Eötvös Loránd Tudomány Egyetem
Természettudományi Kar

Lőrincz Katalin

# Kétmintás statisztikai eljárások

Szakdolgozat

Témavezetők: Móri Tamás, Prőhle Tamás
Valószínűségelméleti és Statisztika Tanszék

Budapest, 2013

# Tartalomjegyzék

**Előszó** **3**

**Köszönetnyilvánítás** **5**

**1. Wilcoxon-Mann-Whitney próba** **6**
1.1. Eredeti Wilcoxon próba . . . . . 6
1.2. Az egybeesések hatása . . . . . 11

**2. A cenzorálás esetei** **14**
2.1. Általánosított Wilcoxon próba jobbról cenzorált adatokra . . . . . 14
2.2. Általánosított Wilcoxon próba kétoldalról cenzorált adatokra . . . . . 17
2.3. Egy nemparaméteres próba csak cenzorált adatokra . . . . . 19

**3. Alkalmazások** **22**
3.1. Az OD görbe . . . . . 22
3.2. Az ROC és az ROL görbe . . . . . 25
3.3. Eltolás detektálás idősorokban . . . . . 28

**4. A Wilcoxon statisztika működése egy idősoron** **31**

**Irodalomjegyzék** **36**

**Függelék** **38**

# Előszó

Jelen dolgozat a rang tesztek témájával foglalkozik, melyeket leggyakrabban az orvosi, a pszichológiai, valamint a szociológia kutatásokban alkalmaznak. Ezen eljárások alanyai személyek, élőlények, illetve eszközök is lehetnek. A kétmintás statisztikai eljárások célja, hogy a vizsgálat alá vont csoportok közötti különbségeket kimutassa vagy jól használható kritériumot találjon a csoportok megkülönböztetésére.

A rang tesztek előnye ([11]) elsősorban az egyszerűség, hiszen a számításokban sokszor a legnagyobb nehézség maga a rangsorolás. Másrészt nem igényelnek speciális feltételeket, általában elegendő a megfigyelések függetlenségét, valamint azt feltenni, hogy egy adott mintában minden megfigyelés ugyanabból a populációból származik. A rang tesztek használhatóságát növeli, hogy akkor is elkészíthetőek a statisztikák, amikor nem állnak rendelkezésre konkrét adatok, csak a mintaelemek sorrendje adott.

Az első fejezet ([1], [2], [6], [12]) az eredeti Wilcoxon-Mann-Whitney statisztikát mutatja be, amelynek rendkívüli előnye, hogy egyszerű és könnyen használható, melynek köszönhetően a fent megjelölt társadalomtudományi területeken kerül leggyakrabban alkalmazásra. Ezen statisztika fontos tulajdonsága, hogy a határeloszlása normális, melynek bizonyítása ebben a fejezetben kerül bemutatásra. A dolgozat kitér a mintaelemek egybeesésének a statisztikára gyakorolt hatására is.

A második fejezetben ([3], [4], [5], [10]) a kétmintás statisztikai eljárásokat cenzorált mintákra terjesztjük ki, amelynek jelentősége, hogy kibővíti ezen eljárások felhasználási területeit. A fejezet külön kitér a jobbról cenzorálás valamint a kétoldalról cenzorálás eseteire, és bemutatásra kerül egy nemparaméteres próba, amely akkor is használható, ha csak cenzorált adatok állnak rendelkezésre.

A harmadik fejezet ([7], [8], [9]) a Wilcoxon-Mann-Whitney statisztika alkalmazásának néhány esetét mutatja be. Az orvostudomány területén gyakori probléma, hogy találjanak egy megfelelő mérőszámot, amelynek segítségével diagnosztizálható egy adott betegség. Az OD-görbe és a Wilcoxon-Mann-Whitney statisztika kapcsolatának köszönhetően megállapítható ehhez egy megfelelő kritikus érték. A fejezetben bemutatásra kerül a Wilcoxon-Mann-Whitney statisztika és az ROC-, illetve az

ROL-görbe kapcsolata. Az említett statisztika érdekes alkalmazása a hirtelen szinteltolások detektálása, melyek fontos információt jelentenek az idősorokban.

A negyedik fejezet egy konkrét példán keresztül mutatja be a Wilcoxon-Mann-Whitney statisztika szinteltolások detektálására való alkalmazását. A havonkénti magyarországi születésszámokon ([12]) beállt változásokat vizsgáljuk 1945 és 2013 között. Az ábrák, és a statisztikák vizuális megjelenítése az R program segítségével készült. A jobb áttekinthetőség kedvéért kinagyított ábrák kerülnek bemutatásra a függelékben.

# Köszönetnyilvánítás

Köszönettel tartozom témavezetőmnek, Móri Tamásnak, a témaválasztásban nyújtott segítségéért és a szakdolgozat megírásához szükséges szakmai támogatásáért.

Köszönettel tartozom továbbá Prőhle Tamás témavezetőmnek a rendszeres konzultációk alkalmával adott támogatásáért, illetve a programozásban nyújtott segítségéért.

# 1. fejezet

# Wilcoxon-Mann-Whitney próba

A Wilcoxon-teszt, vagy ahhoz nagyon hasonló eljárások sok helyen felmerültek az 1900-as évek eleje óta. Ez az oka annak, hogy az irodalomban nem egységes a megnevezése, magát a statisztikát F. Wilcoxon publikálta, de annak szórását, várható értékét és határeloszlását H. B. Mann és D. R. Whitney adta meg.

## 1.1. Eredeti Wilcoxon próba

Legyen $X$ és $Y$ két valószínűségi változó folytonos $F$ és $G$ eloszlásfüggvénnyel. Azt mondjuk, hogy az $X$ valószínűségi változó sztochasztikusan kisebb, mint az $Y$, ha $F(t) > G(t)$ minden $t$-re. A tesztelni kívánt null-hipotézis:

$$H_0: \quad F(t) = G(t) \quad \text{minden } t\text{-re},$$

az ellenhipotézist minden alkalommal külön rögzítjük. Ennek egy lehetséges interpretációja a következő: egy klinikai kísérletben egy eljárás vagy készítmény hatékonyságát szeretnénk vizsgálni, az $X$ valószínűségi változóval azonos eloszlású $n_1$ elemű minta elemei, $x_1, \ldots, x_{n_1}$ a kontrollcsoport, az $Y$ valószínűségi változóval azonos eloszlású $n_2$ elemű minta elemei, $y_1, \ldots, y_{n_2}$ pedig a megkapják a kezelést, vagy alkalmazzák rajtuk a készítményt. Így a null-hipotézis teljesülése azt jelenti, hogy a készítmény hatástalan, az egyoldali ellenhipotézis pedig azt, hogy a készítmény hatásos.

Készítsünk a két mintából egy egyesített rendezett mintát. Minden egyed kapja meg a nagyság szerinti sorszámát, mint rangot. Legyen $W$ az a függvény, ami megszámolja azokat az eseteket, amikor egy $y$ megelőz egy $x$-et a sorban. A feltett folytonos eloszlásból az is következik, hogy $P(x_i = y_j) = 0$, azaz 1 valószínűséggel nincsenek egybeesések a rendezett mintában. Wilcoxon eredeti $T$ statisztikája az

$y_j$-k rangösszege volt, ha $r(y_j)$ jelöli az adott elem rangját, akkor

$$T = \sum_{j=1}^{n_2} r(y_j).$$

Általában

$$W = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T.$$

Mivel csak az $x_i$-k és $y_j$-k egymáshoz viszonyított helyzete érdekel minket a rangsorban, helyettesítsünk minden $y_j$-t 1-gyel és minden $x_i$-t 0-val.

Könnyen látható, hogy ekkor $W$ azt számolja meg, hogy hány alkalommal előz meg egy 1-es egy 0-t. Legyen $\bar{p}_{n_1 n_2}(W)$ azon $0,1$ sorozatok száma, amelyben $n_1$ darab 0 és $n_2$ 1 van, és amelyben $W$-szer előzi meg az 1 a 0-t. Így a következő kifejezéshez jutunk:

$$\bar{p}_{n_1 n_2}(W) = \bar{p}_{n_1-1 n_2}(W - n_2) + \bar{p}_{n_1 n_2 - 1}(W),$$

ahol $\bar{p}_{ij}(W) = 0$ ha $W < 0$ és $\bar{p}_{i0}(W)$, $\bar{p}_{0j}(W)$ nulla vagy egy, attól függően, hogy $W \neq 0$ vagy $W = 0$.

Ha a null-hipotézis teljesül, akkor mind az $(n_1+n_2)!/(n_1!n_2!)$ $0,1$ sorozat egyformán valószínű. Ha $p_{n_1 n_2}(W)$ jelöli annak a $0,1$ sorozatnak a valószínűségét, amiben $W$-szer előzi meg az 1-es a 0-t, akkor

$$p_{n_1 n_2}(W) = \frac{n_1}{n_1+n_2} p_{n_1-1 m}(W - n_2) + \frac{n_2}{n_1+n_2} p_{n_1 n_2 - 1}(W). \tag{1.1.1}$$

Természetesen $p_{n_1 n_2}(W) = p_{n_2 n_1}(W)$.

**1.1.1. Állítás.** *$W$ határeloszlása normális, ha a két minta mérete a végtelenhez tart, és expliciten megadható a szórása és a várható értéke.*

**Bizonyítás:** A határeloszlás bizonyításához a centrális határeloszlás tételt fogjuk használni, amelyhez kiszámoljuk nem csak a szórást és a várható értéket, hanem a magasabb momentumokat is. Mivel annak a valószínűsége, hogy az $i$-edik 1-es megelőzi a $j$-edik 0-t $1/2$, ezért

$$E_{n_1 n_2}(W) = \frac{n_1 n_2}{2}. \tag{1.1.2}$$

Legyen $w = W - n_1 n_2/2$, és számoljuk ki az (1.1.1) segítségével $E_{n_1 n_2}(w^2)$-et.

$$E_{n_1 n_2}(w^2) = \sum_W (W - \frac{n_1 n_2}{2})^2 p_{n_1 n_2}(W)$$

és kifejtve:

$$E_{n_1 n_2}(w^2) = \frac{n_1}{n_1+n_2} E_{n_1-1 n_2}(w^2) + \frac{n_2}{n_1+n_2} E_{n_1 n_2 - 1}(w^2) + \frac{n_1 n_2}{4}, \tag{1.1.3}$$

ahol $E_{n_1n_2}(w)$ jelöli $(W - n_1n_2/2)$ várhatóértékét azok között a sorozatok között, amelyekben $n_1$ darab 0 és $n_2$ 1-es van. Az (1.1.3)-ból közvetlenül adódik, hogy

$$E_{n_10}(w^2) = E_{0n_2}(w^2) = 0. \tag{1.1.4}$$

Teljes indukcióval az (1.1.4) kezdeti feltétellel adódik, hogy:

$$E_{n_1n_2}(w^2) = n_1n_2(n_1 + n_2 + 1)/12, \tag{1.1.5}$$

amivel $W$ szórását megkaptuk. Már csak a határeloszlás van hátra, amelyhez kiszámoljuk a páros momentumokat. A negyedik momentum hasonlóan adódik az (1.1.1)-ből:

$$\begin{aligned} E_{n_1n_2}(w^4) = \frac{n_1}{n_1+n_2}E_{n_1-1n_2}(w^4) + \frac{n_2}{n_1+n_2}E_{n_1n_2-1}(w^4) + \\ + \frac{n_1n_2}{16}(2n_1^2n_2 + 2n_1n_2^2 - n_1^2 - n_2^2 - n_1n_2). \end{aligned} \tag{1.1.6}$$

Az (1.1.6) kezdeti feltétele közvetlen számolással adódik:

$$E_{n_10}(w^4) = E_{0n_2}(w^4) = 0. \tag{1.1.7}$$

Teljes indukcióval bizonyítható, hogy

$$E_{n_1n_2}(w^4) = \frac{n_1n_2(n_1+n_2+1)}{240}(5n_1^2n_2 + n_1n_2^2 - 2n_1^2 - 2n_2^2 + 3n_1n_2 - 2n_1 - 2n_2) \tag{1.1.8}$$

kielégíti az (1.1.6)-ot az (1.1.7) kezdeti feltétellel.

