Szent István Egyetem Állatorvos-tudományi Kar
Biomatematikai és Számítástechnikai Tanszék

Biomatematika 13.

Varianciaanalízis (ANOVA)

Fodor János

Last Revision Date: November 4, 2006 Version 1.25

Table of Contents

1. Bevezetés

Nagyon sok esetben felmerülnek olyan kérdések, hogy:

- hat-e a műtét típusa a túlélési időre?
- hat-e a kezelés típusa a túlélési arányra egy bizonyos betegség esetén?
- hat-e a művelési mód a terméseredményekre?
- hat-e a táp típusa a testsúlyra?

Ilyen típusú kérdések esetén mindig felmerül az a gyanú, hogy a mért vagy megfigyelt **különbséget**

nem az általunk vizsgált effektus okozta. Lehet, hogy a beteg gyorsabb felépülése nem a gyógyszer, kezelés, operáció típusától függ, hanem egyszerűen a jobb kondíciótól.

Lehet, hogy azon a parcellán, amelyen a jobb eredményt érték el, a talaj minősége lényegesen jobb volt, mint a többin, így ez okozta a jobb termésered-ményt.

Az ilyen típusú kérdések megválaszolására a **varianciaanalízis** módszere szolgál, amely tulajdonképpen a t-próba kiterjesztése több mintára. Azt kell eldöntenünk, hogy **kettőnél több populáció átlagai**

azonosak-e vagy sem.

2. ANOVA

Az F próbát két variancia összehasonlítására használtuk, de ez a próba három vagy annál több csoport átlagának összehasonlítására is alkalmas. Ezt a technikát **varianciaanalízisnek** hívják (Analysis of Variance, ANOVA).

Például három csoport esetén csak azt tudja kimutatni, hogy a három átlag nem egyenlő; azt már nem, hogy hol a különbség. Erre a célra más alkal-

mas próbát kell használnunk.

Felmerülhet a kérdés: miért nem alkalmazzuk a t-próbát páronként (két-két átlagot összehasonlítva egyszerre)? Azért, mert sok t-próbát kellene lefuttatni (minden lehetséges párra egyet). Ekkor az igaz null hipotézis elvetésének esélye nő, hiszen az összes lehetséges páronkénti összehasonlítás nagy száma miatt véletlenül is kaphatunk szignifikáns eltéréseket. Például, ha 3 átlagot hasonlítunk össze, 3 t-próbára van szükség; 5 átlagra 10 t-próba, míg 10 átlagra 45 t-próba kell. Az F-próba viszont szimultán teszteli az átlagok egyenlőségét.

Az F próba három vagy annál több átlag összehasonlítására történő alkalmazásának feltételei:

1. A populációk eloszlása (megközelítőleg) normális.
2. A minták egymástól függetlenek.
3. A populációk varianciái egyenlők.

Még ha átlagokat is hasonlítunk össze, a próbában varianciákat használunk.

A populáció varianciájának kétféle becslését készítjük el. Az elsőt **csoportok közötti** varianciának nevezik, és ez az átlagok szórásnégyzetét jelenti. A második

a **csoportokon belüli** variancia, és ezt az összes adat alapján határozzuk meg. Ha nincs különbség az átlagok között, akkor a csoportok közötti és csoportokon belüli varianciák nagyjából egyenlők, és az F próbastatisztika értéke nagyjából 1. Amikor az átlagok lényegesen eltérőek, a csoportok közötti variancia lényegesen nagyobb, mint a csoportokon belüli, és az F próbastatisztika értéke jóval nagyobb mint 1. Mivel a varianciákat hasonlítjuk össze, ezért hívják az eljárást varianciaanalízisnek.

2.1. A hipotézis

H_0 : $\mu_1 = \mu_2 = \ldots = \mu_k$

H_1 : Legalább egy átlag különbözik a többitől.

2.2. A szabadsági fokok

A számláló szabadsági foka (d.f.N.): $k - 1$.

A nevező szabadsági foka (d.f.D.): $N - k$.

Itt k a csoportok száma, N pedig az összes megfigyelés száma ($N = n_1 + n_2 + \ldots + n_k$). Az egyes csoportokra vonatkozó minták nem feltétlenül azonos

elemszámúak. Az F-próba mindig jobboldali.

2.3. Példa: vérnyomáscsökkentés

Példa. Egy kutató három különböző technikát szeretne összehasonlítani magas vérnyomású személyek vérnyomásának csökkentésére. Az egyes személyeket véletlenszerűen osztja be három csoportba: az első csoport tagjai **gyógyszert** szednek; a második csoportba tartozók speciális **tornát** végeznek; a harmadiké speciális **diétát** követnek. Négy hét után feljegyzik az egyes személyek vérnyomásának csök-

kenését. $\alpha = 0.05$ szinten teszteljük azt a hipotézist, hogy nincs különbség a három módszerrel elért átlagos vérnyomáscsökkenések között.

