Drobná překvapení – velká poučení

Ondřej Vencálek
Univerzita Palackého v Olomouci

online Miku(k)láš 2020, 7. prosince 2020

Motto:

Nejčastějšími statistickými úlohami, které řeší „venkovský statistický obvoďák“, jsou problémy s dvěma výběry ...

J. Tvrdík
článek v Informačním bulletinu ČStS 3/2012

Příklad z praxe

Problém:

- Dvě skupiny (děti versus mladiství)
- Sleduje se čas strávený u monitoru (za den) ... `st` (screen time)
- Zajímá nás srovnání skupin

Data:

```
> head(x)
  group       st
1     2 2.071429
2     1 2.285714
3     2 7.142857
4     2 1.571429
5     2 4.428571
6     2 2.785714
> summary(as.factor(x\$group))
  1   2
355 324
```

Příklad z praxe

Problém:

- Dvě skupiny (děti versus mladiství)
- Sleduje se čas strávený u monitoru (za den) ... `st` (screen time)
- Zajímá nás srovnání skupin

Data:

```
> head(x)
  group       st
1     2 2.071429
2     1 2.285714
3     2 7.142857
4     2 1.571429
5     2 4.428571
6     2 2.785714
> summary(as.factor(x\$group))
  1   2
355 324
```

t-test, Mann-Whitney nebo něco jiného?

Povědomí o Mannově-Whitneyho testu (hrubě):

- srovnávám dvě skupiny
- chtěl jsem použít t-test
- ale rozdělení není normální
- tak musím použít Mann-Whitney (dvouvýběrového Wilcoxona)

Povědomí o Mannově-Whitneyho testu (uhlazeněji):

- „Mannův-Whitneyho test je neparametrickou alternativou t-testu pro dva výběry ve chvíli, kdy není splněn některý z jeho předpokladů, respektive máme-li o platnosti některého z jeho předpokladů pochyby.“

  z webu `https://portal.matematickabiologie.cz/`

Překvapení

Na hladině významnosti 0,05

- t-test ukazuje **nesignifikantní** rozdíl (p=**0,206**)
- Wilcoxonův test ukazuje **signifikantní** rozdíl (p=**0,008**)

```
Two Sample t-test

data:  x1 and x2
t = 1.2651, df = 677, p-value = 0.2063
alternative hypothesis: true difference in means is not equal to 0
```

```
Wilcoxon rank sum test with continuity correction

data:  x1 and x2
W = 64262, p-value = 0.008159
alternative hypothesis: true location shift is not equal to 0
```

Překvapení

Na hladině významnosti 0,05

- t-test ukazuje **nesignifikantní** rozdíl (p=**0,206**)
- Wilcoxonův test ukazuje **signifikantní** rozdíl (p=**0,008**)
- ```
  Two Sample t-test

  data:  x1 and x2
  t = 1.2651, df = 677, p-value = 0.2063
  alternative hypothesis: true difference in means is not equal to 0
  ```
- ```
  Wilcoxon rank sum test with continuity correction

  data:  x1 and x2
  W = 64262, p-value = 0.008159
  alternative hypothesis: true location shift is not equal to 0
  ```