A $w$ határeloszlásának bizonyításához, amikor $n_1, n_2$ végtelenhez tart, szükségünk van az $r$-edik momentumokra. Ugyanazt az eljárást alkalmazva, mint amivel a második és a negyedik momentumot megkaptuk, és felhasználva, hogy az összes páratlanadik momentum 0, azt kapjuk, hogy

$$\begin{aligned} E_{n_1n_2}(w^{2r}) = \frac{1}{n_1+n_2}\sum_{\alpha=0}^{r}\binom{2r}{2\alpha}\frac{1}{4^\alpha}\left\{n_1n_2^{2\alpha}E_{n_1-1n_2}(w^{2r-2\alpha}) + \right. \\ \left. + n_1^{2\alpha}n_2E_{n_1n_2-1}(w^{2r-2\alpha})\right\}. \end{aligned} \tag{1.1.9}$$

Azt már tudjuk, hogy $r = 1, 2$-re $E_{n_1n_2}(w^{2r})$ az $n$ és $m$ $3r$-ed fokú polinomja valamint osztható $n_1n_2(n_1 + n_2 + 1)$-gyel. Feltéve, hogy $E_{n_1n_2}(w^{2\alpha})$ $\alpha < r$-re az $n$ és $m$ $3\alpha$-fokú polinomja, ami osztható $n_1n_2(n_1 + n_2 + 1)$-gyel, megmutatjuk, hogy található egy olyan $3r$ fokú polinom, ami osztható $n_1n_2(n_1 + n_2 + 1)$-gyel, ami kielégíti az előző egyenletet és azt a kezdeti feltételt, hogy $E_{n_10}(w^{2r}) = E_{0n_2}(w^{2r}) = 0$. Az utolsó feltétel triviálisan teljesül, ha $E_{n_1n_2}(w^{2r})$ osztható $n_1n_2(n_1 + n_2 + 1)$-gyel. Behelyettesítjük a polinomot az (1.1.9)-be és megmutatjuk, hogy a tagok egyenként kiszámolhatóak. Átrendezéssel kapjuk, hogy

$$\begin{aligned} E_{n_1n_2}(w^{2r}) - \frac{n_1}{n_1+n_2}E_{n_1-1n_2}(w^{2r}) - \frac{n_2}{n_1+n_2}E_{n_1n_2-1}(w^{2r}) = \\ = \frac{1}{n_1+n_2}\sum_{\alpha=1}^{r}\binom{2r}{2\alpha}\frac{1}{4\alpha}\left\{n_1n_2^{2\alpha}E_{n_1-1n_2}(w^{2r-2\alpha}) + \right. \\ \left. + n_1^{2\alpha}n_2E_{n_1n_2-1}(w^{2r-2\alpha})\right\}. \end{aligned} \tag{1.1.10}$$

Mivel $\lambda < r$-re $E_{n_1n_2}(w^{2\lambda}) = n_1n_2(n_1+n_2+1)P^{3\lambda-3}_{n_1n_2}$ ahol $P^{3\lambda-3}_{n_1n_2}$ az $n_1$ és $n_2$ $(3\lambda-3)$-fokú polinomja, a fenti egyenlet egyszerűsíthető:

$$E_{n_1n_2}(w^{2r}) - \frac{n_1}{n_1+n_2}E_{n_1-1n_2}(w^{2r}) - \frac{n_2}{n_1+n_2}E_{n_1n_2-1}(w^{2r}) = n_1n_2Q^{3r-3}_{n_1n_2}, \quad (1.1.11)$$

ahol $Q^{3r-3}_{n_1n_2}$ az $n_1$ és $n_2$ $(3r-3)$-fokú polinomja. Most legyen

$$E_{n_1n_2}(w^{2r}) = n_1n_2(n_1+n_2+1)\sum_{\substack{i,j=0\\ i+j\le 3r-3}}^{3r-3} a_{ij}n_1^in_2^j,$$

ahol $a_{ij} = a_{ji}$ adottak. Behelyettesítve az (1.1.11)-be:

$$\sum_{i,j} a_{ij}\left[(n_1+n_2+1)n_1^in_2^j - (n_1-1)^{i+1}n_2^j - n_1^i(n_2-1)^{j+1}\right] = Q^{3r-3}_{n_1n_2}$$

és átrendezve

$$\sum_{\substack{i,j=0\\ i+j\le 3r-3}}^{3r-3} a_{ij}\left[n_1^in_2^j + \sum_{\alpha=0}^{i}\binom{i+1}{\alpha}(-1)^{i-\alpha}(n_1^jn_2^\alpha + n_1^\alpha n_2^j)\right] = Q^{3r-3}_{n_1n_2}. \quad (1.1.12)$$

Nézzük a $(3r-3)$-fokú tagokat. Ebben az esetben $i+j = 3r-3$ és $\alpha = i$ azt adja, hogy

$$\sum_{i=0}^{3r-3} a_{i3r-3-i}\left[n_1^in_2^{3r-3-i} + (i+1)(n_1^{3r-3-i}n_2^i + n_1^in_2^{3r-3-i})\right] = Q^{3r-3}_{n_1n_2}$$

vagy

$$3r\sum_{i=0}^{3r-3} a_{i3r-3-i}n_1^in_2^{3r-3-i}. \quad (1.1.13)$$

A $(3r-3)$-fokú tagokat egyenlővé téve a megfelelőkkel $Q^{3r-3}_{n_1n_2}$-ből, kiszámolható $a_{i3r-3-i}$, $(i = 0, 1, \ldots, 3r-3)$.

Most feltesszük, hogy $a_{ij}$ ismert, ha $i+j > 3r-3-(k-1)$, és megkapjuk $a_{ij}$-t, ha $i+j = 3r-3-k$. Nézzük az (1.1.12) azon tagjait, amelyek foka $(3r-3-k)$. Ezek a tagok akkor jelennek meg, amikor

$$i+j = 3r-3, \alpha = i-k; \ldots; i+j = 3r-3, \alpha = i.$$

Mind, kivéve az utolsót, tartalmaz olyan tényezőket, amelyeket már megkaptunk. Az utolsó pedig egyszerűsíthető:

$$(3r-3)\sum_{i=0}^{3r-3-k} a_{i3r-3-k-i}n_1^in_2^{3r-3-k-i}.$$

Így az $a_{i3r-3-k-i}$ együtthatók $i = 0, 1, \ldots, 3r-3-k$-ra számolhatók az ismert $a_{ij}$-k és $Q^{3r-3}_{n_1n_2}$ segítségével. Ez bizonyítja, hogy $E_{n_1n_2}(w^{2r})$ $3r$-fokú polinom $n_1$-ben és $n_2$-ben, és osztható $n_1n_2(n_1+n_2+1)$-gyel.

Most a $3r$-fokú tagok együtthatóit vizsgáljuk. Legyen $\lambda = 1, 2$.

$$E_{n_1n_2}(w^{2\lambda}) = \frac{2\lambda-1)\cdots 5\cdot 3\cdot 2}{12^\lambda}(n_1n_2)^\lambda(n_1+n_2+1)^\lambda + (3\lambda\text{-nál kisebb fokú tagok}).$$

Feltesszük, hogy ez érvényes $\lambda < r$-re és megmutatjuk, hogy $\lambda = r$-re is igaz. Az (1.1.10) jobb oldalába behelyettesítve

$$\begin{aligned}
&\frac{1}{n_1+n_2}\binom{2r}{2}\frac{1}{4}\cdot\\
&\cdot\left\{n_1n_2^2\left[\frac{(2r-3)\ldots 5\cdot 3\cdot 1}{12^{r-1}}(n_1-1)^{r-1}n_2^{r-1}(n_1+n_2)^{r-1}\right]+\right.\\
&+n_1^2n_2\left[\frac{(2r-3)\cdots 5\cdot 3\cdot 1}{12^{r-1}}n_1^{r-1}(n_2-1)^{r-1}(n_1+n_2)^{r-1}\right]+\\
&\left.+\,(3r\text{-nél kisebb fokú tagok})\right\} \qquad (1.1.14)
\end{aligned}$$

vagy

$$\begin{aligned}
&\frac{r(2r+1)}{4}\left\{(n_1+n_2)^{r-2}\left[\frac{(2r-3)\cdots 5\cdot 3\cdot 1}{12^{r-1}}\right]\cdot\right.\\
&\left.\cdot\left[n_1(n_1-1)^{r-1}n_2^{r+1}+n_1^{r+1}n_2(n_2-1)^{r-1}\right]+(3\lambda\text{-nál kisebb fokú tagok})\right\} \qquad (1.1.15)
\end{aligned}$$

ami tovább egyszerűsíthető

$$\frac{3r(2r-1)\cdots 5\cdot 3\cdot 1}{12^r}(n_1n_2)^r(n_1+n_2)^{r-1} + (3r-1\text{-nél kisebb fokú tagok}).$$

Összehasonlítva az (1.1.13) együtthatóival, $n_1n_2$-vel szorozva

$$n_1n_2\sum_{i=0}^{3r-3}a_{i3r-3-i}n_1^in_2^{3r-3-i} = \frac{(2r-1)\cdots 5\cdot 3\cdot 1}{12^r}(n_1n_2)^r(n_1+n_2)^{r-1}$$

vagy

$$\begin{aligned}
E_{n_1n_2}(w^{2r}) = &\frac{(2r-1)\cdots 5\cdot 3\cdot 1}{12^r}(n_1n_2)^r(n_1+n_2+1)^r\\
&+(3r\text{-nél kisebb fokú tagok}). \qquad (1.1.16)
\end{aligned}$$

Megmutatjuk, hogy $E_{n_1n_2}(w^{2r})$ legfeljebb $2r$-fokú $n_1$-ben és $n_2$-ben, amit $r = 1, 2$-re már láttunk. Feltéve, hogy a kisebb momentumokra igaz az (1.1.10) jobb oldalán, ami tovább egyszerűsödik $n_1n_2Q^{3r-3}_{n_1n_2}$-é, ami legfeljebb $2r-1$-fokú $n_1$-ben. Megint összehasonlítjuk az együtthatókat az (1.1.12)-ben. Azokra a tagokra, amelyek $3r-3$-fokúak, már láttuk, hogy $n_1$-ben legfeljebb $2r-2$-fokúak. Azokra a tagokra, amelyeknek a foka $3r-4$, használjuk $i+j = 3r-3$, $\alpha = i-1$-et és $i+j = 3r-4$,

$\alpha = i$-t. Az első eset nem ad $n_1$-ben nagyobb fokú tagot, mint $2r-2$, tehát amikor megoldjuk az $a_{i3r-4-i}$ együtthatókra, azoknak a tagoknak az együtthatói, amelyek $n_1$-ben magasabb fokúak, mint $2r-2$, nullának kell lenniük. Megismételve a folyamatot, nem lesz magasabb fokú tag sem $n_1$-ben, sem $n_2$-ben az (1.1.12) bal oldalán, mint $2r-2$. Ebből adódik, hogy $E_{n_1n_2}(w^{2r})$-ben a legmagasabb előforduló fokszám $n_1$-ben és $n_2$-ben legfeljebb $2r$ lehet.

Most nézzük az arányt

$$\begin{aligned} I = \frac{E_{n_1n_2}(w^{2r})}{(E_{n_1n_2}(w^2))^r} &= \frac{\frac{(2r-1)\cdots 5\cdot 3\cdot 1}{12^r}(n_1n_2)^r(n_1+n_2+1)^r}{[n_1n_2(n_1+n_2+1)/12]^r} + \\ &\quad + \frac{b}{[n_1n_2(n_1+n_2+1)/12]^r} = \\ &= (2r-1)\cdots 5\cdot 3\cdot 1 + \frac{b}{(n_1n_2)^r(n_1+n_2+1)^r}, \end{aligned} \tag{1.1.17}$$

ahol $b = 3r$-nél kisebb fokú tagok, amelyekben $n_1$ és $n_2$ foka legfeljebb $2r$. Tehát

$$\lim_{n_1n_2\to\infty} I = (2r-1)\cdots 5\cdot 3\cdot 1$$

és ebből már a centrális határeloszlás tétellel következik, hogy a határeloszlás normális.

## 1.2. Az egybeesések hatása

Az egybeesések Wilcoxon-statisztikára gyakorolt hatásának vizsgálatához megmutatjuk a Wilcoxon $W$-statisztikája és a Kendall által a rangkorrelációnál bemutatott $S$ mennyiség kapcsolatát. Amikor $r$ darab $(u_k, v_k)$ számpár adott, akkor $S$ a következőképpen számolható:

$$\begin{array}{ll} -1 & \text{ha } (u_h-u_k)(v_h-v_k) < 0 \\ 0 & \text{ha } (u_h-u_k)(v_h-v_k) = 0 \\ 1 & \text{ha } (u_h-u_k)(v_h-v_k) > 0. \end{array}$$

A fenti értékeket adjuk össze minden $(h,k)$ pár esetében, amikor $h<k$. Ha a két minta $x_1,\dots,x_n$ és $y_1,\dots,y_m$, és a fenti definícióban $r=m+n$ esetén behelyettesítjük a két minta értékeit a megadott sorrendben méghozzá $u_1,\dots,u_n,u_{n+1},\dots,u_r$-rel és 0-t vagy 1-et ebben a sorrendben $v_k$-val, ha $u_k = x_i$ néhány $i$-re vagy $u_k = y_j$ néhány $j$-re ebben a sorrendben, akkor a következő adódik:

$$2W + S = nm.$$

A legegyszerűbben ez úgy látható, ha megnézzük $2W+S$ értékét minden $(h,k)$ párra. Ez egyenlő lesz 1-gyel, ha $v_h = 0$ és $v_k = 1$, egyébként pedig 0-val. Az értékek

összege így $nm$. A fenti egyenlőség igaz, ha nincsenek egybeesések a két mintában. Természetes, hogy $W$-t úgy definiáljuk, hogy ezt az egyenlőséget terjesztjük ki arra az esetre, amikor vannak egybeesések. Mivel egy $(x_i, y_j)$ párra az $S$ értéke 0, ha $x_i = y_j$, a $W$ értéke 1/2 kell, hogy legyen. Kendall kiszámolta a várható értéket és a standard szórást $S$-re amellett a feltevés mellett, hogy a $v_1, \ldots, v_r$ mennyiségek meghatározott rangsorára $u_1, \ldots, u_r$ minden $r!$ lehetséges permutációja egyformán valószínű. Ez a feltétel ebben az esetben akkor teljesül, ha az $x_1, \ldots, x_n$ és $y_1, \ldots, y_m$ mintákat véletlenszerűen vettük ugyanabból a populációból, így az eloszlásoknak nem kell többé folytonosnak lenniük. Így $W$ várható értéke és szórása a nullhipotézis mellett megkapható a Kendall-formulából. Ezért

$$E(S) = 0$$

és

$$var(S) =$$

$$= \frac{1}{18}\left\{r(r-1)(2r+5) - \sum_t t(t-1)(2t+5) - \sum_s s(s-1)(2s+5)\right\} +$$
$$+\frac{1}{9r(r-1)(r-2)}\left\{\sum_t t(t-1)(t-2)\right\}\left\{\sum_s s(s-1)(s-2)\right\} +$$
$$+\frac{1}{2r(r-1)}\left\{\sum_t t(t-1)\right\}\left\{\sum_s s(s-1)\right\}, \quad (1.2.1)$$

ahol $\sum_t$ az $u_1, \ldots, u_r$ különböző egybeesésein, $\sum_s$ pedig a $v_1, \ldots, v_r$ egybeesésein történik; $t$ és $s$ pedig sorrendben az elemszámot jelöli az egyenlő számok minden csoportjában $u_1, \ldots, u_r$ és $v_1, \ldots, v_r$ között. A $2W + S = nm$ egyenlőségből:

$$E(W) = \frac{1}{2}nm - E(S) = \frac{1}{2}nm \quad (1.2.2)$$

és

$$var(W) = \frac{1}{4}var(S). \quad (1.2.3)$$

A $v_1, \ldots, v_r$ csoport $n$ darab 0-ból és $m$ darab 1-ből áll, így $s$ értéke az (1.2.1)-ben $n$ és $m$, így

$$\sum_s s(s-1)(2s+5) = n(n-1)(2n+5) + m(m-1)(2m+5),$$

$$\sum_s s(s-1)(s-2) = n(n-1)(n-2) + m(m-1)(m-2),$$

$$\sum_s s(s-1) = n(n-1) + m(m-1).$$

Behelyettesítve az 1.2.1-be és az 1.2.3-ba némi egyszerűsítés után a következőt kapjuk:

$$var(W) = \frac{1}{12}nm(n+m+1) - \frac{1}{12}\sum_t t(t-1)(2t+5) +$$
$$+\frac{n(n-1)(n-2)+m(m-1)(m-2)}{36(n+m)(n+m-1)(n+m-2)}\sum_t t(t-1)(t-2) +$$
$$+\frac{n(n-1)+m(m-1)}{8(m+n)(m+n-1)}\sum_t t(t-1), \tag{1.2.4}$$

ahol $\sum_t$ az $x_1, \ldots, x_n$ és $y_1, \ldots, y_m$ egyesített mintaértékek közötti egybeesésekre történik. Ha nincs egybeesés, akkor ez az eredeti Mann-Whitney-Wilcoxon eredményre egyszerűsödik:

$$E(W) = \frac{1}{2}nm,$$

és

$$var(W) = \frac{1}{12}nm(n+m+1).$$

Az utolsó három egyenlőségből akár indukcióval könnyen látható, hogy $var(W)$ csökken a megfigyelések közötti egybeesések jelenlétével. Ezek az eredmények egy lépést jelentenek arrafelé, hogy a Wilcoxon tesztet bármilyen populációból vett mintára használhassuk.

