Szent István Egyetem Állatorvos-tudományi Kar
Biomatematikai és Számítástechnikai Tanszék

13. előadás

Hipotézisvizsgálatok. ANOVA

Fodor János
Biomatematika 2001

Copyright © 2001 jfodor@univet.hu
Last Revision Date: November 23, 2001 Version 1.10

Table of Contents

1 Hipotézisvizsgálatok (folyt.) 4

1.1 Más módszerek hipotézisvizsgálatokra 5

- P-értékek 5
- Hipotézisvizsgálat konfidenciaintervallumokkal 7

1.2 Kétmintás próbák 9

1.3 Páros t-próba 17

1.4 Az F-próba 21

2 Varianciaanalízis (ANOVA) 32

2.1 A kezelések közötti és a kezeléseken belüli variancia 34

2.2 Az ANOVA táblázat 50

1. Hipotézisvizsgálatok (folyt.)

A hipotézisvizsgálat lépéseinek összefoglalása

1. lépés	A null- és az alternatív hipotézis megfogalmazása
2. lépés	A szignifikancia-szint (α) megválasztása
3. lépés	A próbastatisztika kiválasztása
4. lépés	A döntési szabály megfogalmazása
5. lépés	Döntés a minta alapján

1.1. Más módszerek hipotézisvizsgálatokra

A következő két módszer ekvivalens az eddig említettekkel. Csupán azért tekintjük át ezeket, mert a szakirodalomban és a statisztikai programcsomagokban elég gyakran ilyen formában fordulnak elő.

- P-**értékek**

A módszer menete a következő: Először állítsuk fel a null- és az alternatív hipotézist, majd számítsuk ki annak a valószínűségét, hogy éppen a mintának megfelelő értéket, vagy annál nagyobbat kapjunk.

Ez a valószínűség a P-érték. Ha ez kisebb, mint 5%, akkor elutasítjuk a H_0-t.

Példánknál maradva:

$$H_0: \qquad \mu = 21.5,$$
$$H_1: \qquad \mu > 21.5.$$

Tegyük fel, hogy a minta átlaga 23.1 kg-nak adódott. Számítsuk ki annak a valószínűségét, hogy ekkora, vagy ennél nagyobb értéket kapjunk egy $N(21.5, 2.5^2)$ eloszlású populációból:

$$P(\overline{X} > 23.1) = P\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{23.1 - 21.5}{0.8}\right)$$

$$= P(Z > 2) = 2.275\%.$$

Ez az érték kisebb, mint 5%, ezért elutasítjuk H_0-t.

- **Hipotézisvizsgálat konfidenciaintervallumokkal**

Ez a harmadik módszer a már megismert konfidenciaintervallumok elvén alapul, és a kétoldali hipotézisvizsgálat esetére mutatjuk meg.

Tegyük fel, hogy a minta átlaga 22.5 kg. Ha mintánk alapján elkészítjük az új táppal etetett malacok átlagsúlyára vonatkozó, 95% megbízhatósági szintű konfidenciaintervallumot, akkor az

$$]22.5 - 2 \cdot 0.8, 22.5 + 2 \cdot 0.8[=]20.9, 24.1[.$$

Mivel a 21.5 kg beleesik ebbe az intervallumba, nem mondhatjuk, hogy az új táp hatására legalább 95%-os valószínűséggel változott volna a malacok súlya.

Megjegyzés:

Ha nem ismerjük a szórást, akkor az előbb említett két módszer esetén is a t-eloszlást kell használnunk.

1.2. Kétmintás próbák

Előző példánkban azt vizsgáltuk, hogy az új táp hatására változnak-e a malacsúlyok az előző évek adatai alapján megállapított, elméleti értékhez képest. Nagyon gyakran azonban nem áll rendelkezésünkre ilyen elméleti érték. Ilyen esetekben célszerű egy másik (kontroll) csoporthoz viszonyítanunk az eredményeinket. Szinte mindig ez az eljárás gyógyszerhatás vizsgálatnál.

Gyakran előfordul az is, hogy egyszerűen csak két csoportot (populációt) szeretnénk összehasonlítani.

Például, szeretnénk megtudni, hogy vajon a dohányosok rövidebb ideig élnek-e, mint a nem dohányosok, a Holstein-Frízek tejtermelése nagyobb-e Németországban, mint nálunk stb.

A két összehasonlítandó csoportnak nem tudjuk a populáció átlagait, csak a belőlük kiválasztott minták átlagait tudjuk összehasonlítani, és azt vizsgáljuk, hogy a kettő szignifikánsan különbözik-e.