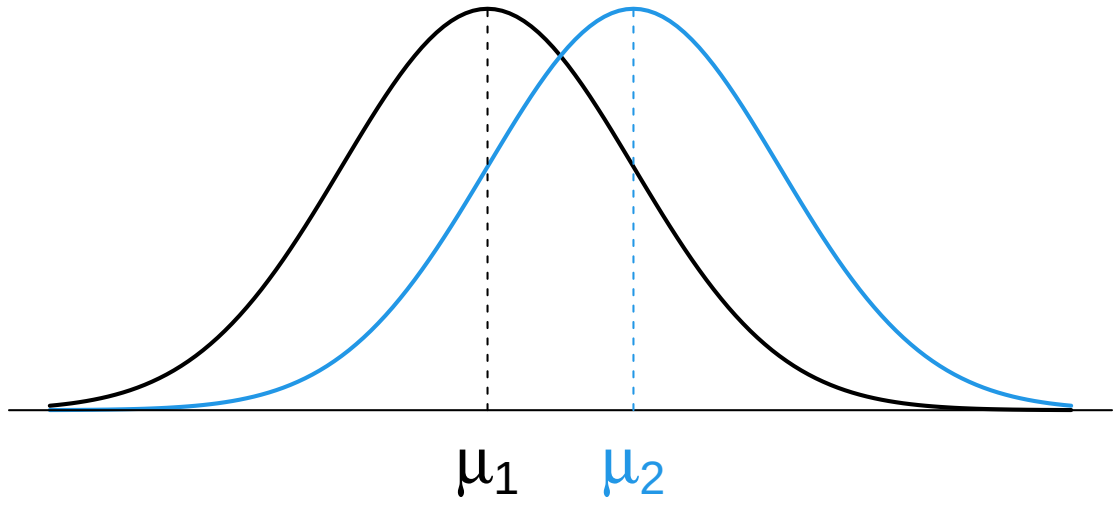
Proč jsem byl překvapený I

- Použití Wilcoxonova testu je zřejmě v pořádku (i když nejsou rozdělení normální).
- Použití t-testu je vzhledem k velkým rozsahům výběru taky v pořádku (díky centrální limitní větě).

... jenže testy „dávají jiný výsledek“...

Proč jsem byl překvapený II

... a tady mé myšlenky zabočily na nesprávnou kolej



Jestliže rozdíl existuje, spíše ho odhalí silnější test (t-test).
Tedy čekal bych p-hodnotu t-testu nižší než u Wilcoxonova testu.
K síle testu se ještě vrátím v závěru.

Formální předpoklady obou testů

Nechť

- $Y_{1,1}, \ldots, Y_{1,n_1}$ je náhodný výběr z rozdělení s distr. funkcí F_1,
- $Y_{2,1}, \ldots, Y_{2,n_2}$ je náhodný výběr z rozdělení s distr. funkcí F_2,
- výběry jsou vzájemně nezávislé.

Předpoklady o tvaru rozdělení

- t-test
  - F_1 je distr. funkce rozdělení $\mathrm{N}(\mu_1, \sigma^2)$
  - F_2 je distr. funkce rozdělení $\mathrm{N}(\mu_2, \sigma^2)$
- Wilcoxonův test
  - F_1 je distr. funkce spojitého rozdělení
  - F_2 je distr. funkce spojitého rozdělení

Nulová hypotéza:

$$F_1 \equiv F_2$$

t-test testuje shodu středních hodnot ...

- Za předpokladu normality a shody rozptylů je shoda rozdělení totéž co shoda středních hodnot, tj.

$$F_1 \equiv F_2 \Leftrightarrow \mu_1 = \mu_2$$

- Obecně* ale platí pouze

$$F_1 \equiv F_2 \Rightarrow \mu_1 = \mu_2$$

* pokud stř. hodnoty existují

Poznámka: U Welchova t-testu, kde předpokládáme normalitu, ale nepředpokládáme shodu rozptylů, testujeme $H_0 : \mu_1 = \mu_2$. Tím ale netestujeme shodu rozdělení!

t-test testuje shodu středních hodnot ...

- Za předpokladu normality a shody rozptylů je shoda rozdělení totéž co shoda středních hodnot, tj.