# 2. fejezet

# A cenzorálás esetei

Az orvosi, a szociológiai és a pszichológiai vizsgálatoknál, az élettartam-adatok elemzéseknél gyakran fordul elő cenzorálás. Akkor beszélünk cenzorálásról, ha nem minden egyedről figyelhető meg pontos adat, akár azért, mert a várt esemény egy bizonyos időpontnál később, akár azért, mert korábban következett be. Amikor egy mintában sok a cenzorált adat, akkor ezek nem hagyhatóak figyelmen kívül, mert ezáltal jelentős információt veszítenénk, és esetleg torzulna a kialakult kép. Ez a torzulás jobbról cenzorálás esetén azt jelenti, hogy a magasabb értékek elhagyásával a megmaradt megfigyelések eloszlása a kisebb értékek felé tér el.

## 2.1. Általánosított Wilcoxon próba jobbról cenzorált adatokra

A fentiek alapján önkényesen jobbról cenzorált mintákra terjesztjük ki a Wilcoxon tesztet, amely függ a megfigyelések szerkezetétől. A vizsgálni kívánt nullhipotézis:

$$H_0: \ F_1(t) = F_2(t)$$

az ellenhipotézis pedig:

$$\begin{aligned} H_1: &\ F_1(t) < F_2(t) \ (t \leq T) \text{ vagy} \\ H_2: &\ F_1(t) < F_2(t) \text{ vagy } F_1(t) > F_2(t) \ (t \leq T) \end{aligned}$$

ahol $F_1, F_2$ a megfigyelések (diszkrét vagy folytonos) eloszlása, és $T$ a megfigyelések felső határa.

Ez a probléma klinikai kísérleteknél fordul elő, amikor összehasonlítjuk két kezelés hatékonyságát arra nézve, hogy mennyire tudják meghosszabbítani az életet és fenntartani a páciens jóllétét. A páciensek sorozatosan lépnek be ezekbe a kísérletekbe, és véletlenszerűen kapják a két kezelés valamelyikét. $T$ idővel a kísérlet

kezdete után vizsgálják a meghibásodási időt (halál vagy visszaesés) és a cenzoráltságot (életben van, vagy a tünetek enyhültek $T$-ben). Jelölje $n_i - r_i$ a meghibásodott egyedek számát és $r_i$ a $T$-ben cenzoráltat. Mivel a páciensek különböző időkben lépnek be a kísérletbe a $[0, T]$ intervallumban, a cenzorálás ideje is különböző a páciensek között. Tegyük fel, hogy $n_1, n_2$ darabszámú egyed kapja az $A$ illetve $B$ kezelést, és a megfigyelésünk:

$$\left.\begin{array}{ll} x_1^{'}, \ldots, x_{r_1}^{'} & r_1 \text{ cenzorált} \\ x_{r_1+1}, \ldots, x_{n_1} & n_1 - r_1 \text{ meghibásodás} \end{array}\right\} A \text{ kezelés}$$

$$\left.\begin{array}{ll} y_1^{'}, \ldots, y_{r_2}^{'} & r_2 \text{ cenzorált} \\ y_{r_2+1}, \ldots, y_{n_2} & n_2 - r_2 \text{ meghibásodás} \end{array}\right\} B \text{ kezelés,}$$

ahol $x_i, y_j$ meghibásodási, $x_i^{'}, y_j^{'}$ pedig cenzorálási idők, mindkettő a belépéstől mérve. Megengedjük az egybeesést a meghibásodások és a cenzorálások között is. A nullhipotézis tehát ebben az esetben az, hogy az $A$ és $B$ kezelés egyformán hatékony, az egyoldali ellenhipotézis az, hogy $A$ hatékonyabb, mint $B$, a kétoldali pedig az, hogy valamelyik kezelés hatékonyabb a másiknál. Definiáljuk:

$$U_{ij} = \begin{cases} -1 & \text{ha } x_i < y_j \text{ vagy } x_i \leq y_j^{'} \\ 0 & \text{ha } x_i = y_j \text{ vagy } (x_i^{'}, y_j^{'}) \text{ vagy } x_i^{'} < y_j \text{ vagy } y_j^{'} < x_i \\ 1 & \text{ha } x_i > y_j \text{ vagy } x_i^{'} \geq y_j \end{cases}$$

és számoljuk ki a $W = \sum_{i,j} U_{ij}$ statisztikát, ahol az összegzés $n_1 n_2$ összehasonlításra végződik. Tehát $W$-hez hozzájárul a két minta minden olyan összehasonlítása, ahol mindkét egyed meghibásodott (kivéve az egybeesést), és minden olyan összehasonlítás, ahol a cenzorált egyed tovább életben maradt, mint amelyik meghibásodott. Könnyen látszik, hogy

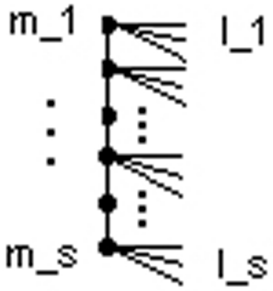
$$W = n_2(n_1 + n_2 + 1) - 2W^{'}$$

ahol $W^{'}$ a második minta rangjainak összege a kombinált rendezett mintában. Továbbá

$$W = 2U^{'} - n_1 n_2$$

ahol $U^{'}$ megszámolja, hogy hány olyan megfigyelés van a második mintában, amely megelőz egyet az első mintából a kombinált rangsorban.

Most nézzük meg $W$ feltételes várható értékét és szórását. Az $n_1, n_2$ megfigyelés szerkezete legyen a következő:



ahol $m_i$ a cenzorálatlan $i$ rangú megfigyelések száma a cenzorálatlan megfigyelések rangsorában; $l_i$ az olyan jobbról cenzorált megfigyelések száma, amelyek az $i$ és az $(i+1)$ rangú megfigyelések közé esnek. A vertikális vonalon a pontok meghibásodott egyedek rangsorának felelnek meg, ezek az $s$ meghatározott meghibásodási pontokon következnek be. Bármely meghibásodási vagy cenzorálási megfigyelési sorozat reprezentálható ilyen diagrammal. Ha van cenzorált megfigyelés az első meghibásodás előtt, akkor ezeket bevehetjük úgy, hogy $l_1$-nek számoljuk őket $m_1 = 0$ értékkel. Általában azonban az ilyen megfigyeléseket a gyakorlatban ki szokták zárni, mivel nem hordoznak információt az $A$ és $B$ eljárás közötti különbségről. Ez a kizárás nincs hatással a próba várható értékére és szórására, mivel azok az adott megfigyelés szerkezetétől függnek.

Tegyük most fel, hogy $H_0$ igaz, és a meghibásodás kockázata ugyanaz a két csoportban. Ha a két minta egyedei a szerkezetben különböző módon vannak jelölve, akkor $\binom{n_1+n_2}{n_1}$ különféle elosztása van az $n_1, n_2$ megfigyelésnek. Jelölje $W$ $H_0$-melletti várható értékét $E(W\,|P, H_0)$ és szórását $var(W\,|P, H_0)$, ahol $P$ a megfigyelések szerkezete. Összesen $\frac{(n_1+n_2)!}{n_1!n_2!}$ egyformán valószínű minta vezet ugyanahhoz a megfigyelési szerkezethez. Természetesen

$$E(W\,|P, H_0) = 0$$

a szimmetria miatt. A szórás képlete pedig:

$$\begin{aligned} var(W\,|P, H_0) = \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-1)} \Bigg\{ \sum_{i=1}^{s} m_i M_{i-1}(M_{i-1}+1) + \\ + \sum_{i=1}^{s} l_i M_i (M_i+1) + \sum_{i=1}^{s} m_i (n_1+n_2-M_i-L_{i-1}) \cdot \\ \cdot (n_1+n_2-3M_{i-1}-m_i-L_{i-1}-1) \Bigg\}, \end{aligned} \tag{2.1.1}$$

ahol

$$M_j = \sum_{i=1}^{j} m_i, \qquad M_0 = 0, \qquad L_j = \sum_{i=1}^{j} l_i, \qquad L_0 = 0.$$

Ha nincsenek egybeesések és cenzorálás, azaz $m_1 = \ldots = m_s = 1$, $l_1 = \ldots = l_s = 0$, és $s = n_1+n_2$, akkor az eredeti Wilcoxon-Mann-Whitney statisztika szórását kapjuk. Ekkor $P$ egyszerűen az $n_1 + n_2$ megfigyelés rangsorolása.

A $W$ statisztika aszimptotikusan normális a null-hipotézis mellett, és

$$Z = \frac{W}{\sqrt{var(W\,|P, H_0)}}$$

értéke aszimptotikusan normális 0 várható értékkel és 1 szórással.

## 2.2. Általánosított Wilcoxon próba kétoldalról cenzorált adatokra

Feltesszük, hogy $n_1$ és $n_2$ megfigyelés van véletlenszerűen elosztva az $A$ és $B$ csoportba, és

$$\left.\begin{array}{ll} x_1^{'}, \ldots, x_{r_1}^{'} & r_1 \text{ jobbról cenzorált} \\ x_{r_1+1}^{''}, \ldots, x_{r_1+s_1}^{''} & s_1 \text{ balról cenzorált} \\ x_{r_1+s_1+1}, \ldots, x_{n_1} & n_1 - r_1 - s_1 \text{ nem cenzorált} \end{array}\right\} \text{megfigyelés az } A \text{ csoportban}$$

$$\left.\begin{array}{ll} y_1^{'}, \ldots, y_{r_2}^{'} & r_2 \text{ jobbról cenzorált} \\ y_{r_2+1}^{''}, \ldots, y_{r_2+s_2}^{''} & s_2 \text{ balról cenzorált} \\ y_{r_2+s_2+1}, \ldots, y_{n_2} & n_2 - r_2 - s_2 \text{ nem cenzorált} \end{array}\right\} \text{megfigyelés a } B \text{ csoportban,}$$

ahol $x_i, y_i$ rendezett megfigyelések, $x_i^{'}, y_i^{'}$ jobbról, vagy a kísérlet végén cenzorált és $x_i^{''}, y_i^{''}$ balról, vagy a kísérlet elején cenzorált megfigyelések. A mintavétel feltétele, hogy pozitív valószínűséggel minden megfigyelés cenzorálatlan, valamint jobbról vagy balról cenzorált, így a minta szerkezete a fenti típusú. Feltesszük, hogy a cenzorálási feltételek ugyanazok mindkét csoportban, és az egybeesést sem a meghibásodásoknál sem a cenzorálásoknál nem zárjuk ki.

A vizsgálni kívánt nullhipotézis ismét:

$$H_0: \; F_1(t) = F_2(t)$$

amit a

$$\begin{array}{l} H_1: \; F_1(t) < F_2(t) \; (t \leq T) \text{ vagy} \\ H_2: \; F_1(t) < F_2(t) \text{ vagy } F_1(t) > F_2(t) \; (t \leq T) \end{array}$$

ellenhipotézisre tesztelünk, ahol $F_1, F_2$ a megfigyelések (diszkrét vagy folytonos) eloszlása, és $T$ a megfigyelések felső határa. Legyen

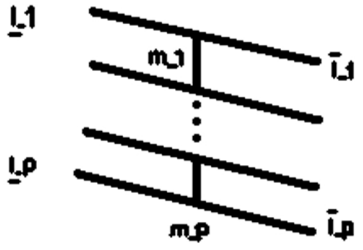
$$U_{ij} = \begin{cases} -1 & \text{ha } x_i < y_j \text{ vagy } x_i \leq y_j^{'} \text{ vagy } x_i^{''} \leq y_j, y_j^{'} \\ 1 & \text{ha } x_i > y_j \text{ vagy } x_i^{'} \geq y_j \text{ vagy } x_i, x_i^{'} \geq y_j^{''} \\ 0 & \text{egyébként} \end{cases}$$

és számoljuk ki a

$$W = \sum_{ij} U_{ij}$$

statisztikát, ahol az összegzés az összes értelmes párosításra történik.

A megfigyelések $P$ szerkezete jól leírható a következő ábrával:




ahol $m_i$ az $i$ rangú cenzorálatlan megfigyelések száma; $\underline{l}_i$ azon balról cenzorált megfigyelések száma, amelyek értéke az $i-1$ és $i$ rangú megfigyelések közé esik; $\overline{l}_i$ pedig azon jobbról cenzorált megfigyelések száma, amelyek értéke az $i$ és $i+1$ rangú megfigyelések közé esik.

