + 7 = 14$,

$S_W^- = 3 + 5 + 6 + 8 + 9 + 10 + 11 + 12 = 64$,

$n = 12$, $\alpha = 0{,}05$, tabelovaná kritická hodnota pro $n = 12$ a $\alpha = 0{,}05$ je 13,

testová statistika = $\min(S_W^+, S_W^-) = \min(14{,}64) = 14$.

Protože 14 > 13, H_0 nezamítáme na hladině významnosti 0,05. Znamená to, že na hladině významnosti 0,05 se nepodařilo prokázat, že aspoň v polovině zemí by se podíl populace nad 60 let odlišoval od 12 %.

Výpočet pomocí systému STATISTICA:

Utvoříme nový datový soubor se dvěma proměnnými a 12 případy. Do proměnné procento napíšeme zjištěné hodnoty a do proměnné konst uložíme číslo 12.

Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných rozdil, Druhý seznam proměnných konst – OK – Wilcoxonův párový test.

	Wilcoxonův párový test (populace_nad_60) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet platných	T	Z	Úroveň p
procento & konst	12	14,00000	1,961161	0,049861

Výstupní tabulka poskytne hodnotu testové statistiky SW^+ (zde označena T), hodnotu asymptotické testové statistiky U_0 a p-hodnotu pro U_0. V tomto případě je p-hodnota 0,049861, tedy nulová hypotéza se zamítá na asymptotické hladině významnosti 0,05. Tento výsledek je v rozporu s výsledkem, ke kterému jsme dospěli při přesném výpočtu. Je to způsobeno tím, že není splněna podmínka pro využití asymptotické normality statistiky SW^+, tj. $n \geq 30$.

Párový Wilcoxonův test
(nerametrická obdoba párového t-testu)

Nechť $(X_1, Y_1), \ldots, (X_n, Y_n)$ je náhodný výběr ze spojitého dvourozměrného rozložení.

Testujeme H_0: $x_{0,50} - y_{0,50} = c$ proti H_1: $x_{0,50} - y_{0,50} \neq c$ (resp. proti jednostranným alternativám).

Utvoříme rozdíly $Z_i = X_i – Y_i$, $i = 1, \ldots, n$ a testujeme hypotézu o mediánu $z_{0,50}$, tj. H_0: $z_{0,50} = c$ proti H_1: $z_{0,50} \neq c$.

Příklad: K zjištění cenových rozdílů mezi určitými dvěma druhy zboží bylo náhodně vybráno 15 prodejen a byly zjištěny ceny zboží A a ceny zboží B: (11,10), (14,11), (11,9), (13,9), (11,9), (10,9), (12,10), (10,8), (12,11), (11,9), (13,10), (14,10), (14,12), (19,15), (14,12). Na hladině významnosti 0,05 je třeba testovat hypotézu, že medián cenových rozdílů činí 3 Kč.

Řešení:Testujeme H_0: $z_{0,50} = 3$ proti oboustranné alternativě H_1: $z_{0,50} \neq 3$, kde $z_{0,50}$ je medián rozložení, z něhož pochází rozdílový náhodný výběr $Z_1 = X_1 – Y_1$, … $Z_{15} = X_{15} – Y_{15}$.Vypočteme rozdíly mezi cenou zboží A a cenou zboží B, čímž úlohu převedeme na jednovýběrový test. Výpočty uspořádáme do tabulky:

č. prodejny	cena zboží A	cena zboží B	rozdíl	|rozdíl-medián|	pořadí
1	11	10	1	2	12
2	14	11	3	0	-
3	11	9	2	1	5,5
4	13	9	4	1	**5,5**
5	11	9	2	1	5,5
6	10	9	1	2	12
7	12	10	2	1	5,5
8	10	8	2	1	5,5
9	12	11	1	2	12
10	11	9	2	1	5,5
11	13	10	3	0	-
12	14	10	4	1	**5,5**
13	14	12	2	1	5,5
14	19	15	4	1	**5,5**
15	14	12	2	1	5,5

(Tučně jsou vytištěna pořadí pro kladné hodnoty rozdíl - medián.)
$S_W^+ = 5,5 + 5,5 + 5,5 = 16,5$,
$S_W^- = 12 + 5,5 + 5,5 + 12 + 5,5 + 5,5 + 12 + 5,5 + 5,5 + 5,5 = 74,5$,
$n = 13$, $\alpha = 0,05$, tabelovaná kritická hodnota = 17, testová statistika = $\min(S_W^+, S_W^-) = \min(16,5; 74,5) = 16,5$. Protože $16,5 \leq 17$, H_0 zamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:
Vytvoříme nový datový soubor se čtyřmi proměnnými A, B, rozdíl, konst a 15 případy. Do proměnných A, B napíšeme ceny zboží A a B, do proměnné rozdíl uložíme rozdíl cen A a B a do proměnné konst uložíme číslo 3.
Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných rozdil, 2. seznam proměnných konst – OK – Wilcoxonův párový test.