$$F_1 \equiv F_2 \Leftrightarrow \mu_1 = \mu_2$$

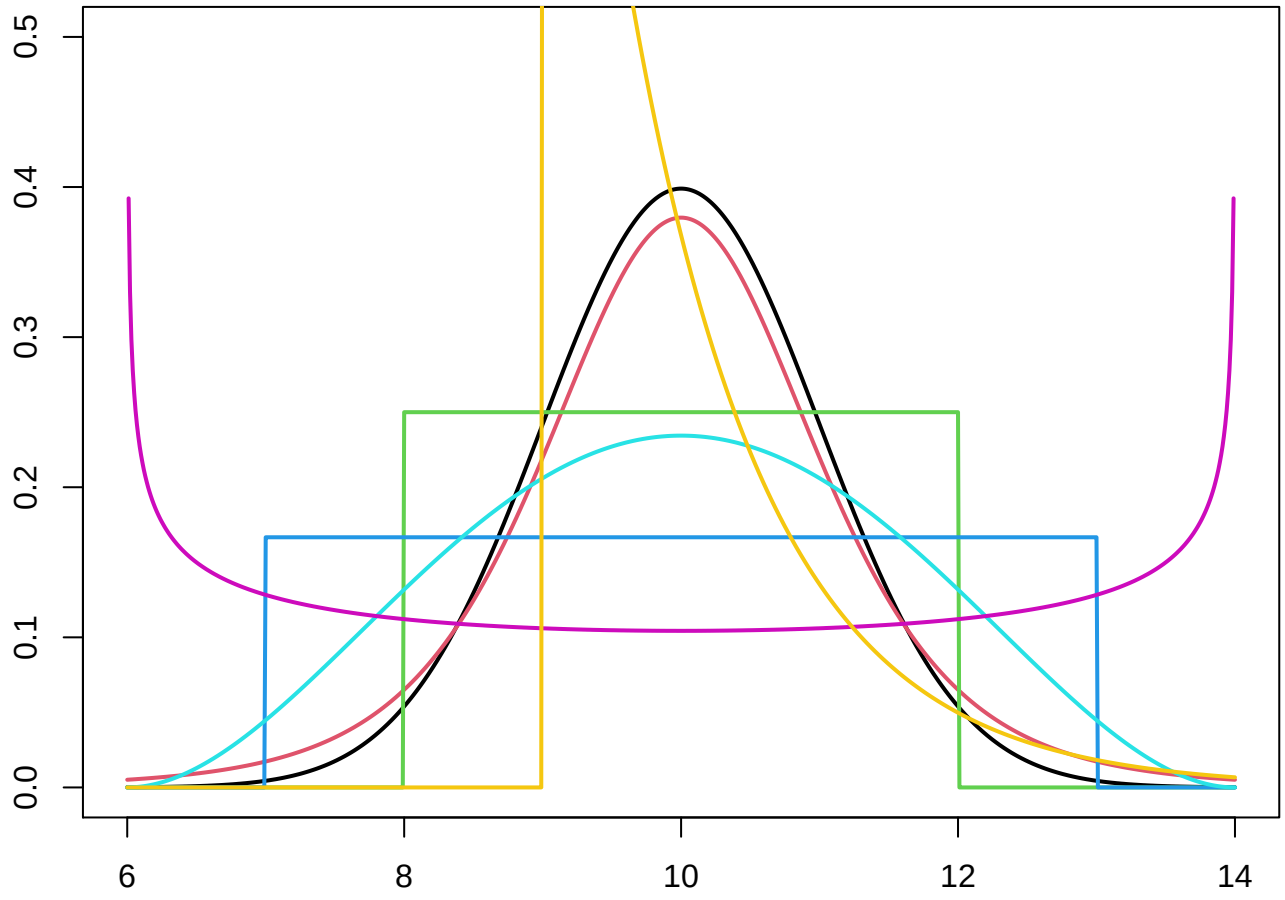
- Obecně* ale platí pouze

$$F_1 \equiv F_2 \Rightarrow \mu_1 = \mu_2$$

* pokud stř. hodnoty existují

Poznámka: U Welchova t-testu, kde předpokládáme normalitu, ale nepředpokládáme shodu rozptylů, testujeme $H_0 : \mu_1 = \mu_2$. Tím ale netestujeme shodu rozdělení!

Příklady různých rozdělení se stejnou střední hodnotou



Pro libovolnou dvojici

- platí $\mu_1 = \mu_2 (= 10)$,
- neplatí $F_1 \equiv F_2$.

Ponaučení I

Wilcoxonův (Mannův-Whitneyho) test není jen „náhražkou“ t-testu, když není splněn předpoklad normality. Testuje totiž jinou hypotézu; hypotéza vyžaduje nejen shodu středních hodnot, ale shodu rozdělení (potažmo všech existujících momentů).

- Je nešťastné chápat Wilcoxonův test jako „náhražku“ t-testu (jeho „alternativu“). Je třeba si uvědomit, že testuje jinou hypotézu.
- Pokud byste místo testování hypotéz raději pracovali Bayesovsky, budete se setkávat s podobným problémem. Zajímá mě rozdělení parametru polohy nebo mě zajímá shoda rozdělení?

