Szóráselemezés

A szórások tulajdonságainak vizsgálata nem fordul elő az elsőéves fizikában. Ezt a fejezetet csak az ismeretek bővítése érdekében hoztuk létre. Hozzájárul ehhez az is, hogy a közreadott mintajegyzőkönyv tartalmaz erre való utalásokat.

Nem követjük a megszokott formát; nem az elején kezdjük, hanem mindjárt megmutatjuk a végeredményt. Ezzel ugyanis érthetővé válik a dolog hasznossága, és nagyobb figyelemmel olvassa a hallgató.

Van tehát egy mérésünk, amelynek minden eleme azonos; csak egy különbség van köztük: a kísérlet végző személy három eltérő működési elvű műszerrel is elvégezte a mérést. Lássuk tehát: ha van mérési bizonytalansága ezeknek a méréseknek, az abból származik-e, hogy más műszerrel mérték!

A feladatot két változatban dolgoztuk fel. A két változat között egyetlen különbség van: a mérési adatok egyikét gyanúsnak találtuk. Ezért végeztünk egy számítást úgy, hogy benne hagytuk, és egy másikat, amelynél kizártuk a feldolgozásból ugyanazt az adatot. A szóráselemzés bebizonyította, hogy ezt az adatot valóban ki kell zárni, mert kilóg a mérések közül, feltételezhetően műszerkezelési hiba következtében került az eredmények közé. Az eljárás hasonló ahhoz, amikor a rendellenes adatot kizárjuk a kiértékelésből (angolul **outlier rejection**).

A rendellenességet jelző cellákat piros szín jelzi az első változaton; a helyesbített változaton ugyanazok a a cellák zöld színűek.

Tekintsük először a mérési adatokat! Vajon mi jelzi, hogy valami nincs rendben? A 170 cm-es eredmény első következménye kilenc sorral lejjebb látható. Az ébreszti fel a gyanúnkat, hogy ennek a mérési sorozatnak szokatlanul nagy a szórása. Az átlagokon nem veszünk észre semmit. A másik két műszerrel mért adatok szórásnégyzete viszont kiugróan eltér ettől.

	sorszám	mérőrúd	mérőszala	ultrahangos távolságmérő		
	1	182	177	183		
	2	179	179	183		
	3	170	179	183		
	4	181	184	177		
	5	177	176	178		
	6	186	183	183		
	7	183	178	179		
	8	184	183	180		
	9	184	184	180		
	10	183	182	182		
átlag		180,9	180,5	180,8		
szórásnégyzet		21,4333	9,16667	5,28889		
szórás (egyszerű)		4,62961	3,02765	2,29976	2,18174	
összátlag	180,733	0,27778	0,54444	0,04444	0,43333	
sB^2		12,4231		F-próba	0,03488	
between classes		szabdsági fok 29-3 (29-1 adat és 3-1 oszlop)				

Egytényezős varianciaanalízis

ÖSSZESÍTÉS

Csoportok	*Darabszám*	*Összeg*	*Átlag*	*Variancia*
mérőrúd	10	1809	180,9	21,4333
mérőszala	10	1805	180,5	9,16667
ultrahang	10	1808	180,8	5,28889

VARIANCIAANALÍZIS

Tényezők	*SS*	*df*	*MS*	*F*	*p-érték*	*F krit.*
Csoportol	0,86667	2	0,43333	0,03622	0,96447	3,35413
Csoportor	323	27	11,963			
Összesen	323,867	29				