	Wilcoxonův párový test (ceny zbozi) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet platných	T	Z	Úroveň p
rozdil & konst	15	16,50000	2,026684	0,042696

Testová statistika (zde označená jako T) nabývá hodnoty 16,5, asymptotická testová statistika (označená jako Z) nabývá hodnoty 2,026684, odpovídající asymptotická p-hodnota je 0,042696, tedy na asymptotické hladině významnosti 0,05 nulovou hypotézu zamítáme.

Příklad (na asymptotickou variantu Wilcoxonova testu):
30 náhodně vybraných osob mělo nezávisle na sobě bez předchozího nácviku odhadnout, kdy od daného signálu uplyne právě 1 minuta. Byly získány následující výsledky (v sekundách):
53 48 45 55 63 51 66 56 50 58 61 51 64 63 59 47 46 58 52 56 61 57 48 62 54 49 51 46 53 58.
Na asymptotické hladině významnosti 0,05 testujte hypotézu, že medián rozložení, z něhož daný náhodný výběr pochází, je 60 sekund proti oboustranné alternativě (nulová hypotéza vlastně tvrdí, že polovina osob délku jedné minuty podhodnotí a druhá nadhodnotí).

Řešení:
Testujeme H_0: $x_{0,50} = 60$ proti oboustranné alternativě H_1: $x_{0,50} \neq 60$.
Obvyklým způsobem stanovíme statistiku $S_W^+ = 55$.
Asymptotická testová statistika:

$$U_0 = \frac{S_W^+ - E\left(S_W^+\right)}{\sqrt{D\left(S_W^+\right)}} = \frac{S_W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{55 - \frac{30(30+1)}{4}}{\sqrt{\frac{30(30+1)(2.30+1)}{24}}} = -3{,}65$$

Kritický obor:

$$W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right) = \left(-\infty, -u_{0,975}\right\rangle \cup \left\langle u_{0,975}, \infty\right) = \left(-\infty, -1{,}96\right\rangle \cup \left\langle 1{,}96, \infty\right).$$

Testová statistika se realizuje v kritickém oboru, tedy H_0 zamítáme na asymptotické hladině významnosti 0,05. S rizikem omylu nejvýše 5% jsme tedy prokázali, že pravděpodobnost nadhodnocení jedné minuty není stejná jako pravděpodobnost podhodnocení.

Výpočet pomocí systému STATISTICA:

Utvoříme nový datový soubor se dvěma proměnnými a 30 případy. Do proměnné odhad napíšeme zjištěné hodnoty a do proměnné konst uložíme číslo 60.

Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných odhad, 2. seznam proměnných konst – OK – Wilcoxonův párový test.

	Wilcoxonův párový test (odhad minuty) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet platných	T	Z	Úroveň p
odhad & konst	30	55,00000	3,650880	0,000261

Testová statistika (zde označená jako T) nabývá hodnoty 55, asymptotická testová statistika (označená jako Z) nabývá hodnoty 3,65088, odpovídající asymptotická p-hodnota je 0,000261, tedy na asymptotické hladině významnosti 0,05 nulovou hypotézu zamítáme.

Dvouvýběrový Wilcoxonův test a jeho asymptotická varianta (neparametrická obdoba dvouvýběrového t-testu)

Nechť $X_1, ..., X_n$ a $Y_1, ..., Y_m$ jsou dva nezávislé náhodné výběry ze dvou spojitých rozložení, jejichž distribuční funkce se mohou lišit pouze posunutím. Označme $x_{0,50}$ medián prvního rozložení a $y_{0,50}$ medián druhého rozložení. Na hladině významnosti 0,05 testujeme hypotézu, že distribuční funkce těchto rozložení jsou shodné neboli mediány jsou shodné proti alternativě, že jsou rozdílné, tj.
H_0: $x_{0,50} - y_{0,50} = 0$ proti H_1: $x_{0,50} - y_{0,50} \neq 0$.

Postup provedení testu:

a) Všech n + m hodnot $X_1, ..., X_n$ a $Y_1, ..., Y_m$ uspořádáme vzestupně podle velikosti.

b) Zjistíme součet pořadí hodnot $X_1, ..., X_n$ a označíme ho T_1.
Součet pořadí hodnot $Y_1, ..., Y_m$ označíme T_2.

c) Vypočteme statistiky $U_1 = mn + n(n+1)/2 - T_1$, $U_2 = mn + m(m+1)/2 - T_2$.
Přitom platí $U_1 + U_2 = mn$.

d) Pokud $\min(U_1,U_2) \leq$ tabelovaná kritická hodnota (pro dané rozsahy výběrů m, n a dané α), pak nulovou hypotézu o totožnosti obou distribučních funkcí zamítáme na hladině významnosti α.
V tabulkách: $n = \min\{m,n\}$ a $m = \max\{m,n\}$.