Ponaučení I

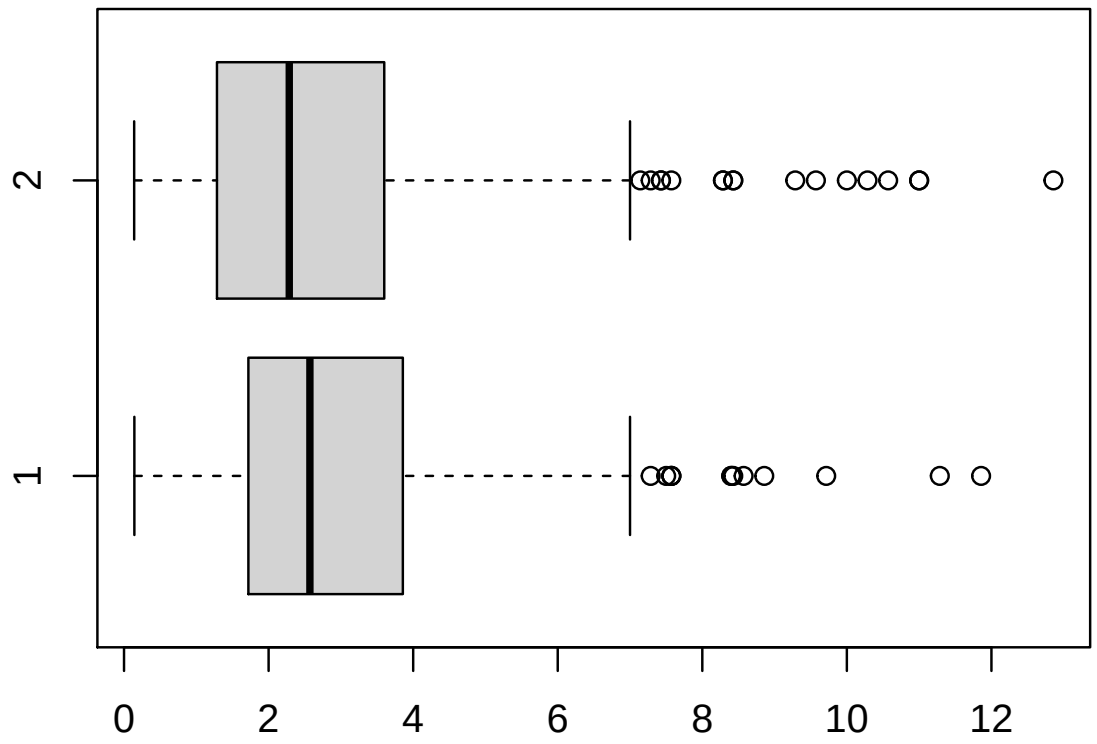
Wilcoxonův (Mannův-Whitneyho) test není jen „náhražkou“ t-testu, když není splněn předpoklad normality. Testuje totiž jinou hypotézu; hypotéza vyžaduje nejen shodu středních hodnot, ale shodu rozdělení (potažmo všech existujících momentů).

- Je nešťastné chápat Wilcoxonův test jako „náhražku“ t-testu (jeho „alternativu“). Je třeba si uvědomit, že testuje jinou hypotézu.
- Pokud byste místo testování hypotéz raději pracovali Bayesovsky, budete se setkávat s podobným problémem. Zajímá mě rozdělení parametru polohy nebo mě zajímá shoda rozdělení?

Zpět k datům anebo co mi (ne)pomohlo

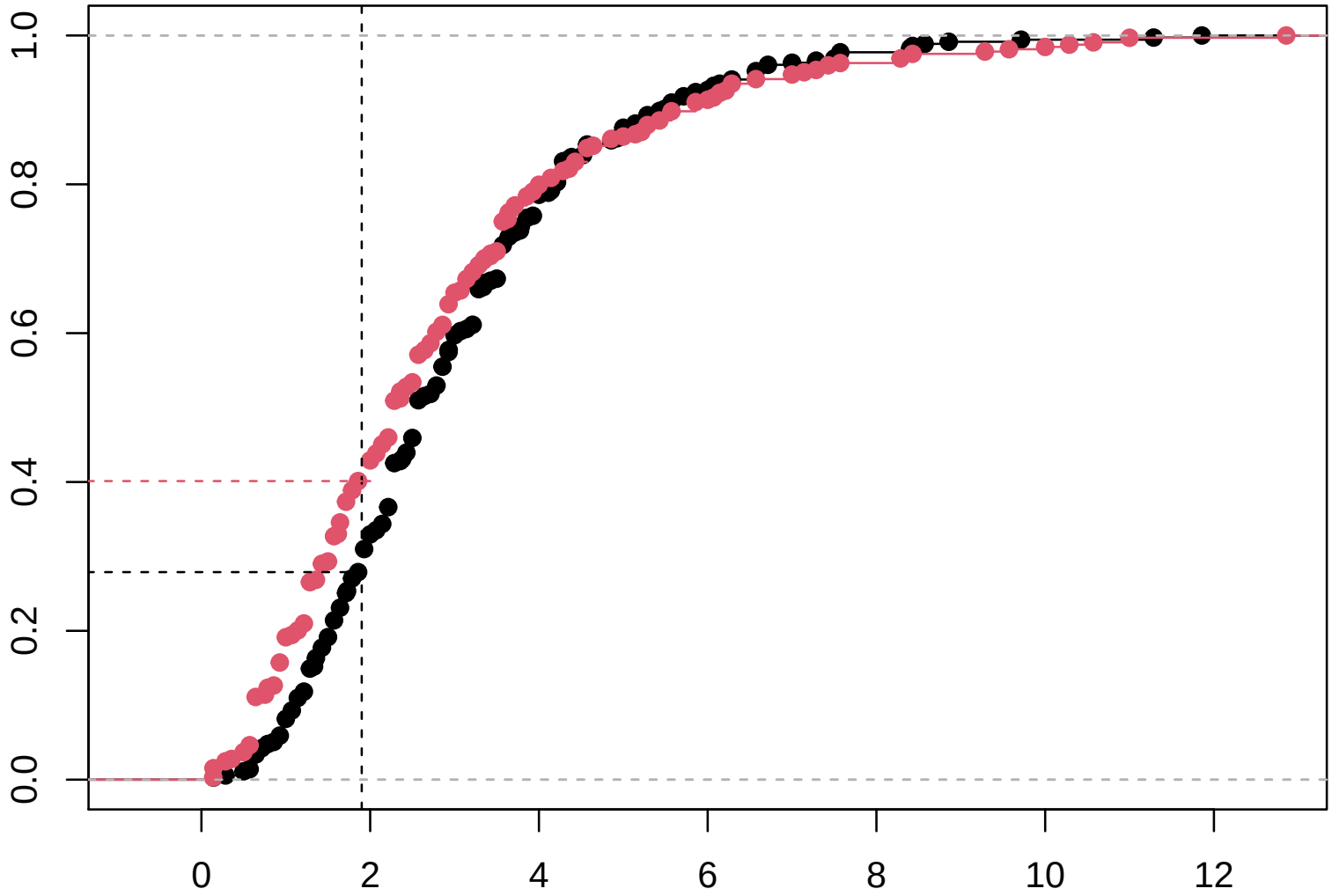
```
> tapply(x$st,x$group,summary)
$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1429  1.7202  2.5714  3.0422  3.8571 11.8571

$`2`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1402  1.2857  2.2857  2.8491  3.5857 12.8571
```