SS->MS 0,43334 within classes

SS->MS 11,963 323,667 az A18-b18-tól eltérések négyzetösszeg

Megtekintve a varianciaanalízis táblázatát, az SS, az MS (**mean square**) értéke gyanúra ad okot, és az F-próba is rosszabb a kritikus értéknél. Az első sorban a csoportokon belüli, az alatta levő sorban a véletlen által okozott (a csoportokon kívüli) négyzetes eltérést látjuk, legalul az összes négyzetes

eltérést. A szóráselemzés azon alapul, hogy a négyzetes eltérések *teljes* összegét két részre bontjuk, ez összege annak, amelyet a *kezelések* okoznak, valamint annak, amely a kezelések következményeként *nem magyarázható* meg. Ez utóbbi a véletlen szóródás, szokás hibának is nevezni. Angolul a négyzetes eltérések neve ugyanilyen sorrendben: **total**, **between** és **error** (más elnevezéssel: between classes és within classes). Képlettel: SST=SSB+SSE. A fenti táblázatban a szórásfelbontás numerikusan a következő: 323,867=0,8667+323. A szórásnégyzet tehát csaknem egészen a hibából származik, és nem a kezelésekből. Nosza, hagyjuk ki a számunkra gyanús adatot, és ismételjük meg a számítást:

	sorszám	mérőrúd	mérőszala	ultrahangos távolságmérő			Egytényezős varianciaanalízis						
	1	182	177	183			ÖSSZESÍTÉS						
	2	179	179	183			*Csoportok*	*Darabszám*	*Összeg*	*Átlag*	*Variancia*		
	3		179	183			mérőrúd	9	1639	182,111	7,61111		
	4	181	184	177			mérőszala	10	1805	180,5	9,16667		
	5	177	176	178			ultrahang	10	1808	180,8	5,28889		
	6	186	183	183									
	7	183	178	179									
	8	184	183	180			VARIANCIAANALÍZIS						
	9	184	184	180			*Tényezők*	*SS*	*df*	*MS*	*F*	*p-érték*	*F krit.*
	10	183	182	182			Csoportol	13,7008	2	6,85038	0,93257	0,40631	3,36902
átlag		182,111	180,5	180,8			Csoportor	190,989	26	7,34573			
szórásnégyzet		7,61111	9,16667	5,28889									
szórás (egyszerű)		2,75882	3,02765	2,29976	2,18174		Összesen	204,69	28				
összátlag	181,103	9,13846	3,6415	0,92081	6,85038								
sB^2		7,34573		F-próba	0,93257		SS->MS	6,85035	within classes				
between classes		szabdsági fok 29-3 (29-1 adat és 3-1 oszlop)					SS->MS	7,34573	204,69	az A18-b18-tól eltérések négyzetösszeg			

Íme, helyreállt a rend! A három műszerrel mért szórásnégyzetek azonos nagyságrendbe kerültek (balról a harmadik oszlop: 7,61111).

Kérdés, hogyan választhatjuk ki azt az adatot, amely az adatösszesség halmazából *kívül esőnek* (outlier) tekinthető. Erre több módszer is létezik. Javasolható például a mediántól legtávolabbi (gyanúsan kicsi, vagy gyanúsan nagy) adatot kritikailag megvizsgálni.

Például a fenti munkalap esetén ezt érdemes kipróbálni:

=ABS(MEDIÁN(C5:E14)-MIN(C5:E14))= 12 > =ABS(MEDIÁN(C5:E14)-MAX(C5:E14))=4
tehát a táblázat legkisebb eleme távolabb esik a közepétől, mint a legnagyobb eleme. Ha töröljük a gyanúsnak vélt adatot (a 170 cm-t), sokkal kiegyenlítettebb adathalmazhoz jutunk:

=ABS(MEDIÁN(C5:E14)-MIN(C5:E14))= 6 > =ABS(MEDIÁN(C5:E14)-MAX(C5:E14))=4

Most, hogy már ismerjük a dolog fontosságát, tekintsük át a feladat végrehajtásának menetét!

Az EXCEL táblázatra való utalásainknál az adatokat a három kezeléshez igazodva három oszlopba rendeztük, mégpedig C5:C14 (mérőrúd) D5:D14 (mérőszalag) és E5:E14 (ultrahangos műszer). A harminc mérésből egyet kizártunk, tehát 29 adatunk van. A csoportok szerinti szabadsági fok k=3-1
A három csoport nem tartalmaz azonos számú adatot; az első szabadsági foka 9-1, a többié 10-1
Az egész populáció szabadsági foka (29-1)-(3-1)=26
A számítás természetesen az átlagok és a szórásnégyzetek kiszámításával kezdődik:

C15, D15, E15 a kezelések átlaga =ÁTLAG(C5:C14), és így a többi is.