Illusztrációként nézzük meg a következő példát:

Két termelőtől kapott sertéshús zsírtartalmát szeretnénk összehasonlítani. Az egyik termelőtől 22

elemű mintánk van, a minta **átlaga 27%**, **szórása 4%**. A másik esetben **25** elemű minta alapján az átlagos zsírtartalom **24%**, szórása **3%**.

A számok alapján úgy tűnik, hogy az első termelő esetén nagyobb a zsírtartalom. Kérdés, hogy ez a különbség szignifikáns-e, vagy csak a mintavételi hiba okozta a különbséget?

Nullhipotézis: **nincs különbség** a két zsírtartalom között.

Alternatív hipotézis: **van különbség** a két zsírtartalom között.

Formálisan felírva:

$$\begin{aligned} H_0 : &\qquad \mu_1 = \mu_2, \\ H_1 : &\qquad \mu_1 \neq \mu_2. \end{aligned}$$

Ha a hipotéziseket nem ebben a formában, hanem az átlagok különbségére írjuk fel, akkor formailag az egymintás esetet kapjuk vissza:

$$\begin{aligned} H_0 : &\qquad \mu_1 - \mu_2 = 0, \\ H_1 : &\qquad \mu_1 - \mu_2 \neq 0. \end{aligned}$$

Ha a mintáink elég nagyok, akkor a mintaátlagok normális eloszlásúak lesznek. (Nem túl nagy minták

esetén, az egymintás esethez hasonlóan, itt is a t-eloszlást kell használni a normális eloszlás helyett.)

Mintaátlag	Várható értéke	Szórása
$\overline{X}_1$	μ_1	$\sigma_1/\sqrt{n_1}$
$\overline{X}_2$	μ_2	$\sigma_2/\sqrt{n_2}$

Ha a minták függetlenek, akkor a mintaátlagok különbsége is normális eloszlású. Ha a szórásukat ismerjük, akkor

$\overline{X}_1 - \overline{X}_2$ várható értéke $\mu_1 - \mu_2$,

szórása $\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$.

Példánkban $\overline{X}_1 - \overline{X}_2$ eloszlása $N(0, (1.04)^2)$.

Ha igaz a nullhipotézis, akkor a megfigyelt különbségnek 2 szóráson belül kell lennie. Esetünkben a különbség 27%-24%=3%. Ez kívül esik a

$$[-2.08\%, 2.08\%]$$

intervallumon. Ezek szerint az 3%-os különbség túl nagy ahhoz, hogy a mintavételi hibával magyarázható legyen, ezért elutasítjuk a H_0-t.

A második módszerünkkel ugyanez a következőképpen néz ki: A 3% 3/1.04=2.88 szórásnyira van a 0-tól. A 2.88-os z érték azt jelenti, hogy ezen érték fölé a normális eloszlás görbe alatti területének csupán 0.2%-a (P-érték) esik, azaz annak a valószínűsége, hogy ekkora legyen a különbség csupán 0.2%.

Harmadik módszerünket alkalmazva, azt az intervallumot adjuk meg, amelybe 95% valószínűséggel beleesik a két populáció átlag különbsége. Ez az intervallum:

$$[3 - 2 \cdot 1.04, 3 + 2 \cdot 1.04] = [0.92, 5.08].$$

Ebbe az intervallumba a 0 nem esik bele, azaz azt mondhatjuk, hogy a két termelő által szállított hús zsírtartalma nem azonos.

Megjegyzések:

1. Kis mintaelemszámok esetén mind a három esetben a t-eloszlás táblázatából kikeresett értékeket kell használni. A szabadsági fok $n_1 + n_2 - 2$ lesz, mivel két becsült értéket, az átlagot és a szórást, használunk.

2. A különbség szórását csak akkor becsülhetjük az előbb ismertetett módon, ha feltehető, hogy a

két populáció szórása ismert. Ha a szórások ismeretlenek, de feltehető, hogy egyenlőek, akkor a különbség szórását az

$$s = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}$$

képlettel számíthatjuk ki.

1.3. Páros t-próba

Tegyük fel, hogy egy mintán vizsgáljuk valamilyen kezelésnek a hatását. Ilyen esetekben nem a mintaátlagokat hasonlítjuk össze, hanem a kezelés **előtti**

és **utáni** érték **különbségéről** állapítjuk meg, hogy szignifikánsan különbözik-e nullától.