Zpět k datům anebo co mi pomohlo

graf: srovnání empirických distribučních funkcí



černě: děti (8-13 let), červeně: mladiství (14-18 let)

Ponaučení II

Vizualizace, vizualizace, vizualizace!

Praktik naléhá: Tak jaký test mám použít?

Nejprv mi řekni, co chceš vědět. Jestli si nejsi jistý, zkus si představit, k jakým závěrům můžeme dojít.

- Wilcoxon nezamítá: „rozdělení jsou shodná“, měl by nezamítat taky t-test (stř. hodnoty jsou stejné).*
- t-test nezamítá: „střední hodnoty jsou stejné“, ale o shodě rozdělení to nic neříká (tohle je naše situace).

*Je-li rozdíl mezi rozděleními malý a/nebo mám málo pozorování, může se stát, že Wilcoxon nezamítne nulovou hypotézu, ale t-test ano.

Odpověď praktikovi

- Ptáš-li se (jednoduše) „*Tráví mladiství u monitorů v průměru stejně času jako děti?*“, pak odpověď je (zjednodušeně) ANO.
- Ptáš-li se (jednoduše) „*Tráví mladiství u monitorů stejně času jako děti?*“, pak odpověď je (zjednodušeně) NE.

Jak to: i když se „průměrná“ doba strávená u monitorů u mladistvých významně neliší oproti dětem, je rozdělení ve skupině mladistvých statisticky významně jiné než ve skupině dětí. Největší rozdílnost pozorujeme v podílu těch, kteří u monitorů tráví méně než dvě hodiny. Méně než 2 hodiny tráví u monitorů přibližně 43 % mladistvých (ve věku 14-18 let), ale jen 33 % dětí (ve věku 8-13 let).

Ještě jedno překvapení anebo o síle testů

$n_1 = n_2 = 30$

$\mu_1 - \mu_2$	t-test zamítl	Wilcoxon zamítl	p(t-test) $<$ p(Wilcoxon)
0.5σ	47795	45766	61273
1σ	96796	96027	69287
1.5σ	99988	99980	73908
2σ	100000	100000	80069
2.5σ	100000	100000	89304
3σ	100000	100000	96549

Závěr

- Wilcoxonův test není jen náhražkou t-testu při nedodržení normality.
- Shoda středních hodnot obecně neimplikuje shodu rozdělení.
- Vhodná vizualizace je za tisíc slov (a čísel).
- Do nedávna jsem podceňoval sílu Wilcoxona, a vy?