C16, D16, E16 a kezelések szórásnégyzete =VAR(C5:C14), és így a többi (ezek a korrigált empirikus szórásnégyzet értékeit tartalmazzák)

B18 cellába kerül az egész populáció átlaga : =ÁTLAG(C5:E14)

A szórások alatti sorba helyeztük el a csoportátlagokat: C18, D18, E18
Például C18 tartalma: =9*((C15-B18)²). Itt arról van szó, hogy az első kezelés (az első csoport) átlaga kilenc alkalommal tér el az egész populáció átlagától. Tekintettel arra, hogy a vizsgálataink a varianciára irányulnak, ezért nem az eltéréseket, hanem az eltérések négyzetét összegezzük. A másik két csoport tíz adatot tartalmaz, például a D18 cella tartalma: =10*((D15-B18) ²)

Amikor ismerjük valamennyi kezelés értékét, már számítható az MS (Mean Square) értéke; nálunk az F18 cellába kerül: =(C18+D18+E18)/2
C19-ben van a szórásnégyzetek átlaga =(C16*(9-1)+D16*(10-1)+E16*(10-1))/26

Az angol kifejezések magyar megfelelői esetünkben:
(MSB between csoportok közötti része a szórásnégyzetnek) ->SS osztva a szabadsági fokkal
(MSE error csoportokon belüli része a szórásnégyzetnek) ->SS osztva a szabadsági fokkal

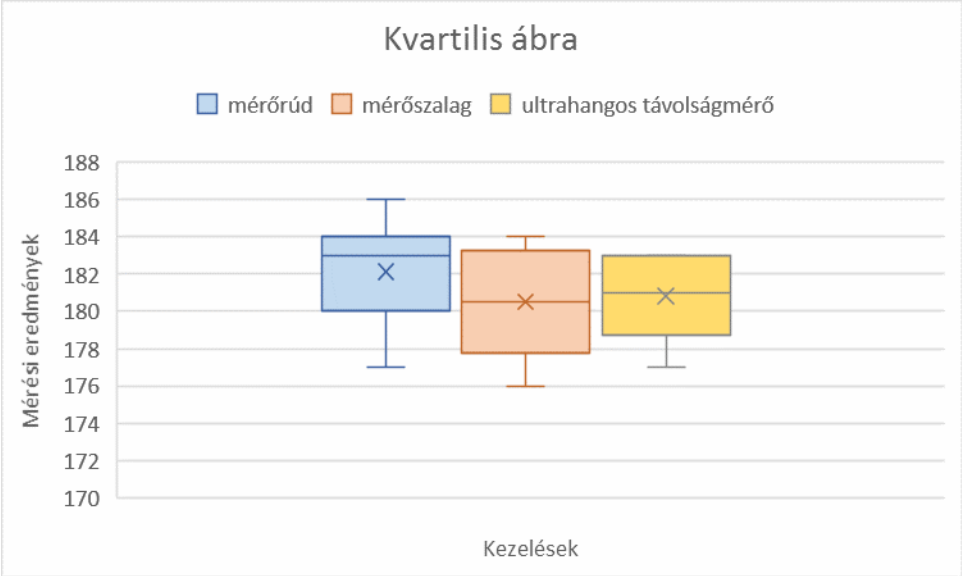
- SS (Sum of Sqares) csoportok között: C18+D18+E18=9,1385+3,6415+0,9208=13,7008, a szabadsági fokkal osztva MS=13,7008/2=6,8504 – ezt okozza az, hogy a csoportok átlaga, vagy szórása különbözik
- SS (Sum of Squares) csoportokon belül a következőképpen számítható. A kezelések szórásnégyzetét megszorozzuk a hozzá tartozó szabadsági fokkal
- =(C16*(9-1)+D16*(10-1)+E16*(10-1))=7,611*8+9,166*9+5,288*9= 190,989, a szabadsági fokkal osztva MS= (7,611*8+9,166*9+5,288*9)/26=7,3457 - ezt okozza az, hogy a csoportokon belül is van szóródás, amely nem magyarázható meg azzal, hogy valamely adat valamelyik csoporthoz tartozik.
- SS (sum of Sqares) Total = a kétféle számítás összege: 13,708+190,989=204,69
- Az MS értékeit összeadni nincs értelme, mert eltérő szabadsági fokra vonatkoznak.