Példa.

Egy lázcsillapító gyógyszer hatásosságát vizsgáljuk. A betegek hőmérsékletét kétszer, a lázcsillapító bevétele előtt illetve után mérjük meg. A mért értékeket és a változást a következő táblázat tartalmazza.

testhőmérséklet gyógyszer		különbség
előtt	után	
39.2	38.6	0.6
38.7	37.2	1.5
37.9	36.8	1.1
38.8	37.9	0.9
39.4	38.2	1.2
38.2	38.0	0.2
38.6	36.9	1.7
38.8	37.8	1.0
39.0	37.9	1.1
38.5	37.6	0.9

A különbségek átlaga: $\overline{d} = 1.02$, szórása $s_d = 0.42$.

A hipotéziseket formálisan felírva:

$$H_0: \qquad \mu_d = 0,$$
$$H_1: \qquad \mu_d \neq 0.$$

Vagyis a nullhipotézis szerint a gyógyszer hatástalan.

Ettől kezdve mindent úgy kell csinálnunk, mint az egymintás próba esetén. Ezek szerint az $\overline{X}_d$ mintaátlag eloszlása normális, várható értéke 0, szórása pedig $0.42/\sqrt{10} \approx 0.134$. A szabadsági fok esetünkben $n - 1 = 9$, az ehhez és a 95%-os valószínűséghez tartozó t érték 2.26.

Így a különbségek átlagának, ha igaz a nullhipotézis,

bele kellene esnie a

$$[0 - 2.26 \cdot 0.134, 0 + 2.26 \cdot 0.134] = [-0.302, 0.302]$$

intervallumba. Ez nem teljesül, ezért elutasítjuk a nullhipotézist. Tehát azt mondhatjuk, hogy a lázcsillapító gyógyszer hatásos.

1.4. Az F-próba

A t-próba tárgyalásánál már volt arról szó, hogy a próbát másképp kell elvégezni, ha a két populáció szórása (szórásnégyzete) megegyezik, és másképp akkor, ha nem. Most azt nézzük meg, hogy hogyan

lehet eldönteni, hogy a szórásnégyzetek megegyeznek-e.

Legyen σ_1^2 az első, σ_2^2 a második populáció varianciája (szórásnégyzete). Írjuk fel a null- és az alternatív hipotézist:

$$\begin{aligned} H_0 : &\qquad \sigma_1^2 = \sigma_2^2, \\ H_1 : &\qquad \sigma_1^2 \neq \sigma_2^2. \end{aligned}$$

Ha H_0 igaz, akkor a két populáció szórásnégyzetének hányadosa 1. Két minta alapján becsüljük ezt a hányadost. A becslést **F-statisztikának** nevezzük,

ahol

$$\mathbf{F} = \frac{\mathbf{s_1^2}}{\mathbf{s_2^2}},$$

és s_1^2 az első, s_2^2 a második minta korrigált tapasztalati szórásnégyzete.

Ha ez az F érték elég közel van 1-hez, akkor azt mondhatjuk, hogy az eltérést csupán a véletlen mintavételből származó hiba okozta, így elfogadhatjuk a H_0-t, egyébként pedig elutasítjuk.

Most már csak az a kérdés, hogy mit jelent az, hogy elég közel van az 1-hez.