Az adatok:

Gyógyszer	Torna	Diéta
10	6	5
12	8	9
9	3	12
15	0	8
13	2	4
$\bar{X}_1 = 11.8$	$\bar{X}_2 = 3.8$	$\bar{X}_3 = 7.6$
$s_1^2 = 5.7$	$s_2^2 = 10.2$	$s_1^2 = 10.3$

Megoldás. **1. lépés: A null és az alternatív hipotézis felállítása.**

H_0 : $\mu_1 = \mu_2 = \mu_3$.

H_1 : Legalább egy átlag eltér a többitől.

2. lépés: A kritikus érték meghatározása.

d.f.N. $= k - 1 = 3 - 1 = 2$

d.f.D. $= N - k = 15 - 3 = 12$

A kritikus érték 3.89.

3. lépés: A próbastatisztika értékének kiszámítása.

(a) Mindegyik csoport átlagának és varianciájának kiszámítása (lásd a fenti táblázatban)

(b) A "nagy átlag" ($\bar{X}_{GM}$, az összes adat átlagának) kiszámítása: $\bar{X}_{GM} = 7.73$.

(c) A csoportok közötti variancia (s_K^2) kiszámítása.

$$
\begin{aligned}
s_K^2 &= \frac{\sum n_i(\bar{X}_i - \bar{X}_{GM})^2}{k-1} \\
&= \frac{5 \cdot (11.8 - 7.73)^2 + 5 \cdot (3.8 - 7.73)^2 + 5 \cdot (7.6 - 7.73)^2}{3-1} \\
&= \frac{160.13}{2} = 80.07.
\end{aligned}
$$

(d) A csoportokon belüli variancia (s_B^2) kiszámítása.

$$\begin{aligned} s_B^2 &= \frac{\sum(n_i-1)s_i^2}{\sum(n_i-1)} \\ &= \frac{(5-1)5.7+(5-1)10.2+(5-1)10.3}{(5-1)+(5-1)+(5-1)} \\ &= \frac{104.8}{12} = 8.73. \end{aligned}$$

(e) Az F próbastatisztika kiszámítása.

$$F = \frac{s_K^2}{s_B^2} = \frac{80.07}{8.73} = 9.17$$

4. lépés: A döntés. Mivel 9.17 $>$ 3.89, a null hipotézist elutasítjuk.

A 3. lépésben (c) esetében kapott tört számlálóját a **csoportok közötti négyzetösszegnek** is nevezik és SQ_K-vel jelölik, míg a (d) esetben kapott tört számlálóját a **csoportokon belüli négyzetösszegnek** hívják, és SQ_B-vel jelölik. Aztán SQ_K-t kell elosztanunk a számláló szabadsági fokával ahhoz, hogy megkapjuk a csoportok közötti varianciát. SQ_B-t pedig $N - k$-val, hogy a csoportok közti varianciát. E két varianciát néha átlagos négyzeteknek

is nevezik, és az MQ_K illetve az MQ_B szimbólumokkal jelölik.

A következő táblázatban foglaljuk össze a varianciaanalízis lényegét.

Forrás	Négyzetek összege	d.f.	Átlagos négyzetek	F
Közötti	SQ_K	$k-1$	$MQ_K = \frac{SQ_K}{k-1}$	$\frac{MQ_K}{MQ_B}$
Belüli	SQ_B	$N-k$	$MQ_B = \frac{SQ_B}{N-k}$	
Teljes	$SQ_K + SQ_B$	$N-1$		

Az előző példára vonatkozó ANOVA tábla a következő:

Forrás	Négyzetek összege	d.f.	Átlagos négyzetek	F
Közötti	160.13	2	80.07	9.17
Belüli	104.80	12	8.73	
Teljes	264.93	14		

A fentebb vizsgált próbát egyszempontú ANOVA-nak nevezik, mert csak egy független változó szerepel benne.

Léteznek többszempontú változatok is, de ezekkel nem foglalkozunk, csak megemlítünk egy példát:

Szeretnénk tesztelni két talajtípus és két műtrágya hatását egy bizonyos növény növekedési sebességére. Ekkor a két független változó: a talajtípus és a műtrágya fajtája. A függő: a növény mérete. A többi faktor (hőmérséklet, napfényes órák száma, öntözés, stb) ugyanaz.

2.4. Példa: oldat tömménysége

Példa. Tegyük fel, hogy van 12 növényünk különböző helyeken. Háromféle tápszeres oldattal öntözzük ezeket. Kíváncsiak vagyunk arra, hogy a töménységnek van-e valami hatása a növekedésre.