Asymptotická varianta dvouvýběrového Wilcoxonova testu:

Pro velká n, m (n, m > 30) lze využít asymptotické normality statistiky U_1.

Platí-li H_0, pak $U_0 = \dfrac{U_1 - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \approx N(0,1)$, kde $U_1 = \min(U_1, U_2)$.

Kritický obor:

pro oboustrannou alternativu $W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$,

pro levostrannou alternativu $W = \left(-\infty, -u_{1-\alpha}\right\rangle$,

pro pravostrannou alternativu $W = \left\langle u_{1-\alpha}, \infty\right)$

H_0 zamítáme na asymptotické hladině významnosti α, když $U_0 \in W$.

Předpoklady použití dvouvýběrového Wilcoxonova testu:

- dané dva náhodné výběry jsou nezávislé
- rozložení, z nichž dané dva náhodné výběry pocházejí, jsou spojitá
- distribuční funkce těchto rozložení se mohou lišit pouze posunutím
- sledovaná veličina má aspoň ordinální charakter

Příklad:

Bylo vybráno 10 polí stejné kvality. Na čtyřech z nich se zkoušel nový způsob hnojení, zbylých šest bylo ošetřeno starým způsobem. Pole byla oseta pšenicí a sledoval se její hektarový výnos. Je třeba zjistit, zda nový způsob hnojení má týž vliv na průměrné hektarové výnosy pšenice jako starý způsob hnojení.

hektarové výnosy při novém způsobu: 51 52 49 55

hektarové výnosy při starém způsobu: 45 54 48 44 53 50

Test proveďte na hladině významnosti 0,05.

Řešení:

Na hladině významnosti 0,05 testujeme H_0: $x_{0,50} - y_{0,50} = 0$ proti oboustranné alternativě H_1: $x_{0,50} - y_{0,50} \neq 0$.

usp. hodnoty	44	45	48	**49**	50	**51**	**52**	53	54	**55**
pořadí x-ových hodnot				4		6	7			10
pořadí y-ových hodnot	1	2	3		5			8	9	

$T_1 = 4 + 6 + 7 + 10 = 27$, $T_2 = 1 + 2 + 3 + 5 + 8 + 9 = 28$

$U_1 = 4.6 + 4.5/2 - 27 = 7$, $U_2 = 4.6 + 6.7/2 - 28 = 17$

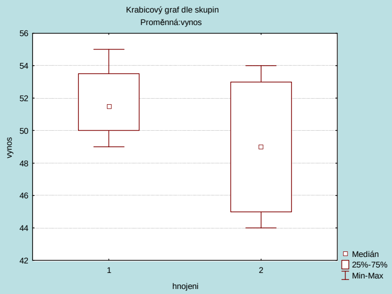
Kritická hodnota pro $\alpha = 0,05$, $\min(4,6) = 4$, $\max(4,6) = 6$ je 2. Protože $\min(7,17) = 7 > 2$, nemůžeme na hladině významnosti 0,05 zamítnout hypotézu, že nový způsob hnojení má na hektarové výnosy pšenice stejný vliv jako starý způsob.

Výpočet pomocí systému STATISTICA:

Utvoříme nový datový soubor se dvěma proměnnými a 10 případy. Do proměnné vynos napíšeme zjištěné hodnoty a do proměnné hnojeni napíšeme 4x číslo 1 pro nový způsob hnojení a 6x číslo 2 pro starý způsob hnojení.
Statistiky – Neparametrická statistika – Porovnání dvou nezávislých vzorků – OK – Proměnné – Seznam závislých proměnných vynos, Nezáv. (grupov.) proměnná hnojeni – OK – M-W U test.
Upozornění: Ve STATISTICE je dvouvýběrový Wilcoxonův test uveden pod názvem Mannův – Whitneyův test.

	Mann-Whitneyův U test (vynos) Dle proměn. hnojeni Označené testy jsou významné na hladině p <,05000									
Proměnná	Sčt poř. skup. 1	Sčt poř. skup. 2	U	Z	Úroveň p	Z upravené	Úroveň p	N platn. skup. 1	N platn. skup. 2	2*1str. přesné p
vynos	27,00000	28,00000	7,000000	1,066004	0,286423	1,066004	0,286423	4	6	0,352381

Ve výstupní tabulce jsou součty pořadí T_1, T_2, hodnota testové statistiky min(U_1, U_2) označená U, hodnota asymptotické testové statistiky U_0 (označená Z), asymptotická p-hodnota pro U_0 a přesná p-hodnota (ozn. 2*1str. přesné p – ta se používá pro rozsahy výběrů pod 30). V našem případě přesná p-hodnota = 0,352381, tedy H_0 nezamítáme na hladině významnosti 0,05.
Výpočet je vhodné doplnit krabicovým diagramem.



Je zřejmé, že výnosy při starém způsobu hnojení jsou vesměs nižší než při novém způsobu a také vykazují mnohem větší variabilitu.