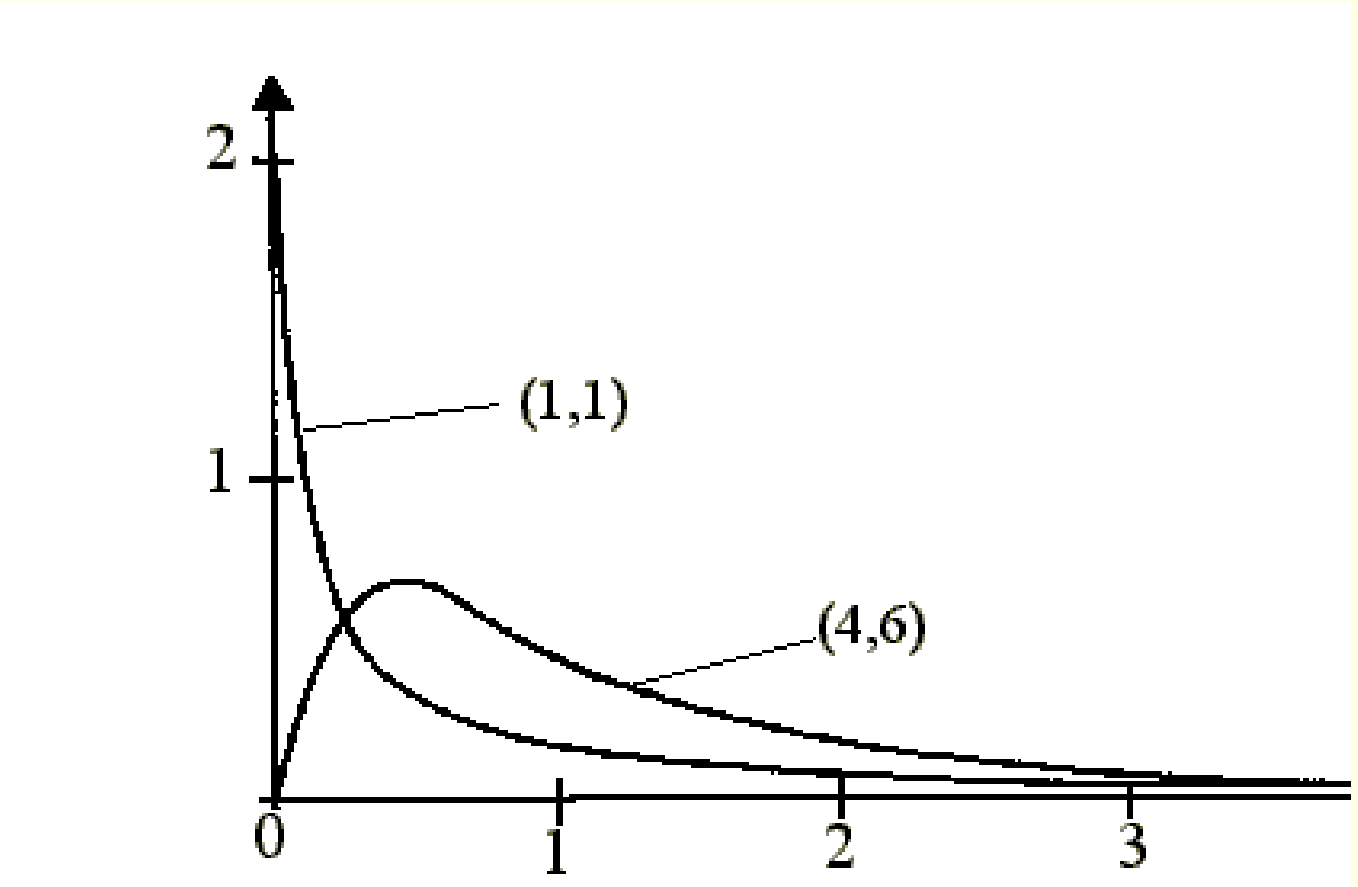
Minél nagyobbak a mintáink, annál jobban megközelíti a minták szórásnégyzete a populáció szórásnégyzetét. Ilyen esetekben az 1-től csak kis eltérést engedünk meg.

Ha a mintáink viszonylag kicsik, akkor pedig még nagyobb eltérés esetén is elfogadjuk a nullhipotézist.

Az is előfordulhat, hogy az egyik minta kicsi, a másik pedig nagy. Ebből is kitűnik, hogy **mind a két minta elemszámától függ**, hogy az 1 körüli mekkora intervallumban fogadjuk el a H_0 hipotézist.

Az F-statisztika eloszlása különböző mintaelemszámok

esetén más és más lesz. Hasonlóan a t-statisztikához, itt is a **szabadsági fok** mutatja meg, hogy melyik F-eloszlást kell választanunk. A két minta szabadsági foka esetünkben $(n_1 - 1, n_2 - 1)$, ahol n_1 az első minta, n_2 pedig a második minta elemszáma. A következő ábrán különböző szabadsági fokokkal rendelkező F-eloszlásokat láthatunk.