Sorsoljuk ki, hogy melyik kapjon **tiszta vizet**, és melyikeket öntözzük **tömény** illetve **híg** oldattal. Úgy végezzük el a sorsolást, hogy a harmada tiszta, a harmada tömény, a harmada híg öntözővizet kapjon.

Beleteszünk egy kalapba 12 cédulát, amelyek közül négyen **"t"** (tömény oldat), négyen **"v"** (víz), négyen

pedig **"h"** (híg oldat) betű van, és minden növénynél húzunk egyet. Ily módon azt próbáljuk elérni, hogy az egyéb faktorok (termelési különbségek, növények kondíciója) kiegyensúlyozzák egymást.

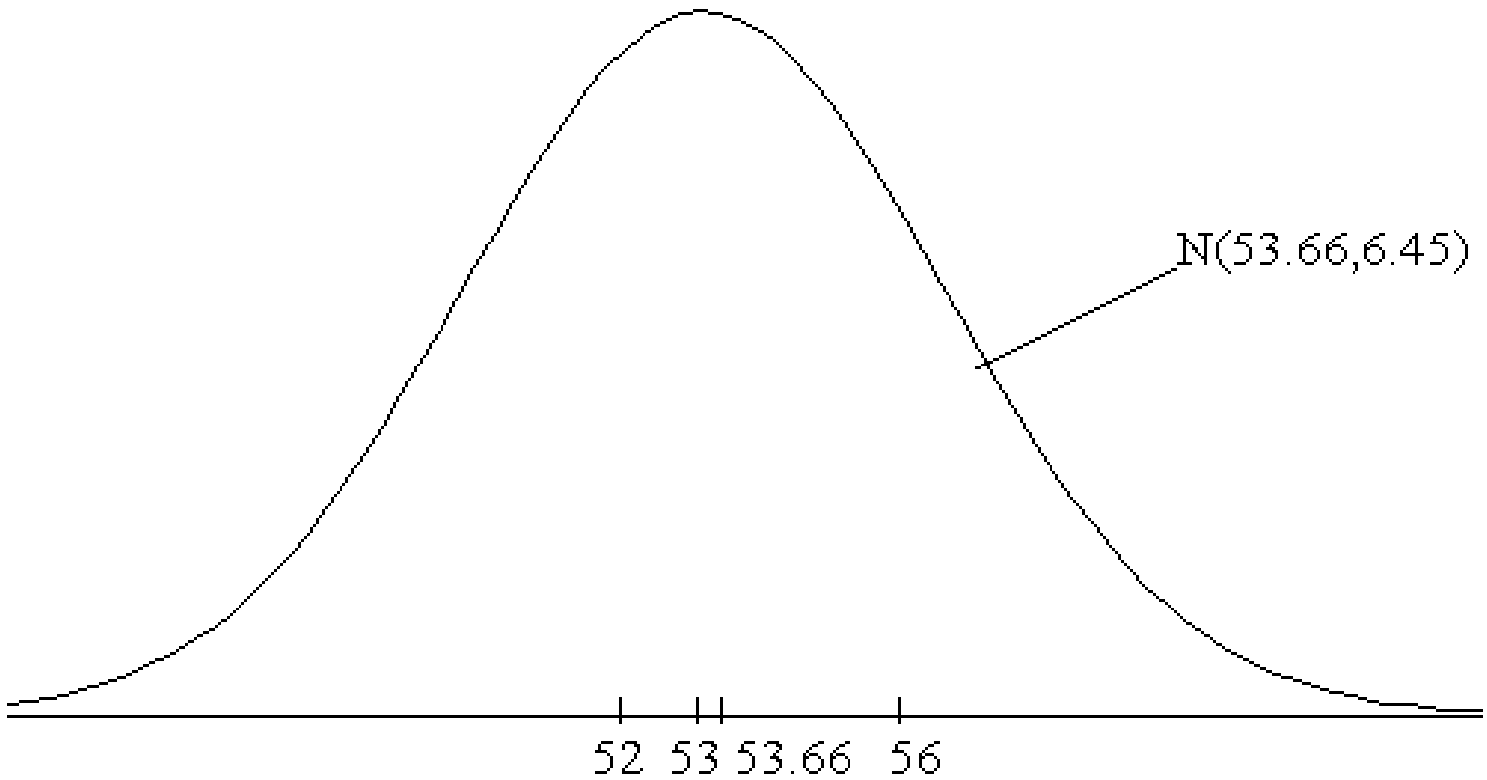
Nézzük meg a statisztikai analízist, ha a kísérletben a növények magassága (cm-ben mérve) az alábbiak szerint alakult:

	kezelés		
	tömény oldat	híg oldat	tiszta víz
	56	57	54
	48	50	46
	66	47	60
	54	58	48
átlag:	56	53	52
variancia:	56	28.66	40
szórás:	7.48	5.35	6.32