Ha az F-próbát kívánjuk alkalmazni, akkor a szórásnégyzetek hányadosát kell kiszámítanunk. Ez az F19 cellában van nálunk: =F18/C19 = 6,8504/7,3457=0,93257
Az adott szabadsági fokok által meghatározott F-próba kritikus értéke 3,369 > 0,93257 – ez azt jelenti, hogy adott valószínűségi szinten a csoportok (kezelések) leválaszthatóak egymástól; más szóval: nem származnak azonos populációból.
A számláló (6,8504) szabadsági foka 2, a nevező (7,3457) szabadsági foka 26 – ehhez a táblázatban 95% valószínűségi szintre találjuk a 3,369 értéket. 95% valószínűsége van tehát, hogy helyesen döntöttünk, és 5% valószínűsége van annak, hogy tévedtünk.
A táblázat több helyen is megtalálható, például a NIST szerverén, de ki is számítható:
=F.INVERZ.JOBB(0,05;2;26)=3,369016359

A példa, amelyet ismertetünk, annak figyelembe vételével készült, hogy méréstechnikai gyakorlattal nem rendelkező elsőéves hallgatók igen bizonytalan mérési eredményeket produkálnak. A tapasztalatnak megfelelően itt is csak egyszeres szórásnak (kiterjesztési tényező: **1**) megfelelő eltérések esetén feltételezhetjük, hogy adataink azonos eloszlásból származnak. Ez excel (és a legtöbb más szoftver) háromszoros kiterjesztési tényezővel; sokkal szigorúbb követelmények elfogadását kéri számon (95%-os egyezést). Önmagunk számára sokkal engedékenyebb értékeket is elfogadhatunk, de csak a számítási gyakorlatok kedvéért. Például egy engedékenyebb kritikus érték: =F.INVERZ.JOBB(0,4;2;26)=0,949355 gyöngébb szignifikancia szintet rendel a eredményeinkhez.

Megjegyzés. Az F-próba értékeinek táblázata – amely valójában az F eloszlás értékeit tartalmazza – csak 1 és végtelen közé eső értékeket tartalmaz, az EXCEL viszont 1-nél kisebb eredményt számolt

ki nekünk. Valójában ilyenkor a szórásnégyzetek hányadosának reciprok értékével kellene számolnunk. Példánkban eltértünk ettől, mert ezúttal nem mindegy, hogy a szóródást a csoporton belüli, vagy a csoportok közötti szórás okozza-e.
Jelen célkitűzésünk megvizsgálni, hogy a kezelésnek van-e befolyása az eredményre. Két szélsőséges eset fordulhat elő. Az egyik: ha az adatok eloszlását nem befolyásolja, hogy melyik csoportba tartoznak; azaz a két eloszlás teljesen megegyezik. Az eltérésüket kifejező szórás ekkor tart a nullához, a csoporton belüli szórás viszont nem. Ez utóbbit a nullával elosztva végtelent kapunk; az F-próba végtelen. A másik: ha az adatok szóródását csakis az okozza, hogy melyik kezelésből származnak, akkor ez lényegesen nagyobb, mint a csoporton belüli véletlenből származó szóródás. A hányadosuk az előzőnek ellentéte.
Jelölje a csoportok közötti szórásnégyzetet az MSB (between), a csoporton belüli szórásnégyzetet az MSE (error). Az első esetben MSB<MSE és MSE/MSB → végtelen.
A második esetben MSB > MSE és MSB/MSE → végtelen. *Az F-próba értéke ilyenkor nulla és végtelen közé esik.*
Egészen eltérő esetet jelent, ha olyan populációkat (eloszlásokat) hasonlítunk össze, amelyek közt az eltérést **nem** az okozza, hogy tudjuk: melyik kezelésből származnak. Ilyenkor, ha az összehasonlított adatok szórása azonos, akkor az F-próba tart az 1-hez. Fordítva: ha az adatok szórása egyre inkább eltérő, akkor az F-próba tart a végtelenhez. *Értékkészlete ilyenkor az 1 és a végtelen közé esik.*