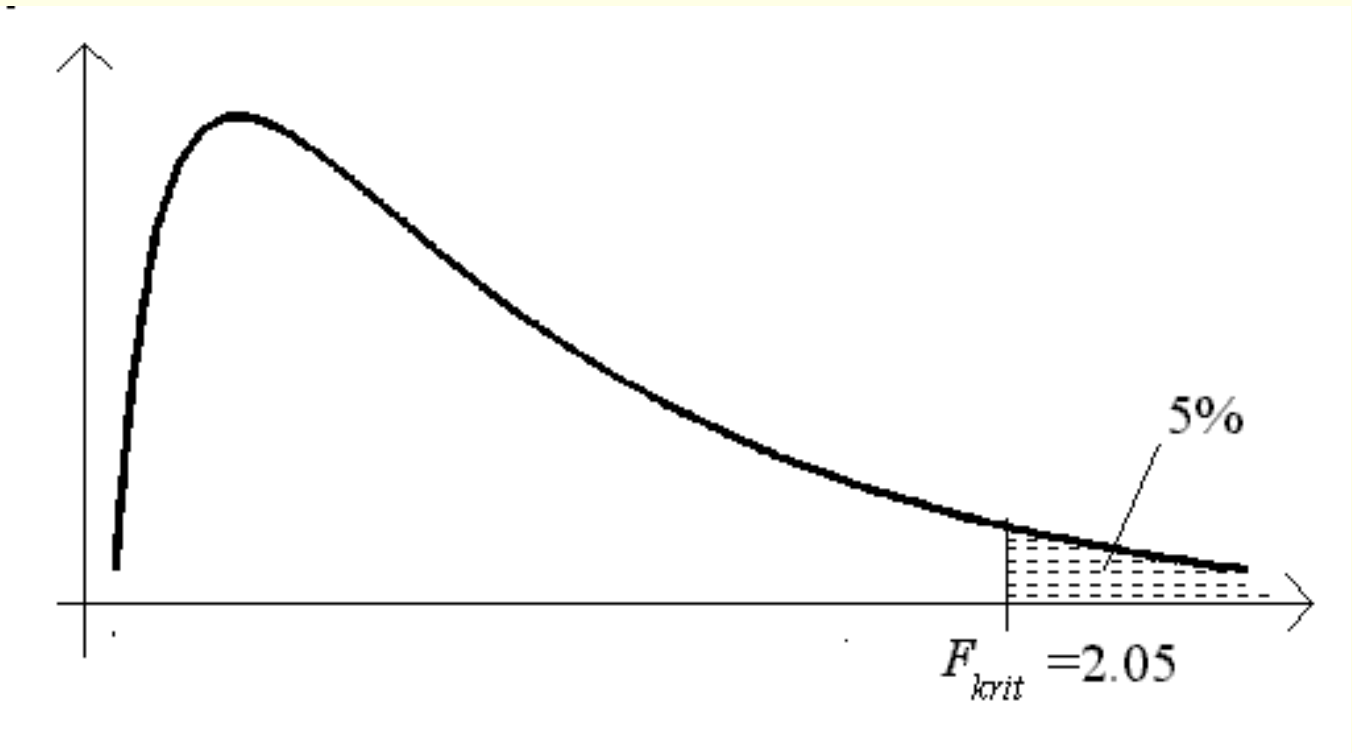
Két F-eloszlás különböző szabadsági fokkal

Ezek után már csak annyi a teendő, hogy az F-eloszlás ismeretében megmondjuk hogy mekkora a

valószínűsége annak, hogy a két minta szórásnégyzetének hányadosa akkora vagy annál nagyobb értéket vegyen fel, mint amekkorát a számolás során kaptunk. Ha ez a valószínűség kisebb, mint 5%, akkor elvetjük a nullhipotézist.

A táblázatokat $s_1 > s_2$ esetre közlik. Mivel a megfelelő F-eloszlást a két szabadsági fok **együtt** határozza meg, a táblázatokat adott – általában **5%**-os – szignifikanciaszintre közlik a különböző statisztika könyvek. Ilyen esetben azt kell megnézni, hogy a számunkra megfelelő szignifikanciaszinthez és szabadsági fokhoz tartozó F értékhez hogyan viszonyul az általunk

kiszámított érték. Ha nagyobb, akkor - mint az alábbi ábrán is látható - elutasítjuk H_0-t, ellenkező esetben elfogadjuk.



$(21, 24)$ szabadsági fokú F-eloszlás

Példa.

Tekintsük újra a kétmintás t-próba esetén vizsgált példát.

Két termelőtől kapott sertéshús zsírtartalmát szeretnénk összehasonlítani. Az egyik termelőtől **22** elemű mintánk van, a minta **átlaga 27%**, **szórása 4%**. A másik esetben **25** elemű minta alapján az átlagos zsírtartalom **24%**, szórása **3%**.

Tehát most

$$n_1 = 22, \qquad s_1 = 4, \qquad \overline{x_1} = 27,$$
$$n_2 = 25, \qquad s_2 = 3, \qquad \overline{x_2} = 24.$$

Tehát $F = 16/9 = 1.78$, a szabadsági fok pedig $(21, 24)$.