Most megnézzük $W$ eloszlását, feltételes várható értékét és szórását. Jelölje a várható értéket $E(W\,|P, H_0, n_1, n_2)$ és a szórást $var(W\,|P, H_0, n_1, n_2)$, ahol $P$ a megfigyelések szerkezete. A duplán cenzorált esetben is összesen $\frac{(n_1+n_2)!}{n_1!n_2!}$ egyformán valószínű minta vezet ugyanahhoz a megfigyelési szerkezethez. Természetesen

$$E(W\,|P, H_0) = 0$$

a szimmetria miatt. A szórás képlete pedig: .

$$\begin{aligned} var(W\,|P, H_0, n_1, n_2) &= \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-1)} \cdot \\ \cdot \Bigg\{ \sum_{i=1}^{p} m_i (M_{i-1} + \underline{L}_i)(M_{i-1} + \underline{L}_i + 1) &+ \sum_{i=1}^{p} \overline{l}_i (M_i + \underline{L}_i)(M_i + \underline{L}_i + 1) + \\ &+ \sum_{i=1}^{p} m_i (M_p + \overline{L}_p - M_i - \overline{L}_{i-1}) \cdot \\ &\cdot (M_p - \overline{L}_p - 3M_{i-1} - \overline{L}_{i-1} - 2\underline{L}_i - m_i - 1) + \\ + \sum_{i=1}^{p} \underline{l}_i (M_p + \overline{L}_p - M_{i-1} - \overline{L}_{i-1})(M_p + \overline{L}_p - M_{i-1} - \overline{L}_{i-1} - 1) &\Bigg\}, \quad (2.2.1) \end{aligned}$$

ahol

$$M_j = \sum_{i=1}^{j} m_i, \quad M_0 = 0, \quad \overline{L}_j = \sum_{i=1}^{j} \overline{l}_i, \quad \overline{L}_0 = 0, \quad \underline{L}_j = \sum_{i=1}^{j} \underline{l}_i, \quad \underline{L}_0 = 0.$$

A $W$ statisztika határeloszlása ebben az esetben is normális, és

$$Z = \frac{W}{\sqrt{var(W\,|P, H_0, n_1, n_2)}}$$

aszimptotikusan standard normális eloszlású.

## 2.3. Egy nemparaméteres próba csak cenzorált adatokra

Azokon a területeken, mint az orvostudomány és a társadalom tudomány, gyakran fordul elő az az eset, amikor az egyedeket az élettartamuk alatt csak egyszer tudjuk megfigyelni. Ilyen lehet például az, amikor azt vizsgálják, hogy egy adott képesség mikor alakul ki. Mivel egyszer vizsgálják a populációt, ezért egyrészt azoknál az egyedeknél akik már képesek a feladatra, nem derül ki, hogy mikor tanulták meg, így ezek az adatok balról cenzoráltak a vizsgálatnál, másrészt azoknál az egyedeknél, akiknél még nem alakult ki a képesség, a várakozási idő jobbról cenzorált.

Vezessük be a következő jelöléseket: $\underline{X} = (x_1, \ldots, x_n)$ vektor jelöli az igazi független, azonos eloszlású túlélési időt, $F$ abszolut folytonos eloszlásfüggvénnyel és $f$ sűrűségfüggvénnyel. Legyen $\underline{T} = (T_1, \ldots, T_n)$ a független azonos eloszlású megfigyelési idő $G$ abszolut folytonos eloszlásfüggvénnyel, és $g$ sűrűségfüggvénnyel. Tegyük még fel, hogy $\underline{X}$ és $\underline{T}$ is függetlenek. Nem figyeljük meg $\underline{X}$-et és $\underline{T}$-t, csak azt, hogy $T_i$-ben az $i$-edik egyed már korábban képessé vált ($D_i = I(x_i \leq T_i)$), de az idejét ($x_i$-t), hogy mikor történt vagy később mikor fog történni, nem. A megfigyelésünk tehát: $[T_i, D_i = I(x_i \leq T_i)]$. Ezen a megfigyelésen alapul $G$ likelihood-függvénye:

$$\prod_{i=1}^{n} g(T_i) F(T_i)^{D_i} \left[1 - F(T_i)\right]^{1-D_i}$$

és a túlélési idő eloszlásának, $F$-nek a likelihood-függvénye:

$$L_n(F) = \prod_{i=1}^{n} F(T_i)^{D_i} \left[1 - F(T_i)\right]^{1-D_i}.$$

A nemparaméteres maximum likelihood-becslő (NPMLE) $\hat{F}_n$ az $L_n(F)$-ből kapható egy egyszerű nem iteratív algoritmuson keresztül. Tegyük fel, hogy $\underline{T}$ komponensei növekvő sorrendben vannak: $T_1 \leq T_2 \leq \ldots \leq T_n$, és $D_1, \ldots, D_n$ eszerint. Legyen $i = 0, 1, \ldots, n$-re

$$H_i = \sum_{j=1}^{i} D_j$$

a $D_j$-k kumulatív összege, és tekintsük a $H_i$-k ábrázolásából nyert lépcsős függvényt.

Most nézzük a függvény legnagyobb konvex minoránsát, azaz a lépcső pontjainak összekötésével nyert töröttvonalat azzal a megkötéssel, hogy a kapott töröttvonalnak konvexnek kell lennie. Így $\hat{F}_n(t) = \hat{F}_n(T_i)$ jobbról folytonos minden $t \in (T_i, T_{i+1})$-re, ahol $T_0 = 0$ és $T_{n+1} = \infty$. Tekintsük $\hat{F}_n$ integrálját $\sqrt{n}$-nel normálva. Ha $\tau < \infty$

$$\sqrt{n} \int_0^\tau \left[\hat{F}_n(t) - F(t)\right] dt$$

és

$$\sqrt{n} \int_0^\tau \left[\hat{F}_n^2(t) - F^2(t)\right] dt$$

normális határeloszlású 0 várható értékkel, és a szórás:

$$\int_0^\tau \frac{F(t)\,[1 - F(t)]}{g(t)} dt$$

és

$$4 \int_0^\tau \frac{F^3(t)\,[1 - F(t)]}{g(t)} dt.$$

Egy alkalmasan megválasztott integrálható $h$ függvénnyel

$$\sqrt{n} \int_0^\tau \left[\hat{F}_n^2(t) - F^2(t)\right] h(t) dt$$

aszimptotikusan normális ha $n \to \infty$, a várható értéke 0, a szórás pedig:

$$4 \int_0^\tau \frac{F^3(t)\,[1 - F(t)]}{g(t)} h^2(t) dt.$$

Ebből az eredményből a $\hat{G}_n$ tapasztalati eloszlásfüggvény egyenletesen erős konzisztenciájával következik, hogy

$$\sqrt{n} \int_0^\tau \left[\hat{F}_n^2(t) - F^2(t)\right] d\hat{G}_n(t)$$

aszimptotikusan normális 0 várható értékkel és

$$h \int_0^\tau F^3(t)\,[1 - F(t)]\, g(t) dt$$

szórással, ha $f$ és $g$ a $[0, \tau]$-n elválasztott a 0-tól, $f$ korlátos és $g$ deriváltja is az.

Nézzünk most egy kétmintás problémát a következő adatokkal:
$(t_{1i}, D_{1i}, i = 1, 2, \ldots, n_1)$, $(T_{2i}, D_{2i}, i = 1, 2, \ldots, n_2)$. Legyen

$$H_0: \; F_1 = F_2$$

azaz az a nullhipotézis, hogy a két minta élettartam eloszlása azonos. Legyen $\hat{F}_{1n_1}$ és $\hat{F}_{2n_2}$ az NPMLE értéke $F_1$-nek és $F_2$-nek a mintából számolva és legyen $\hat{G}_n$ a két minta megfigyelési idejeinek feltételezett közös eloszlásának a tapasztalati eloszlás függvénye, ahol $n = n_1 + n_2$. A $H_0$-ra vonatkozó teszt statisztika alapja:

$$\int_0^\tau \left[\hat{F}^2_{1n_1}(t) - \hat{F}^2_{2n_2}(t)\right] d\hat{G}_n(t)$$

ami $\sqrt{n}$-nel szorozva a korábbiak szerint aszimptotikusan normális eloszlású a $H_0$ mellett ha $n_1, n_2 \to \infty$ úgy, hogy $n_1/n_2 \to p \in (0,1)$. A várható érték 0 és a szórás

$$\frac{4}{p(1-p)} \int_0^\tau F^3(t)\left[1 - F(t)\right] g(t)dt.$$

Itt $\tau$ a megfigyelések felső határa, $F$ pedig a feltételezett közös értéke $F_1$-nek és $F_2$-nek. Legyen $\hat{F}_n$ a $F$ NPMLE-je a kombinált $n = n_1 + n_2$ elemű mintára. Ekkor

$$W = \frac{\int_0^\tau \left[\hat{F}^2_{1n_1}(t) - \hat{F}^2_{2n_2}(t)\right] d\hat{G}_n(t)}{\left\{\frac{4n^2}{n_1 n_2} \int_0^\tau \hat{F}^3_n(t)\left[1 - \hat{F}_n(t)\right] d\hat{G}_n(t)\right\}^{1/2}}$$

eloszlása megközelítőleg standard normális elég nagy $n_1$ és $n_2$ esetén. Ha $G$ értéke különböző a két mintában, azaz $G_1 \neq G_2$, akkor $W$ szórása expliciten függ $G_1$-től és $G_2$-től a megfelelő $g_1$ és $g_2$ sűrűségfüggvényeken keresztül, és ilyenkor az itt javasolt statisztika kevésbé megfelelő.

# 3. fejezet

# Alkalmazások

## 3.1. Az OD görbe

Tegyük fel, hogy adott két valószínűségi változó $X$ és $Y$, és $c \in \mathbb{R}$ tetszőleges konstans. Tekintsük azt a grafikont, melynek horizontális koordinátája $P(X \leq c)$, a vertikális koordinátája pedig $P(Y \leq c)$. Ábrázoljuk a $T(c)$ pontot $c$ minden ($c \in (-\infty, \infty)$) lehetséges értékére. Az így kapott grafikon az OD-grafikon egy típusa, a populáció OD-grafikon.

Tekinsük most a $T(c)$ pontot. Mivel mindkét koordinátája 0 és 1 között van, $T(c)$ mindig az egységnégyzetben lesz, és ha $X$ és $Y$ folytonos valószínűségi változók, akkor $T(c)$ a $(0,0)$-ból induló, $(1,1)$-be érkező folytonos görbe lesz. A $T$ mindkét változójában monoton növekedő függvény. Ha $X$ és $Y$ véges diszkrét valószínűségi változók, amelyek $k$ értéket vehetnek fel nem nulla valószínűséggel, akkor két további érték, a $-\infty$ és $\infty$ hozzávételével az OD-diagramnak $k+1$ különböző pontja lesz, melyekből kettő rögzített: $T(-\infty) = (0,0)$ és $T(\infty) = (1,1)$.

Az OD-grafikonok egyik igen hasznos tulajdonsága, hogy a rang megőrző transzformációkra invariánsak. Ez azt jelenti, hogy ha $m$ egy szigorúan növő függvény, amely $X$ és $Y$ minden lehetséges értékére definiált, akkor az $(X, Y)$ OD-grafikon ugyanaz lesz, mint az $(m(X), m(Y))$ OD-grafikon.

Az OD-grafikonnal összehasonlítható sztochasztikusan két valószínűségi változó pontosan akkor, ha a görbe nem keresztezi a $(0,0)$-t az $(1,1)$ ponttal összekötő egyenest. Ha az $(X, Y)$ OD-görbe teljesen a pozitív digonálison, vagy alatta van, akkor azt mondjuk, hogy $X$ sztochasztikusan kisebb, mint $Y$. Ha $X$ sztochasztikusan kisebb, mint $Y$, és $m$ monoton növő függvény, akkor $E(m(X)) \leq E(m(Y))$.

Legyen $X$ és $Y$ két valószínűségi változó, melyekből mintát veszünk. Legyen $W$ az a valószínűségi változó, amely a megfigyelések értékeit reprezentálja. Jelölje $S_X$ azt az eseményt, hogy a megfigyelést az $X$ eloszlásból vettük. Tegyük fel, hogy $P(S_X)$ előzetes valószínűség $1/2$, és a megfigyelés értéke $c$. Ha $X$ és $Y$ véges diszkrét

valószínűségi változó, akkor

$$P(S_X|\,W=c)=\frac{1}{1+P(Y=c)/P(X=c)}.$$

Másrészt, ha $X$ és $Y$ folytonos, akkor

$$P(S_X|\,W=c)=\frac{1}{1+f_Y(c)/f_X(c)},$$

ahol $f_X$ és $f_Y$ a megfelelő eloszlásfüggvények. Legyen $L(c)$ a likelihood arány:

$$L(c)=P(Y=c)/P(X=c),$$

illetve folytonos esetben

$$L(c)=f_Y(c)/f_X(c).$$

Így véges diszkrét, és folytonos esetre is:

$$P(S_X|\,W=c)=\frac{1}{1+L(c)}.$$

A két valószínűségi változó, $X$ és $Y$ hasznos tulajdonsága lehet, ha a poszteriori valószínűség, $P(S_X|\,W=c)$ a $c$ monoton függvénye. Ha $X$ és $Y$ ilyen tulajdonságú, akkor monoton poszteriornak nevezzük őket. Ez azt jelenti, hogy ha $X$ sztochasztikusan kisebb, mint $Y$, és a poszterior valószínűség $c$ monoton növő függvénye, akkor a megfigyelések legnagyobb $c$ értéke valószínűbben tartozik az $X$ eloszláshoz, mint az $Y$-hoz. Továbbá $X$ és $Y$ monoton poszterior pontosan akkor, ha az $(X,Y)$ OD-görbe konvex.