Az ábrán a kereszt (×) az átlagot jelöli, a vízszintes vonal a mediánt, a bekeretezett négyszög a kvartiliseket, az ábra végén a vonalkák a terjedelem szélső értékeit (**box and whisker plot**). Ilyen adatokat csak meglehetősen alacsony valószínűségi szinten tudunk azonos populációból származó eredményeknek tekinteni.

Megoldás keresése a SCILAB programban

A hallgatói mérés és az adatok kiértékelése általában két különböző időpontban történik. Ezért megnézzük annak a lehetőségét, hogy hogyan lehet külső forrásból származó adatokat beolvasni. Erre sokféle lehetőség kínálkozik. Most azt a változatot nézzük meg, amelynél a számértékeket tabulátorral elválasztott fájlból olvassuk be.

Nem foglalkozunk a kétféle megoldással; mindjárt azt az esetet nézzük át, amelynél a kívülálló adatot már eltávolítottuk. Ennek következménye, hogy az oszlopok (vagy sorok) egyenlőtlen méretűek. Jelen esetben az első oszlopból egy adatot töröltünk. Ennek az adatnak a helyén egy figyelmeztetés jelenik meg: Nan (N**ot a number**) a számérték helyén. Természetesen ezt a problémát is tudjuk kezelni, mindjárt látjuk, hogyan. A beolvasandó táblázat a következő:

182 177 183

179 179 183

181 179 183

177 184 177

186 176 178

183 183 183

184 178 179

184 183 180

183 184 180

182 182

Az első oszlopban tehát egy hiányzó értékkel kell számolnunk (**missing value**). Olvassuk hát be az adatainkat!

```
// Kijelöli a munkakönyvtárat

chdir ('E:\oktatas\scilab\Tmpdir');

filename = fullfile( 'hos_F.txt');

//Beolvasás (a tabulátor a 9-es karakter, ez most a mezőelválasztó)

rr= csvRead(filename, ascii(9));
```

```
 rr  =

    182.    177.    183.
    179.    179.    183.
    181.    179.    183.
    177.    184.    177.
    186.    176.    178.
    183.    183.    183.
    184.    178.    179.
    184.    183.    180.
    183.    184.    180.
    Nan     182.    182.
```

Sajnos, a program észrevette a hiányzó értéket. A teljes táblázat oszlopai és sorainak a számára azért szükségünk lehet:

```
nr = size(rr, "r")
nc = size(rr, "c")
```

Nem kell tudnunk, melyik oszlop tartalmaz hiányzó értéket. Olvassuk be mindegyiket, majd alakítsuk át olyan vektorrá, amely már csak numerikus értékeket tartalmaz (t= **treatment**):

```
t0 = rr(:,1)
t1= thrownan(t0)
t0 = rr(:,2)
t2= thrownan(t0)
t0 = rr(:,3)
t3= thrownan(t0)
```