A $(21, 24)$ szabadsági fokhoz és az 5%-os szignifikanciaszinthez tartozó F érték **2.31**. (Vigyázat, most kétoldali próbáról van szó! Ezért a táblázatból a 2.5%-os szignifikanciaszinthez tartozó értéket kell kikeresni!)

Mivel az általunk számolt érték ennél kisebb, elfogadjuk a nullhipotézist, azaz azt, hogy a két termelő által szállított hús zsírtartalmának szórásnégyzete megegyezik.

A z- és t-próbákhoz hasonlóan ezt a vizsgálatot is el lehet végezni konfidenciaintervallumok segítségével. Ehhez ki kell számolni a két szórásnégyzet hányadosának 95%-os megbízhatósági szintű konfidenciaintervallumát, és megnézni, hogy az 1 beleesik-e vagy sem. Ha beleesik, akkor elfogadjuk a H_0-t, ha nem, akkor elutasítjuk.

2. Varianciaanalízis (ANOVA)

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 termése-redményt.

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.

2.1. A kezelések közötti és a kezeléseken belüli variancia

A varianciaanalízis logikáját a következő példán keresztül tekintjük át.

Példa.

Egy mezőgazdasági kutató intézetben néhány agrármérnök azt állította, hogy egy bizonyos növény jobban nő, ha egy bizonyos tápanyagot tartalmazó oldattal öntözik.

Kijelöltek két földdarabot (parcellát), amelyeken ezt a növényt termesztették, és az egyiken az oldattal öntöztek, a másikon tiszta vízzel. Azt az eredményt kapták, hogy az oldattal öntözött növények magasabbra nőttek, mint a tiszta vízzel öntözöttek.

Más kutatók megismételték a kísérletet, és arra a megállapításra jutottak, hogy az oldatnak semmiféle hatása nincs.

Az eredeti kísérlettel az volt a probléma, hogy két különböző parcellán termesztették a növényeket. Az általában minden ilyen jellegű kísérletre igaz, hogy

sohasem lehet teljesen egyforma körülményeket teremteni a kísérleti csoportoknak, ezen kívül a kísérletbe bevont egyedek is különbözőek. Az a parcella, amelyen az oldattal öntöztek, valószínűleg jobb minőségű volt, mint a másik.

Az lenne az ideális eset, ha a kísérlethez teljesen egyforma körülményeket lehetne teremteni. Ez a gyakorlatban kivitelezhetetlen, úgyhogy meg kell becsülnünk, hogy **mekkora a termelési körülményekből adódó szórás**. Ha ezt meg tudjuk tenni, akkor meg tudjuk mondani, hogy a tiszta vízzel illetve az

oldattal öntözött növények közötti különbség mértéke lényegesen nagyobb-e a más termelési viszonyokból adódó különbségeknél.

Ahhoz, hogy elkerülhessük az előbb vázolt kísérletben elkövetett hibát, célszerű a kísérletet többször megismételni, véletlen mintákat választani.

Tegyük fel, hogy van 12 növényünk különböző helyeken. **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 ön-

tözővizet 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. Abban reménykedünk, hogy ily módon az egyéb faktorok (termelési különbségek, növények kondíciója) kiegyensúlyozzák egymást.