Tegyük fel, hogy független véletlen mintát veszünk az $X$ és $Y$ eloszlásokból. Jelölje $P(X<Y)$ annak az eseménynek a valószínűségét, hogy az $X$ megfigyelés kisebb, mint az $Y$. Hasonlóan definiálható $P(X\le Y)$ és $P(X=Y)$ is. Tekintsük az egységnégyzet és az OD-görbe által határolt felső területet, és jelölje ezt $A(X,Y)$. Ha $f_Y$ az $Y$ sűrűségfüggvénye, akkor:

$$\begin{aligned}A(X,Y)&=\int_0^1 P(X\le c)dP(Y\le c)=\\&=\int_{-\infty}^{\infty}P(X\le c)f_Y(c)dc=P(X\le Y).\end{aligned}\tag{3.1.1}$$

Természetesen ha $X$ és $Y$ folytonos, akkor $P(X=Y)=0$, amiből az is következik, hogy $P(X<Y)=P(X\le Y)=A(X,Y)$. Ha $X$ és $Y$ véges diszkrét valószínűségi változók, akkor $A(X,Y)$ a trapéz-szabállyal számolható:

$$\begin{aligned}A(X,Y)=\sum_{i=1}^{k}P(Y=c_i)P(X\le c_i)+\frac{1}{2}\sum_{i=1}^{k}P(Y=c_i)P(X=c_i)=\\=P(X<Y)+\frac{1}{2}P(X=Y),\end{aligned}\tag{3.1.2}$$

ahol $c_1, \ldots, c_k$ azon értékek halmaza, melyeket $X$ és $Y$ nemnulla valószínűséggel felvesz.

A fent kiszámolt $A(X,Y)$ jelöli azt a mértéket, amely azt méri, hogy mennyire van az $X$ az $Y$ alatt, így értéke 0 és 1 között mozog. A maximumát pontosan akkor veszi fel, ha az $X$ teljesen az $Y$ alatt van átfedés nélkül, és pontosan akkor lesz 0, ha teljesen felette van. Másrészt ha $X$ és $Y$ azonos eloszlású, akkor $A(X,Y) = 1/2$. Továbbá $A(X,Y) + A(Y,X) = 1$. Ezen tulajdonságok miatt $A(X,Y)$ nagyon hasznos két populáció közötti különbség mérésére, hiszen ha értéke valamelyik szélsőérték közelében van, az fontos különbséget jelent a populációk között. Más megközelítésben $A(X,Y)$ használható arra is, hogy milyen pontosan különbözteti meg egy adott teszt a két populációt. Orvosi teszteknél gyakran egy kritikus értéket keresnek arra, hogy diagnosztizáljanak egy adott betegséget. A már meglévő adatok alapján a betegek és az egészségesek alkotják a két populációt, ilyenkor ha $A(X,Y)$ közel van 1-hez, akkor majdnem biztos lesz a diagnózis, ha pedig 1/2, akkor véletlenszerű lesz a jó döntés.

Tegyük fel, hogy van egy $N_X$ méretű véletlen mintánk $X$-ből, és $N_Y$ méretű az $Y$-ból. Jelölje $p(X \leq c)$ azon megfigyelések arányát $N_X$-ből, melyek kisebbek vagy egyenlőek, mint a $c$ konstans. Definiáljuk $p(Y \leq c)$-t hasonlóan. Minden $c$-re $t(c)$ legyen az a pont, melynek első koordinátája $p(X \leq c)$, a második koordinátája pedig $p(Y \leq y)$. Az így kirajzolódó görbe a minta OD-görbe. Minden $c$-re a $t(c)$ koordinátái torzítatlan becslést adnak a $T(c)$ megfelelő koordinátáira, azonban ez a görbe csak véges sok pontot tartalmaz. Ha az egyesített mintában nincsenek egybeesések, akkor pontosan $N_X + N_Y + 1$ különböző pontot. Ha véletlen a mintavételezés, akkor $N_X N_Y$ különböző párosítása van a két mintának.

Jelölje $p(X < Y)$, $p(X = Y)$ és $p(X \neq Y)$ azon párok arányát ebből az $N_X N_Y$ párból, melyekre $X < Y$, $X = Y$ illetve $X \neq Y$. Nyilvánvaló, hogy ezek torzítatlanul becsülik $P(X < Y)$-t, $P(X = Y)$-t és $P(X \neq Y)$-t. Legyen $a(X,Y)$ a minta OD-görbe fölötti terület, amelyet trapéz szabállyal számolhatunk:

$$a(X,Y) = p(X < Y) + \frac{1}{2} p(X = Y).$$

Mindkét oldal várható értékét véve adódik, hogy $A(X,Y)$ torzítatlan becslése az $a(X,Y)$. Összességében tehát a minta OD-görbe fölötti terület torzítatlan becslése a populáció OD-görbe fölötti területnek.

A most definiált $a(X,Y)$ közeli kapcsolatban van a Mann-Whitney U-statisztikával, hiszen ha $X$ és $Y$ folytonos, akkor

$$a(X,Y) = \frac{U}{N_X N_Y}.$$

## 3.2. Az ROC és az ROL görbe

Tekintsünk most egy jelérzékelő kísérletet, ahol zajos környezetben alkalmi jelek érkeznek. A megfigyelő minden próbánál érzékel egy hangot, egy jelet vagy egy zajt, és az érzékelt hang erősségének függvényében dönt arról, hogy mit hallott. Azonban, mivel a hang erőssége próbáról próbára változik, ez általában nem megbízható alap, hogy a megfigyelő eldöntse, hogy jel érkezett vagy zaj. Tegyük fel, hogy a megfigyelő először megállapít egy kritikus értéket, azután, ha ennél erősebb hang érkezik, akkor igent mond, azaz jelnek tekinti, minden más esetben pedig nemet. Jelölje $P(igen|\, jel)$ illetve $P(igen|\, zaj)$ az igen válasz valószínűségét jel illetve zaj esetén, hasonlóan $P(nem|\, jel)$ és $P(nem|\, zaj)$ a nem válasz valószínűségét. Ezek a valószínűségek egy $2 \times 2$-es táblázatba foglalhatóak:

| valóság/tipp | jel | zaj |
|---|---|---|
| jel | $P(igen\|\, jel)$ | $P(nem\|\, jel)$ |
| zaj | $P(igen\|\, zaj)$ | $P(nem\|\, zaj)$ |

Ha a megfigyelő csökkenti a kritikus értéket, akkor az igen válaszok valószínűsége nőni fog. Ábrázoljuk a $P(igen|\, jel)$ és a $P(igen|\, zaj)$ valószínűségeket minden olyan kritikus értékre, amelyet a megfigyelő elfogad, oly módon, hogy a zajhoz tartozó kerül a horizontális tengelyre, a jelhez tartozó pedig a vertikálisra. Extrém alacsony kritérium esetén az ábrázolt pont közel lesz az $(1, 1)$ ponthoz, nagyon magasnál pedig a $(0, 0)$-hoz. Így egy folytonos görbét kapunk az egységnégyzetben, ami a $(0, 0)$-ból az $(1, 1)$-be ér. Ezt a görbét hívják Igen-Nem ROC-görbének.

Hasonlóan definiálható az $(I_n, I_s)$ ROC-görbe, ahol $I_s$ és $I_n$ az a két valószínűségi változó, melyek a jel illetve a zaj esemény által keltett hang erősségét jelölik. Legyen $c$ a megfigyelő kritériuma, ekkor:

$$P(igen|\, jel) = P(I_s \geq c)$$

és

$$P(igen|\, zaj) = P(I_n \geq c).$$

Így az Igen-Nem ROC-görbe képe pontosan ugyanaz, mint az $(I_n, I_s)$ ROC-görbe, és mindkettő az OD-görbék egy fajtája. Továbbá az Igen-Nem ROC-görbe alatti terült megegyezik az $(I_n, I_s)$ ROC-görbe alatti, és az $(I_n, I_s)$ OD-görbe feletti területtel. Így az ROC-görbe alatti terület is $A(I_n, I_s)$. Ez azt is jelenti, hogy ha ez az érték 1, akkor az $I_n$ eloszlás teljesen az $I_s$ alatt van, így a megfigyelő hibátlanul fogja megkülönböztetni a zajtól a jelet. Természetesen azonban, mivel ez egy elméleti konstrukció, az $(I_n, I_s)$ ROC-görbét a gyakorlatban nem tudjuk felrajzolni, csak véges sok pontját.

Egy kicsit más megközelítésben ugyanez: előrejelző rendszerek minőségét szeretnénk leírni. A ROC-görbe alatti terület oly módon jellemzi az előrejelző rendszer minőségét, hogy leírja a rendszer arra vonatkozó képességét, hogy előre jelezze az előre meghatározott események bekövetkeztét vagy nem bekövetkeztét.

A ROC-görbe megalkotásakor az előrejelzéseket bináris módon fejezzük ki, mint figyelmeztetések vagy nem figyelmeztetések, amelyek azt jelzik, hogy az előre meghatározott esemény várhatóan bekövetkezik-e vagy sem. Amikor az előrejelzést egy esemény bekövetkezési valószínűségében fejezik ki, a kiadott figyelmeztetés valószínűsége különböző küszöbértékek között mozog. Minden küszöbértéknél megvizsgálják az előrejelzések és a megfigyelések közötti kapcsolatot. Ez a kapcsolat leírható egy két dimenziós vektorral, amelyet a találati arány, azaz a helyesen előrejelzett események aránya, és a hibás riasztási arány definiál. Ez a két koordináta megad egy pontot az ROC-térben, és együttesen definiálják az ROC-görbét.

Egy kapcsolódó eljárás az ROL, melynél ideális esetben az események intenzitása folyamatos skálán mozog, és a riasztásokat rögzített kritériumok alapján adják ki. Az ROC-görbéhez hasonlóan az előrejelzések bináris állítások arra vonatkozóan, hogy egy előre meghatározott esemény várhatóan megtörténik-e vagy sem. A megfigyeléseket is bináris módon fejezzük ki, azaz egy adott megfigyelés egy esemény, ha a megfigyelés meghalad egy előre meghatározott küszöbértéket. A ROL-görbe esetén az az intenzitás, amely megkülönbözteti az eseményeket a nem eseményektől, egy korlátos skálán belül mozog.

Hasonlóan az ROC-görbéhez, az előrejelzések és a megfigyelések közötti kapcsolatot minden küszöbértékre megvizsgálják. Ismét egy két dimenziós vektor definiálja ezt a kapcsolatot, melynek első koordinátája a helyes riasztás arány, azaz azon esetek aránya, amikor egy riasztást követően megtörtént az esemény, a második koordinátája pedig a tévesztés arány, azaz azon esetek aránya, amikor a nem riasztást követően megtörtént az esemény. Ezek az arányok a ROL-térben adnak meg koordinátákat, és együttesen definiálják az ROL-görbét. Az ROL-görbe tehát egy előrejelző rendszer szakértelmének mértékét adja meg, amellyel a rendszer meg tudja jósolni, hogy az előrejelzés tárgya meghalad egy előre definiált küszöbérték sort. Az ROL-görbe alatti terület meghatározza annak a valószínűségét, hogy az esemény intenzívebb abban az esetben, amikor kiadtak riasztást ahhoz képest, amikor riasztás nem történt.

Az ROC-terület és a Wilcoxon statisztika kapcsolatának vizsgálatához bevezetjük a következő jelöléseket: legyen $n$ az összes esetek száma, $e$ az előre meghatározott események száma, $e' = n - e$ pedig a nem események száma. Soroljuk az előrejelzéseket csökkenő sorrendbe, azaz előre kerülnek azok, melyeknek a valószínűsége a legnagyobb. Tegyük fel továbbá, hogy nincsenek egybeesések. Nevezzük most találatnak azokat az eseteket, amikor egy eseményre adtak ki előrejelzést, és az meg is történt, hamis riasztásnak pedig azokat, amikor egy esemény a rá kiadott riasztás ellenére sem történt meg. Az ROC-görbe alatti terület minden találattal, jelölje $f$, nő, a nyert terület pedig

$$\frac{e' - f}{e'e},$$

amiből az $A$-val jelölt ROC-terület:

$$A = \frac{1}{e'e} \sum_{i=1}^{e} (e' - f_i),$$

ami egyszerűbben

$$A = 1 - \frac{1}{e'e} \sum_{i=1}^{e} f_i.$$

Mivel az események és a nem események száma konstans, a terület egyszerűen azon hamis riasztások összegének függvénye, melyeknek minden találatnál magasabb a valószínűsége. Jelölje $r_i$ a találatoknak megfelelő előrejelzések rangját, így

$$F = \sum_{i=1}^{e} f_i = \sum_{i=1}^{e} (r_i - i) = \sum_{i=1}^{e} r_i - \frac{e(e+1)}{2}.$$

Ha minden találatnak nagyobb a valószínűsége, mint a hibás riasztásoknak, akkor $F = 0$, és így az ROC-terület, $A = 1$.

Egy kicsit más megközelítésben először tegyük külön listába a valószínűségük sorrendjében az eseményeket és a nem eseményeket, aztán fűzzük össze a két listát. Így $F$-et kiszámíthatjuk, mint az így kialakult lista inverziószáma, ami már mutatja a kapcsolatot a Wilcoxon-statisztikával, $U$-val:

$$F = U = e'e(1 - A).$$

Természetesen így az ROC-területre is igazak a korábbi eredmények, $F$ eloszlása is normális, a várható értéke $e'e/2$, a szórása pedig $e'e(n+1)/12$.

Az ROL-görbe esetén teljesen analóg az eljárás, jelölje $w$ a riasztások számát, $w'$ a nem riasztások számát, az eseményeket rangsoroljuk az intenzitásuk szerint csökkenő sorrendbe, és jelölje $B$ az ROL-görbe alatti területet.

$$B = 1 - \frac{1}{w'w} \sum_{i=1}^{w} m_i,$$

tehát a terület azon nem riasztások összegének függvénye, melyek intenzívebb eseményekkel párosulnak, mint a riasztások. Az ROC-területhez hasonlóan adódik:

$$M = \sum_{i=1}^{w} m_i = \sum_{i=1}^{w} q_i - \frac{w(w+1)}{2},$$

továbbá

$$M = U = w'w(1 - B),$$

azaz $M$ is ekvivalens $U$-val.

Azonban a ROC-, és ROL-görbék ilyen alkalmazásánál figyelembe kell venni a Wilcoxon-statisztika feltételeit, amelyek közül a legfontosabb a két minta függetlensége, amely ebben a kontextusban olyan feltételre fordítható, amely szerint az előrejelzések és a megfigyelések szekvenciálisan és térben függetlenek. Ez például meteorológiai előrejelzéseknél megnehezíti az alkalmazást.