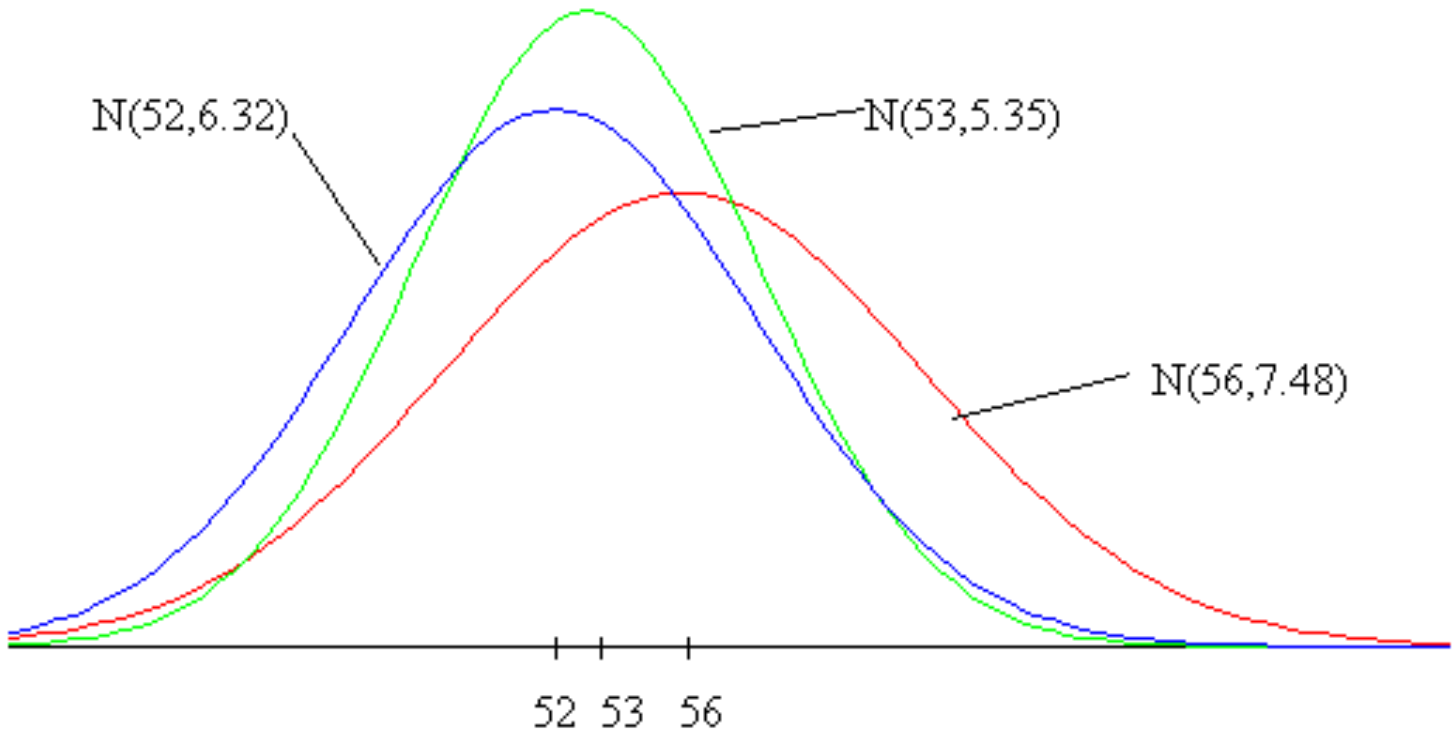
Látható, hogy az oldatokkal öntözött növények átlagos magassága nagyobb, mint a tiszta vízzel öntözötteké.

Az is rögtön szembetűnik, hogy az egyes növények között elég nagy különbségek vannak a mintákon belül is.

Ki kell számolnunk, hogy mekkora a valószínűsége annak, hogy a mért különbségek csupán a véletlen mintavétel következményei. Azaz: el kell döntenünk, hogy a mintákat ugyanabból a populációból vettük-e (nullhipotézis), vagy pedig különbözőkből (alternatív hipotézis). A két lehetőséget szemlélteti a következő ábra.



Ugyanabból a populációból származnak a minták



Különböző populációból származnak a minták

Az **ANOVA táblázat**:

A variancia eredete	Eltérés-négyzetösszeg	szabadsági fok	átlagos négy-zetes eltérés (variancia)	F
Kezelések közötti	3.08	2	1.54	
Kezelésen belüli	93.42	9	10.38	0.148
Teljes	94.96	11		

Ahogy azt már láttuk, a szabadsági fok $(2, 9)$. Az 5%-os szignifikanciaszinthez és ehhez a szabadsági fokhoz tartozó táblázatbeli F érték: **4.26**.

Mivel $0.148 < 4.26$, így elfogadjuk a nullhipotézist.