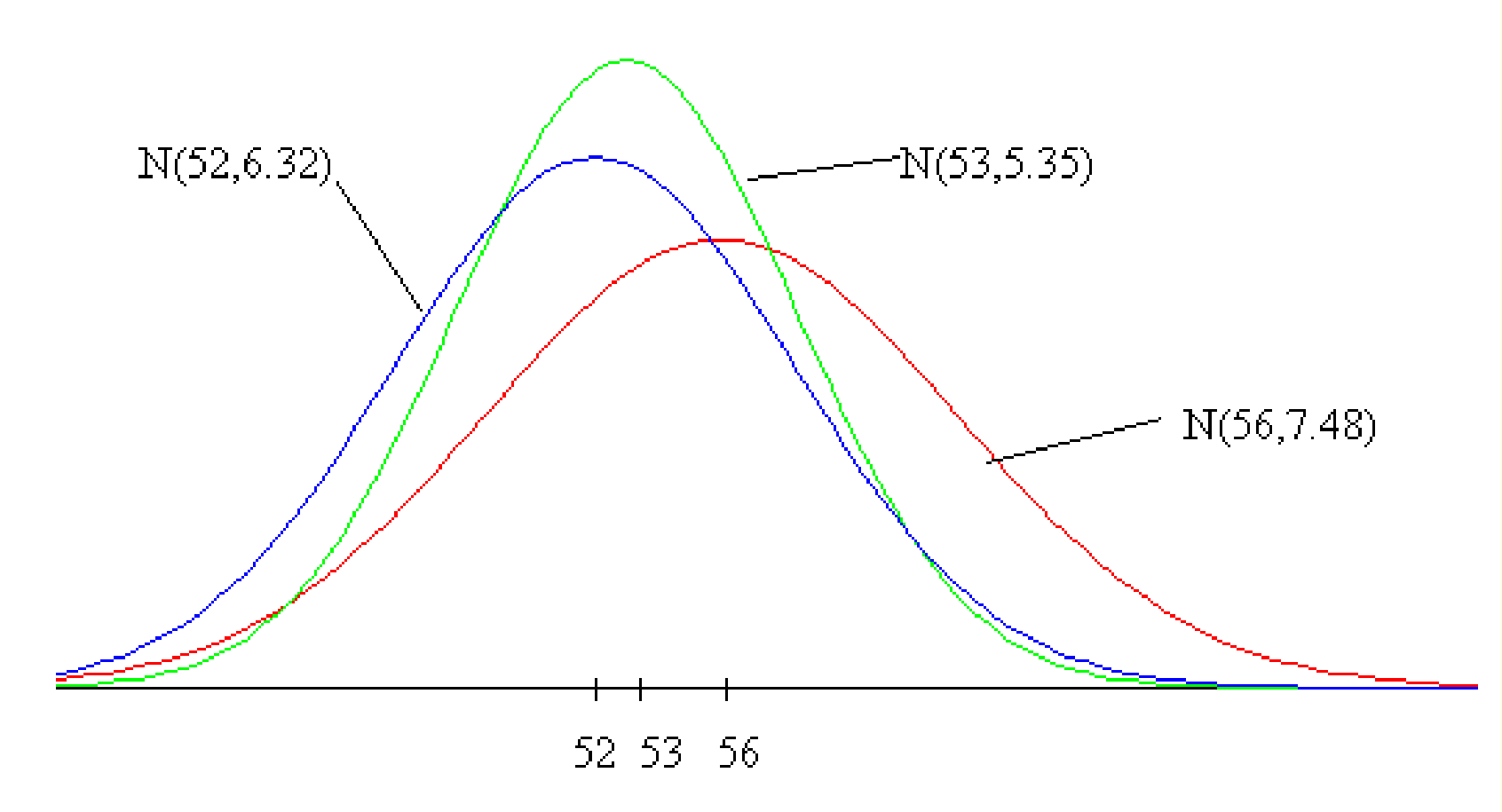
Természetesen előfordulhat olyan eset, hogy az oldattal öntözött növények jobb körülmények között nőnek mint a többiek, és megint félrevezető eredményt kapunk. A statisztikai procedúránk éppen arra jó, hogy ki lehessen számolni az ilyen esetek valószínűségét.