## 3.3. Eltolás detektálás idősorokban

Fontos információt jelentenek az idősorokban a hirtelen szinteltolások, amelyeket éleknek vagy ugrásoknak hívunk. A közvetlen analízishez szükség van megbízható szabályokra a szintugrások detektálásához. Alapvető elvárás az eltolások megkülönböztetése a kisebb fluktuációktól és a nem releváns kiugrások rövid sorozatától. Ezt a feladatot úgy formalizáljuk, hogy az $Y_t$ ($t \in \mathbb{Z}$) megfigyeléseket egyszerű additív komponensekre bontjuk:

$$Y_t = \mu_t + v_t + u_t \quad t \in \mathbb{Z},$$

ahol $\mu_t$ az idősor időfüggő szintje, amiről feltesszük, hogy a megváltozása sima néhány hirtelen ugrással, az $u_t$ egy zaj 0 várható értékkel és az esetleg időfüggő $\sigma_t^2$ szórással. A $v_t$ egy véletlen tényező, ami a kiugró értékeket generálja. Ez többnyire 0, de alkalmanként nagy abszolut értékű. Az így létrejövő kiugró értéket outlier-nek nevezzük.

A $\mu_t$ becslésével sok szűrő eljárás elérhető. Lineáris szűrők, mint a mozgó átlag, hatékonyak Gauss-zaj esetében, de érzékenyek a kiugró értékekre. Futó mediánok segítségével is becsülhető $\mu_t$ szintje mozgó ablakokkal $(y_{t-k}, \dots y_{t+k})$, így a mediánnal $\tilde{\mu}_t = med(y_{t-k}, \dots y_{t+k})$, ahol $t \in \mathbb{Z}$. Az ilyen eljárások előnye, hogy eltávolítják az outlier-eket, és jobban megőrzik az ugrásokat.

Időnként azonban a szintugrások megőrzése a vizualizációra nem elég és szükséges, hogy a szintugrások automatikusan érzékelhetőek legyenek. A robosztus kontroll diagramokhoz szükség van a kontrollálatlan folyamatra vonatkozó erős feltételekre és egy állandósult állapot létezésére, illetve ezek reagálnak a struktúrális változások jónéhány típusára. Elvárás az észlelések átlagos késésének minimalizálása, amíg időnként a pontos eltolás nem is számít, ha az túl nagy.

Vannak azonban olyan detektálási szabályok, amelyeket specifikusan arra fejlesztettek ki, hogy egy adott időtartamon belül érzékeljenek szintugrásokat és csak gyenge feltételeket igényelnek. A kétmintás rang tesztek ígéretes jelöltek erre az egyszerűségük miatt is. Az adatok rangjait egy $k$ szélességű $(y_{t+1}, \dots, y_{t+k})$ mozgó ablakban, és egy hosszabb ablakban határozzák meg, ami tartalmaz $h$ korábbi megfigyelést $(y_{t-h+1}, \dots, y_t)$, balra a $t-$től. Egy emelkedő illetve csökkenő eltolás $t$ és $t+1$ között detektálódik, ha az $y_{t+1}, \dots, y_{t+k}$ rangok vagy a megfelelő transzformáltjaik elég nagyok illetve kicsik. Az eltolás rövid késéssel való detektálására idősorokban rang teszteket vizsgálunk, és úgy módosítjuk őket, hogy megkülönböztessék a bizonyos hosszúságú kilógó adat sorozatokat a hosszútávú eltolásoktól.

Az ugrás detektálásra vonatkozó szabályok formalizálásához és összehasonlításához tegyünk fel egy $\delta$ nagyságú ideális élet a $t \in \mathbb{Z}$ időpont után:

$$\mu_{t+j} = \begin{cases} \mu & \text{ha } j \leq 0 \\ \mu + \delta & \text{ha } j > 0. \end{cases}$$

Egy, a $t$ időben történő pozitív, illetve negatív eltolás detektálásához a $H_0 : \delta = 0$ nullhipotézist teszteljük a $H_1^+ : \delta > 0$ illetve $H_1^- : \delta < 0$ ellenhipotézissel szemben. A következőkben mindössze egy időpontot vizsgálunk az $n = h + k$ darab

$y_{t-h+1}, \ldots, y_t, y_{t+1}, \ldots, y_{t+k}$ megfigyelést tekintve a $\tilde{\mu}_t$ mediánnal, ahol $h \geq k$. A $h$ és $k$ kiválasztásához használt elveket később adjuk meg.

Lineáris rang statisztikákon alapuló teszteket gyakran használnak él detektálásra, különösen népszerű a Wilcoxon és a medián-teszt. Legyenek $y_{t(1)} \leq \ldots \leq y_{t(n)}$ a rangorolt megfigyelések az ablakon belül, és jelölje $r_{-h+1}, \ldots, r_k$ az $y_{t-h+1}, \ldots, y_{t+k}$ rangjait ebben a sorozatban. A legutóbbi $k$ megfigyelés egy általános lineáris rang statisztikája a következő módon számolható:

$$S_+ = \sum_{j=1}^{k} a(r_j),$$

az $a(1), \ldots, a(n)$ értékekkel. Az $S_+$ kiegészítőjét jelölje

$$S_- = \sum_{i=1}^{n-1} a(r_{-i}) = \sum_{i=1}^{n} a(i) - S_+.$$

Így a lineáris rang statisztika:

$$L = \frac{(h+k)\left[h(S_- - \bar{a})^2 + k(S_+ - \bar{a})^2\right]}{\sum_{i=1}^{h+k}(a(i) - \bar{a})^2},$$

ahol $\bar{a} = \frac{1}{n}\sum_{j=1}^{n} a(j)$, ami eloszlásmentes, és aszimptotikusan $\chi_1^2$ eloszlású $H_0$ alatt konstans szórás esetén.

A Wilcoxon teszt esetén $a(i) = i$, $i = 1, \ldots, n$, azaz $S_+ = \sum_{i=1}^{k} r_i$. A normalizált Wilcoxon statisztika pedig

$$W = S_+ - \frac{k(k+1)}{2},$$

mely az értékeit 0 és $\frac{k(3k+1)}{2}$ között veszi fel, ha $h = k$. A Wilcoxon értékek olyan becslésekhez és tesztekhez vezetnek, amelyek majdnem olyan hatásosak Gauss-zaj esetén, mint az átlagokon alapuló módszerek. A medián teszt esetén

$$a_i = \begin{cases} 1 & \text{ha } i = \left\lfloor \frac{n}{2} \right\rfloor, \ldots, n \\ 0 & \text{egyébként.} \end{cases}$$

Ekkor $S_+$ azon értékek száma $y_{t+1}, \ldots, y_{t+k}$ között, amelyek nagyobbak mint a teljes ablak mediánja, és értékei 0 és $\left\lfloor \frac{n+1}{2} \right\rfloor$ között vannak. A medián teszteket még nehézfarkú eloszlások esetén is megbízhatónak tartják. A medián teszt könnyen alkalmazható, hiszen véges minták esetén a két csoport mediánját egyszerűen meghatározhatjuk. Ha a két vizsgált csoport között nincs különbség, azaz $H_0$ teljesül, akkor a közös medián alatt és felett nagyjából hasonló arányban oszlanak meg a megfigyelések. Az adatokat kontingencia táblába foglalva gyakorlatilag a $\chi^2$ próbához jutunk.

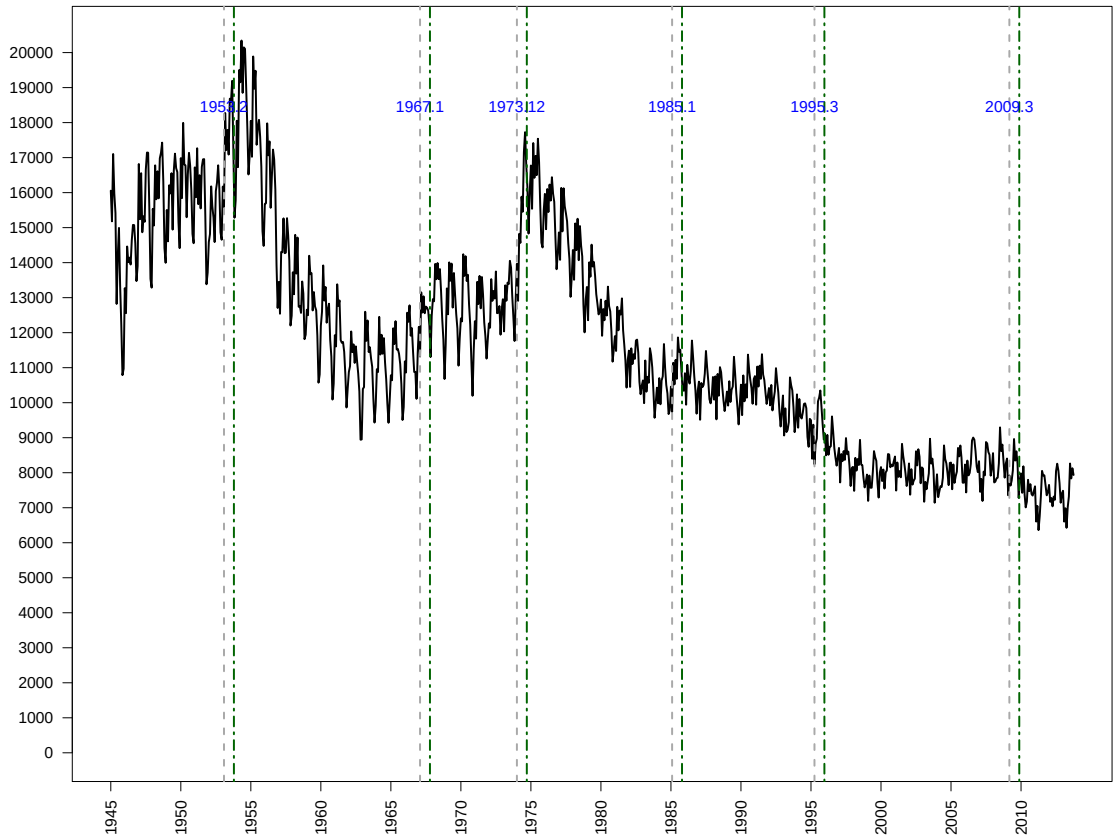
Ahhoz, hogy csak a releváns nagyságú emelkedő, illetve csökkenő eltolást érzékeljük, vonjunk ki, illetve adjunk hozzá egy $\delta_0$ küszöbértéket az $y_{t+1}, \ldots, y_{t+k}$ értékekből, mielőtt a rang tesztet alkalmaznánk. A $\delta_0$ küszöb értékét válasszuk

úgy, hogy nagyobb legyen, mint a zaj standard szórása. Ha $\sigma_t$ időfüggő, akkor a küszöbnek is időfüggőnek kell lennie, azaz legyen $\delta_0 = \delta_0(t)$, hogy skála-ekvivariáns eljárást kapjunk. Feltéve, hogy a standard szórás majdnem konstans a bal ablakban amely $y_{t-h+1}, \ldots, y_t$, kiszámítjuk a $\sigma_t$ robosztus becslését $\hat{\sigma}_t$-t, és a $\delta_0$ küszöböt úgy választjuk, mint ennek egy konstansszorosát, legyen ez $d\hat{\sigma}_t$. Nem vesszük bele azonban $\sigma_t$ becslésébe az $y_{t+1}, \ldots, y_{t+k}$-t, hogy elkerüljük a $t$-beli eltolás $\sigma_t$ torzított becslése miatti álcázását.

# 4. fejezet

# A Wilcoxon statisztika működése egy idősoron

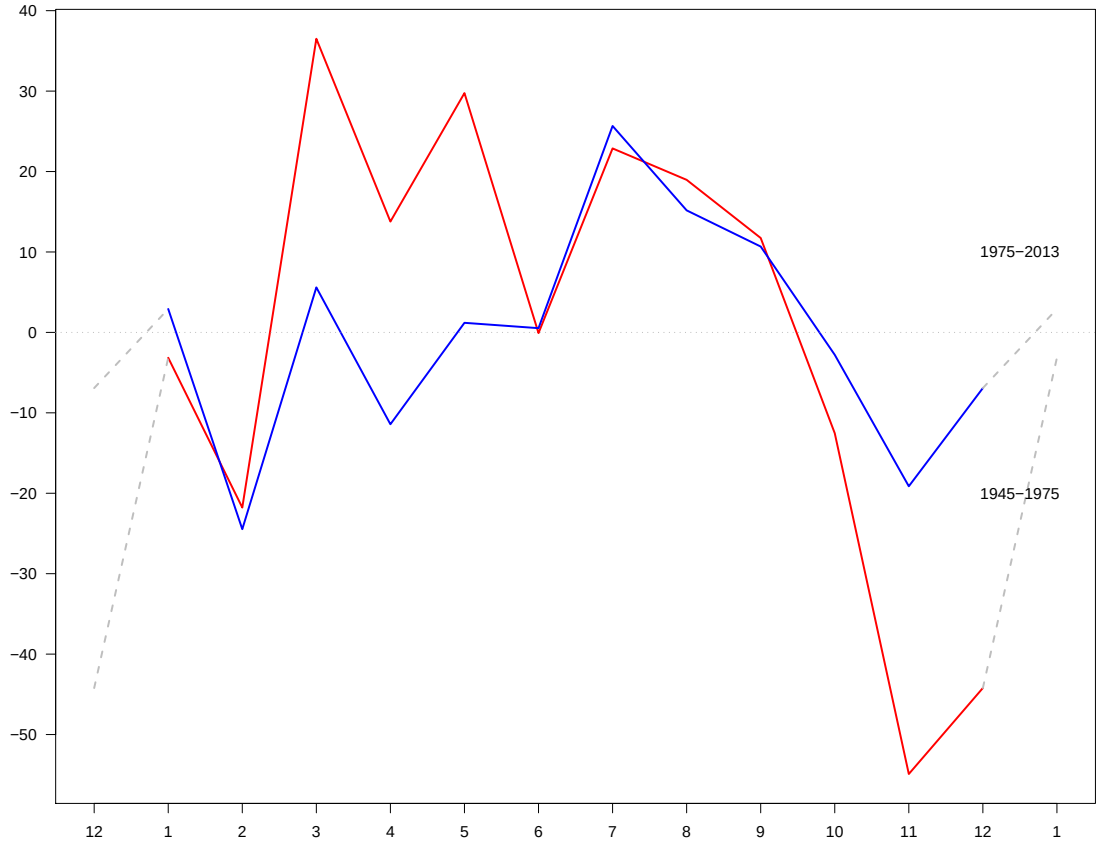
Havonkénti magyarországi születésszámokat nézzük 1945. és 2013. között. A nyers megfigyelési adatábra a következő:




Azt vizsgáljuk, hogy hogyan viselkednek a születésszámok az ezen időszak alatti hat ismert, jelentős demográfiai hatású döntés idején. A hat intézkedést jelöltük az

ábrán függőleges vonallal, illetve mindegyikre azt is, amikor a hatásuk jelentkezhetett. Ezen időszakokra vonatkozó részletesebb adatábrákat a függelék külön-külön tartalmazza.