Megjegyezzük, hogy létezik ezt megkerülő megoldás is. Az átlag és a szórás például így is kiszámítható (lásd: **Data with missing values**):
m1= nanmean(t0)
s1 = nanstdev(t0)
Ezt most csak példa, az alábbiakban a numerikussá változtatott vektorokkal dolgozunk.
Nézzük meg a vektorok hosszát!
n1= size(t1,"r") = 9
n2= size(t2,"r") = 10
n3= size(t3,"r") = 10
Az átlagok:
mean(t1) = 182,111
mean(t2) = 180,5
mean(t3) = 180,8
A szórások:
stdev(t1) = 2,7588242
stdev(t2) = 3,0276504
stdev(t3) = 2,2997584
vagy mindjárt a szórásnégyzet:
variance(t1) = 7,6111111
variance(t2) = 9,1666667
variance(t3) = 5,2888889
Az adatösszesség átlaga:
m0= nanmean(rr) = 181,10345
mv1= n1*((m0-m1)^2) = 9,1384595
mv2= n2*((m0-m2)^2) = 3,6414982
mv3= n2*((m0-m3)^2) = 0,9208086
A szabadsági fok az egész populációra (az összes adat az első zárójelben, a csoportok száma a második zárójelben, a SCILAB nem igényli a zárójeleket, ez itt csak a megértés kedvéért szerpel):
df = (n1+n2+n3-1) – (nc-1) = 26
Mean sqare (csoportok között)
MSB = (mv1+mv2+mv3)/2 = 6,8503831
Mean sqare (csoportok nélkül, angolul erre az error szó használatos)
MSE = ((n1-1)*v1 +(n2-1)*v2+(n3-1)*v3)/df = 7.3457265
Most már számítható az egyszeres osztályozásra az F-próba értéke:
F = MSB / MSE = 0,9325671
esteleg (mert nem egyforma az adatok darabszáma; **unequal**)
f=ftuneq(t1, t2, t3) = 0.9325671
[f p]=ftuneq(t1, t2, t3)
p = 0.4063111 – hát, ez bizony elég gyenge szignifikancia szint. Komolyabb munkához meg kellene ismételni a méréseket.

Néhány megjegyzés az adatok eloszlásáról

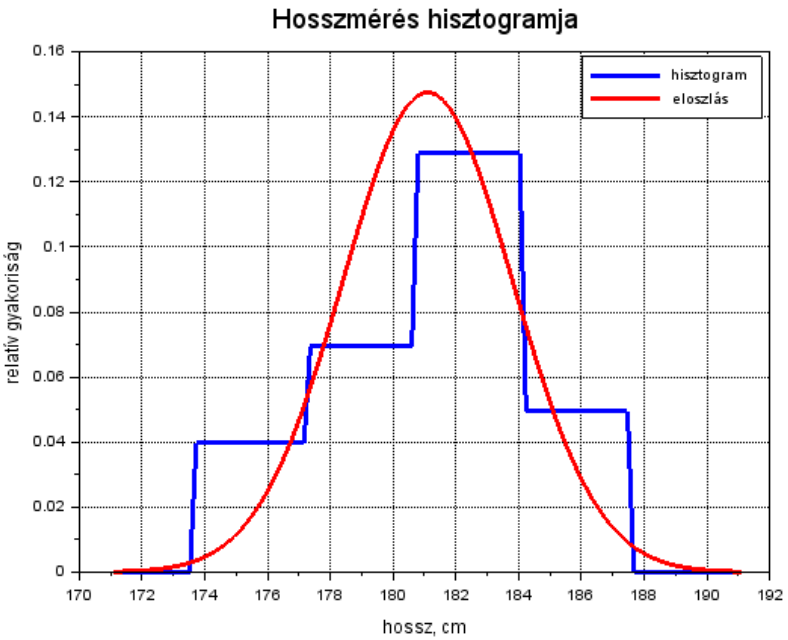
Ha valaki kíváncsi rá, többféle lehetőség is adódik az eloszlások tulajdonságainak megtekintésére.
Nézzük, hogyan lehet a gyakoriságok eloszlását összehasonlítani a normáleloszlásával!