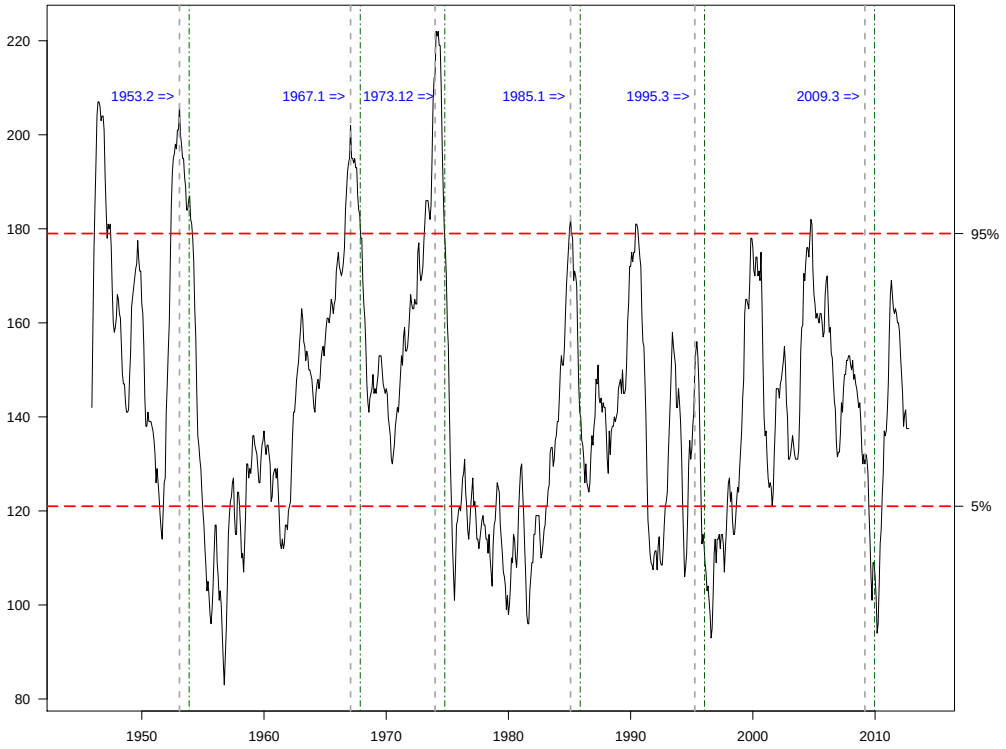
A következő ábrán az éves periódus profil látható 1975. előtt, illetve 1975. után:



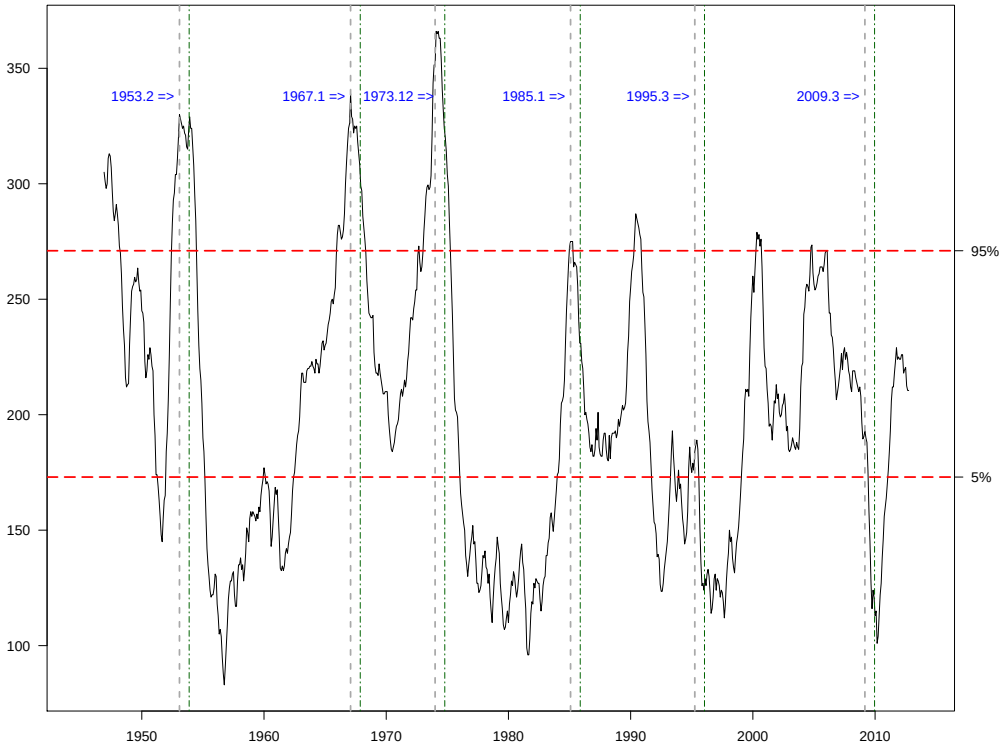

Az 1945-2013 közötti havi nyers adatok és az éves periódus profil ábrái alapján az látható, hogy a megfigyelési időszak alatt egyrészt a születésszám szintje jelentősen megváltozott az 1950-es évek közepén közel 20000 születéshez képest, a 2010-es években volt olyan hónap, amikor ez a szám a 8000-et sem érte el.

Másrészt az is látható, hogy a születésszám periódus profilja is nagy mértékben megváltozott, csökkent az amplitudója, a korábbi markáns csúcsokból és gödrökből gyakorlatilag a februári csökkenés és a koranyári növekedés maradt csak meg.

Ezekre az adatokra használjuk a Wilcoxon-statisztikát, az előző fejezetben bemutatott ablak-módszer segítségével. Az ablak méreteire az éves periodicitás miatt a 12 egészszámú többszörösei tűnnek jó választásnak. Ezért nézzük először a $h = k = 12$, majd a $h = 24$ illetve $k = 12$ választással. Mindkét esetben a konkrét értékeket táblázat helyett szimuláció állítja elő. Az ábrákon vízszintes vonal jelöli a statisztika 5-százalékos alsó, illetve felső kvantilisét. Ezen értékek az első esetben 121 és 179, míg a hosszabb múltra tekintő statisztikánál 173 és 271. Az ablak méretének megváltoztatása azonban nem okoz markáns különbséget, a Wilcoxon-statisztika szerint a születésszám négyszer került szignifikánsan növekedő, és ötször szignifikánsan csökkenő pályára.




4.1. ábra. h=k=12



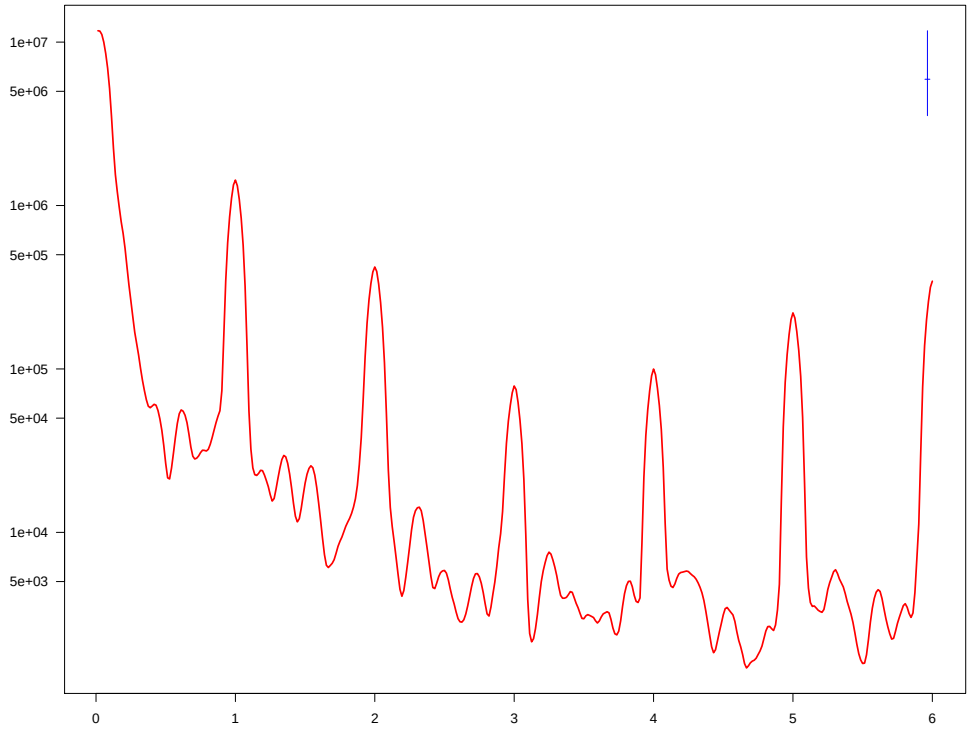

4.2. ábra. h=24,k=12

Az első növekedési periódus az 1940-es évek végén látható, ennek lehetséges oka a háború utáni konszolidáció. A második az 1950-es évek első felében volt a születés szabályozási intézkedések után. A következő növekedési szakasz az 1960-as évek végén látható, köszönhetően a hosszabb idejű, fix összegű havi juttatásoknak, melyek emelése okozott még egy ilyen periódust az 1970-es évek elején. A vizsgált időszakban volt egy ötödik pozitív beavatkozás, de ennek szignifikáns emelő hatása nem volt. A nyers adatokat bemutató ábra tanúsága szerint esetleg annyi, hogy megállította a csökkenést, de a Wilcoxon statisztika inkább azt mutatja, hogy a csökkenés hamarabb megállt.

A csökkenési periódusok közül az első az 1950-es évek második felében látható. A második az 1980-as évek körül volt, a fix összegű havi juttatás hatásának elmúlása és a jövedelem arányos juttatás bevezetése között. A következő csökkenés az 1990-es évek mésodik negyedében látható, nem sokkal a jelentős politikai változások után. Ezt az 1990-es évek harmadik negyedében követte egy csökkenés, miután megváltoztak a juttatásokra való jogosultságok. Az utolsó csökkenési periódus 2010 előtt látható.

Azaz a Wilcoxon statisztika eredménye a várakozásoknak nagyjából megfelel.

Most nézzük a fenti idősor spektrumát.




4.3. ábra. Spektrum

A spektrum azt mutatja, hogy a születésszám az erős trend ellenére határozott 12 havi periodicitást mutat. A trend jelenlétét a 0 körüli magas spektrum értékek mutatják, a 12 havi ciklicitást pedig az 1 pontbeli csúcs. A spektrum értelmezési tartománya a $[0, 6]$, mert azt adtuk meg, hogy a periódushossz 12 megfigyelés. Ennek felel meg, hogy az értelmezési tartomány jobboldali vége $6 = 12/2$-vel cimkézve. A $2, 3, 4, 5, 6$ pontokban azért jelentkeznek további spektrum csúcsok mert a 12 havi ciklicitás a mintavételi időköz pontos többszöröse. A csúcsok magassága pedig a profiltól függ. A spektrum a nyers periodogramból egy 12 megfigyelés széles simítással készült.

# Irodalomjegyzék


[1] H. B. Mann and D. R. Whitney: On a test of whether one of two random variables is stochastically larger than the other, Annals of Math. Stat., Vol.18 (1947)

[2] J. Hemelrijk: Note on Wilcoxon's two-sample test when ties are present, Annals of Math. Stat., Vol.23 (1952)

[3] Edmund A. Gehan: A generalized Wilcoxon test for comparing arbitrarily singly-censored smaples, Biometrika, Vol.52 (1965)

[4] Edmund A. Gehan: A generalized two-sample Wilcoxon test for doubly censored data, Biometrika, Vol.52 (1965)

[5] P. K. Andersen and B. B. Ronn: A nonparametric test for comparing two samples where all observations are either left- or right-censored, Biometrics, Vol.51 (1995)

[6] M. G. Kendall and A. Stuart: The advanced theory of statistics,2. London:Charles Griffin and Co. (1961)

[7] D. Bamber: The area above the ordinal dominance graph and the area below the receiver operating charasteristic graph, Journal of Mathematical Psychology, Vol.12 (1975)

[8] S. J. Mason and N. E. Graham: Areas beneath the relative operating charasteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation, Q.J.R.Meteorol.Soc., Vol.128 (2002)

[9] R. Fried and U. Gather: On rank tests for shift detection in time series, Computational Statistics and Data Analysis, Vol.52 (2007)

[10] Mári Tamás: Élettartam-adatok elemzése (2006)

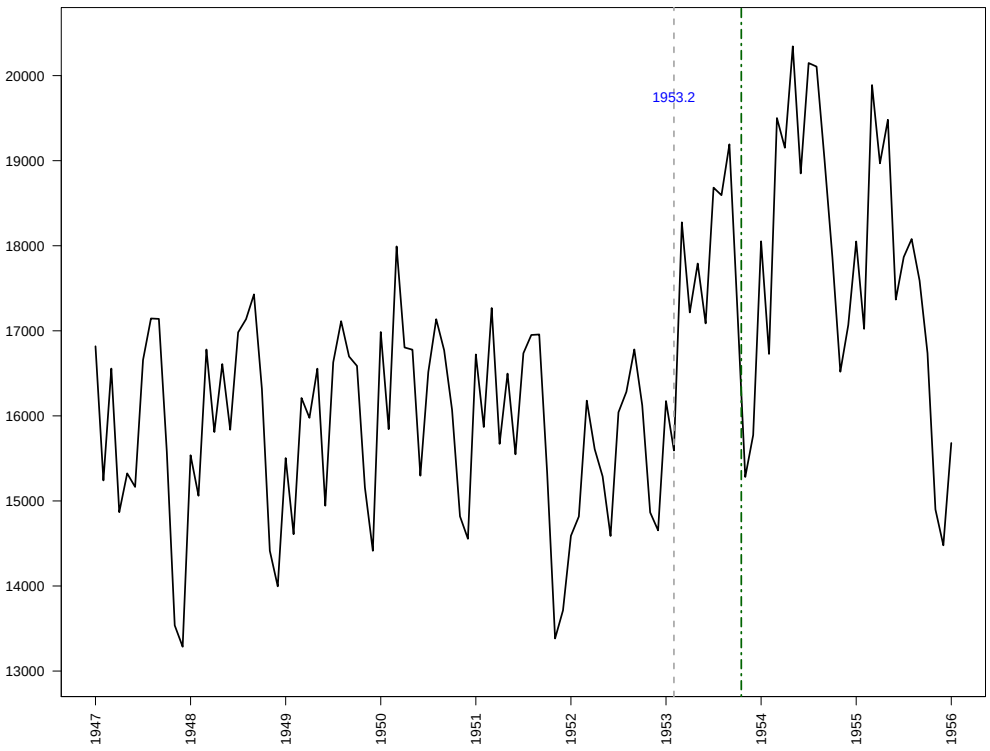
[11] W. H. Kruskal and W. A. Wallis: Use of ranks is one-criterion variance analysis, Journal of the American Statistical Association, Vol.47 (1952)

[12] W. H. Kruskal: Historical notes on the Wilcoxon unpaired two-sample test, Journal of the American Statistical Association, Vol.47 (1957)

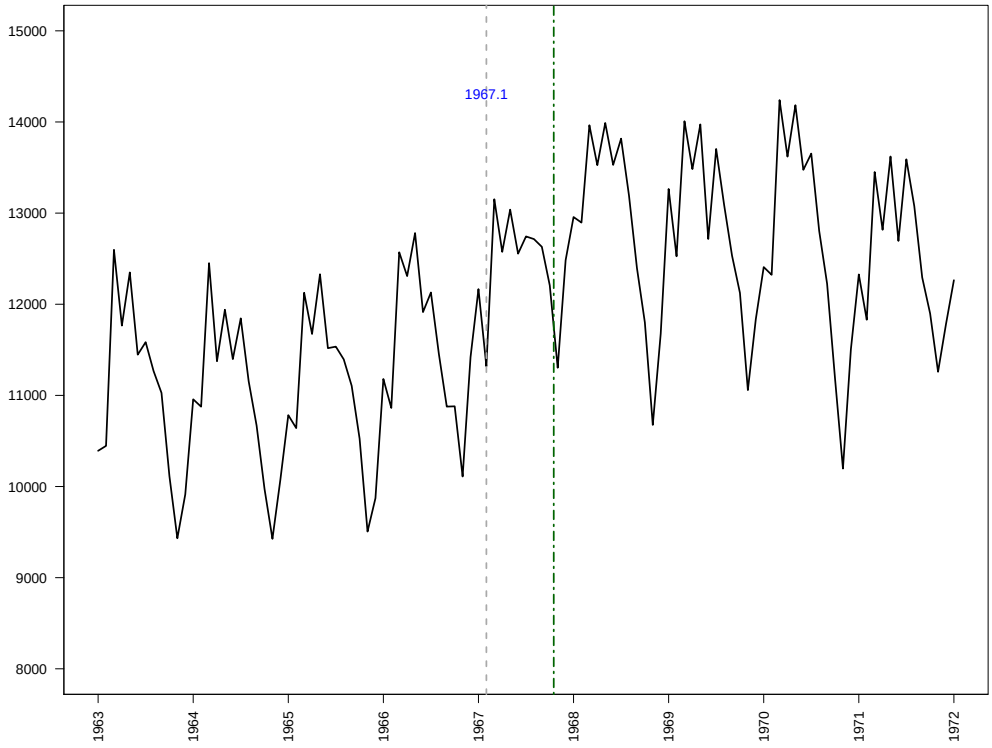
[13] KSH, Központi Statisztikai Hivatal, Népmozgalom, Gyorstájékoztató 2013/191. (Ez a felhasználtak közül a legutolsót jelöli, 1945 és 2013 között számos cím és forma változás volt.)

[14] Prőhle Tamás: Kézirat, 2013.

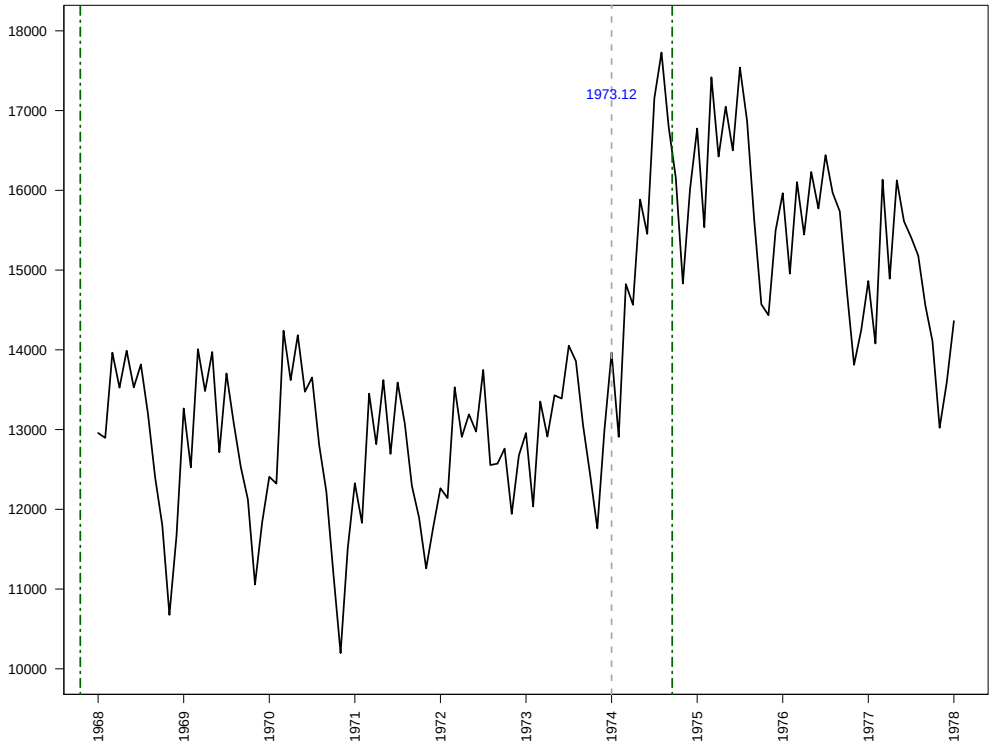
# Függelék



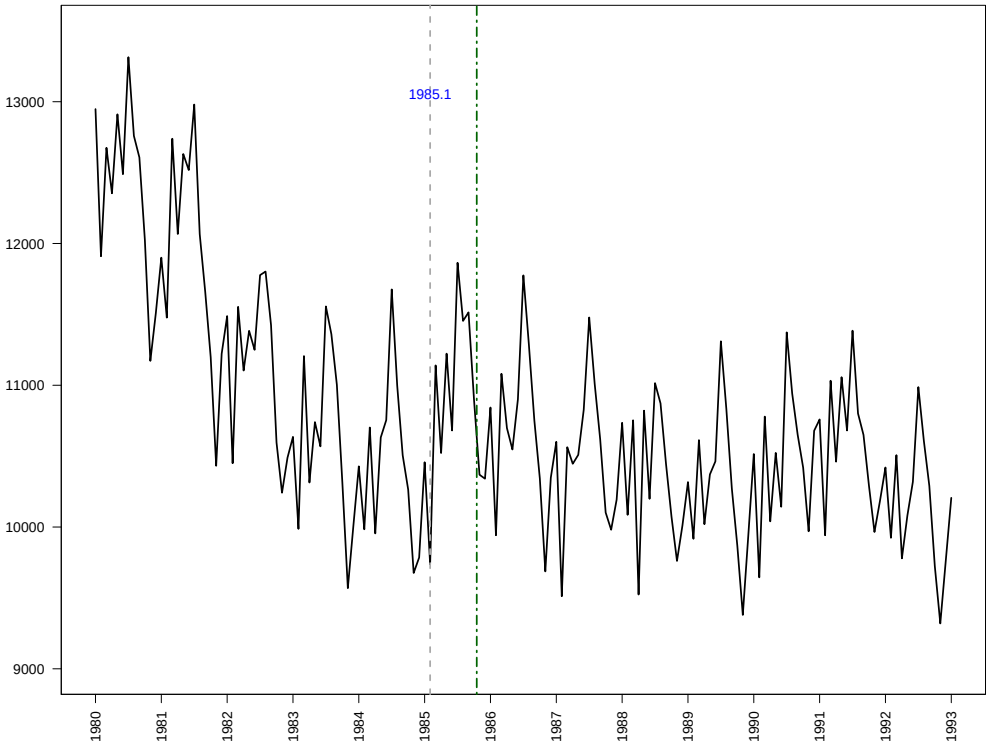

4.4. ábra. 1953



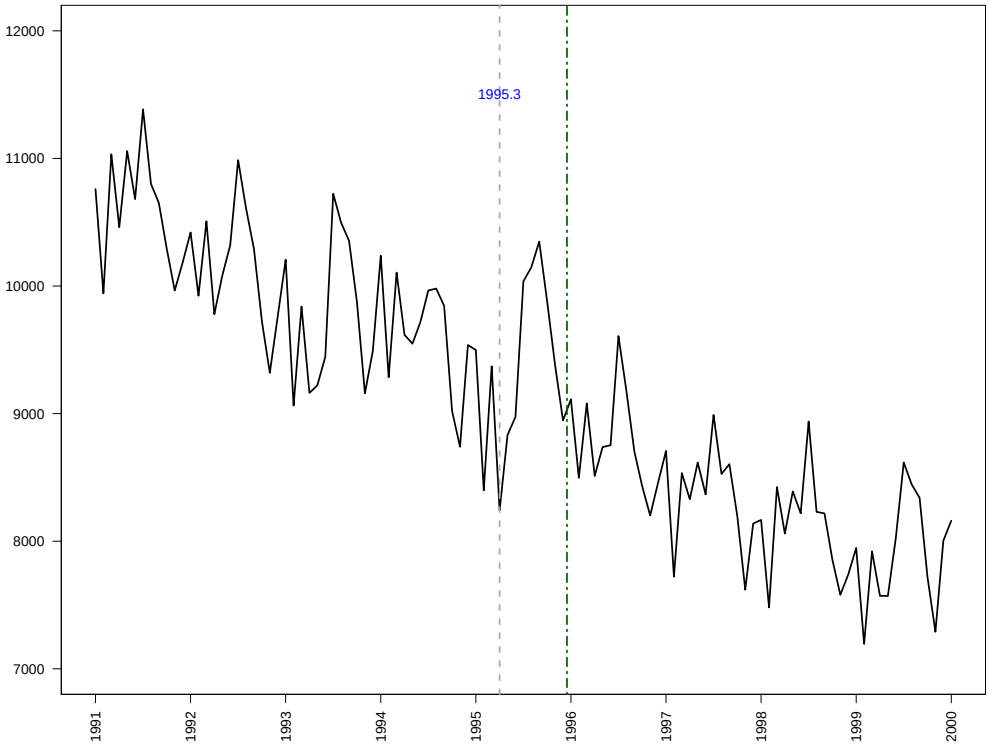

4.5. ábra. 1967



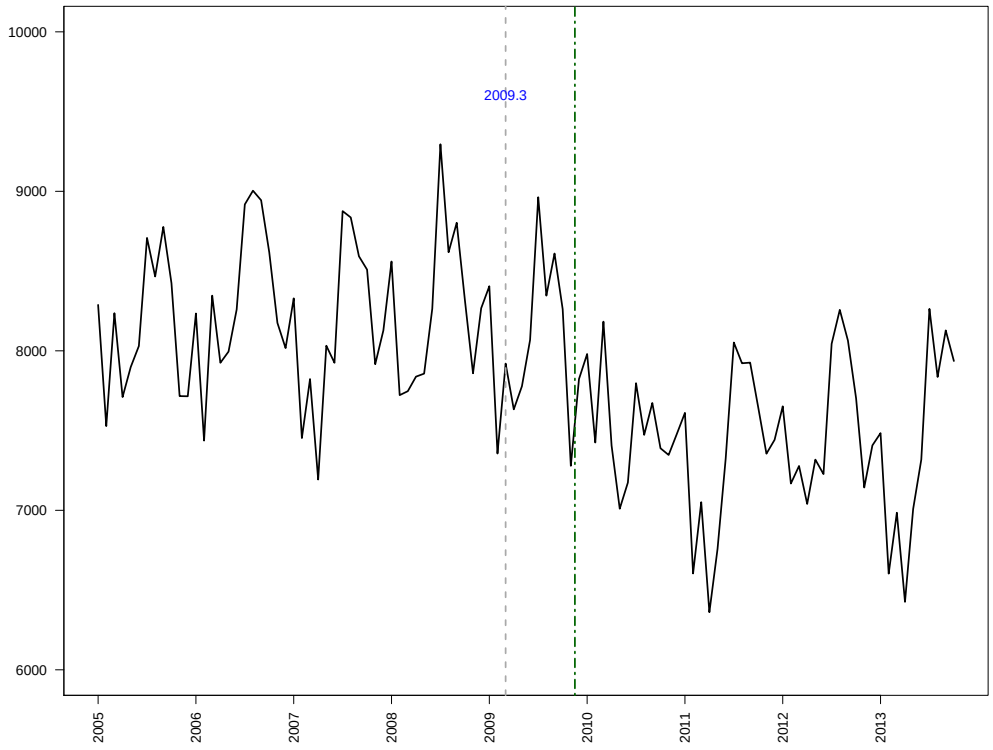

4.6. ábra. 1973




4.7. ábra. 1985




4.8. ábra. 1995


10000
9000
8000
7000
6000
2009.3
2005
2006
2007
2008
2009
2010
2011
2012
2013

4.9. ábra. 2009