Az egyik ilyen lehetőség a hisztogram. Ehhez először is szükségünk van az átlag és a szórás értékére. Kiválasztjuk az adathalmazból a numerikus értékeket: r0= thrownan(rr)
a szórás adatait: s0=nanstdev(rr)
az elemek számát: n0= size(r0,"r")
rendezzük az adatokat (átlag, szórás, és egytől az utolsóig az összes elemet választjuk:
data=distfun_normrnd(m0,s0,[1 n0]);
létrehozzuk a hisztogramhoz szükséges információkat: h=distfun_histocreate("data",data);

létrehozzuk a vízszintes tengelyt (az átlag körül plusz-mínusz 10 értéket veszünk, ebbe belefér valamennyi mért adatunk):
x=linspace(m0-10,m0+10);
a függőleges tengely létrehozása a hisztogram számára: y=distfun_histopdf(x,h);
És már rajzolunk is (vonalvastagság és vonalrácsozat): plot(x,y,'Linewidth',3);xgrid(0, 1, 7)
A normáleloszlás számára szintén létrehozunk ordináta értékeket (ugyanazokhoz az x értékekhez, ugyanazzal az átlaggal és szórással:

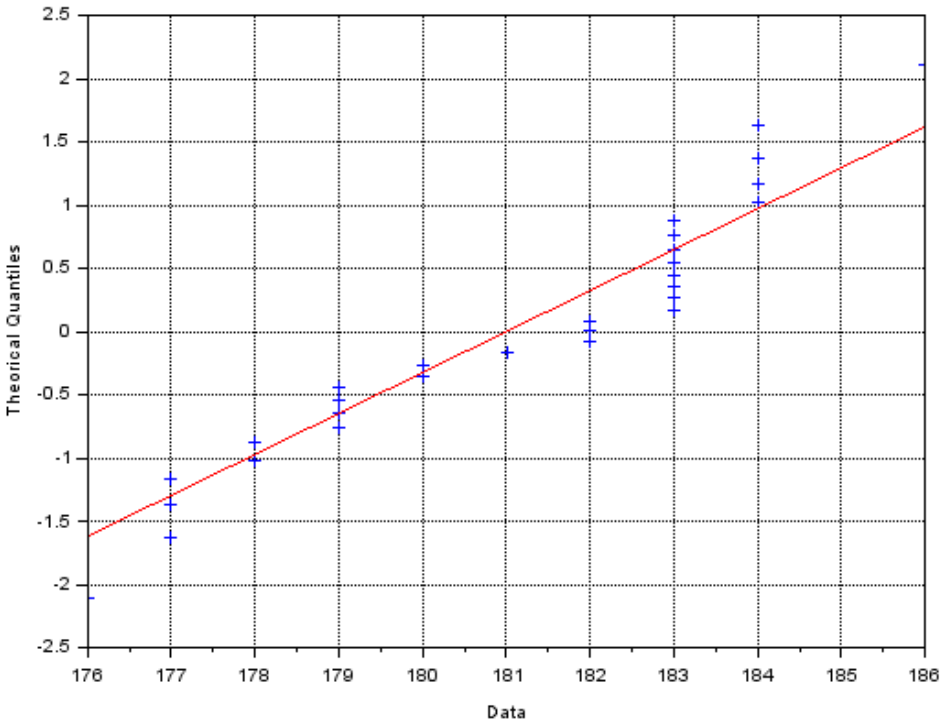
ynorm=distfun_normpdf(x,m0,s0);

és azt is kirajzoljuk (r= red, tehát piros színnel): plot(x,ynorm,'r','LineWidth',3)



Ha becsuktuk a grafikus képet, készíthetünk még egy normalitásvizsgálati ábrát is (r0 tartalmazza mind a huszonkilenc mért értéket):

normplot(r0) ,xgrid(0, 1, 7)



Hát, sajnos, ez az adathalmaz ezer sebből vérzik. Túl sokszor fordul elő a 183 és a 184 cm-es mérési eredmény.