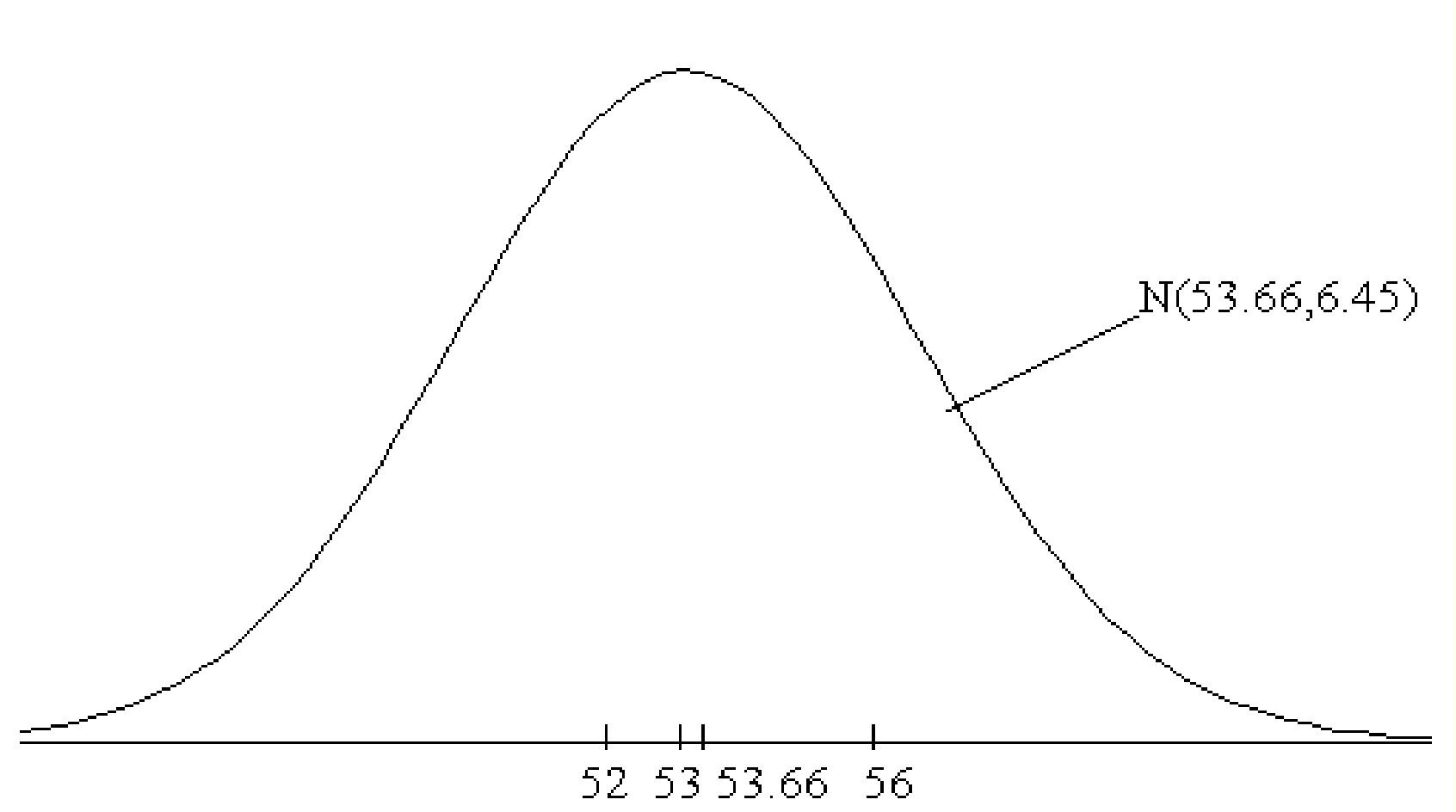
Nézzük meg a statisztikai analízist, ha a kísérletben a növények magassága (cm-ben mérve) az alábbiak szerint alakult:

	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.


N(52,6.32)
N(53,5.35)
N(56,7.48)
52
53
56



Ha formálisan felírjuk a minták elemeinek magasságát:

tömény oldat: $Y_1 = \mu_1 + \varepsilon$,
híg oldat: $Y_2 = \mu_2 + \varepsilon$,
víz: $Y_3 = \mu_3 + \varepsilon$,

ahol μ_1, μ_2, μ_3 a populációk átlaga, ε pedig a véletlen hiba, mely a különböző körülményekből, és a növények különböző kondíciójából származik.

Ha az a magyarázat felel meg a valóságnak, hogy a minták **ugyanabból** a populációból származnak (azaz az oldat nem okoz változást a növények magasságában), akkor $\mu_1 = \mu_2 = \mu_3$. A megfigyelt különbségek csupán a véletlen mintavétel eredményei.

Melyik magyarázat igaz?

Először is meg kell néznünk, hogy az egyes növények között mekkora különbségek vannak. Ha elég kicsik, akkor az átlagok közötti különbségek elég meggyőző bizonyítékok arra, hogy a minták különböző populációkból származnak. Ha viszont nagyok a különbségek, akkor éppen fordított a helyzet.

Először számítsuk ki, hogy mekkora a szórásnégyzet abban az esetben, ha **egy populációról** van szó. Ezt a varianciát a minták szórásnégyzetének átlagával becsülhetjük. (Ez ugyanazt az eredményt adja, mintha összeöntenénk a mintákat, és az így kapott nagyobb

mintának számítanánk ki a szórásnégyzetét.)

Így a növények közötti szórásnégyzet becslése: $(56+40+28.66)/3 = 41.55$. Ekkor a szórás: **6.45**.

Most már el tudjuk dönteni, hogy a minták átlagai (56, 53, 52) közti különbségek mekkorák ehhez a szóráshoz képest.

Tudjuk, hogy ha egy μ átlagú és σ szórású normális eloszlású populációból kiválasztunk véletlenszerűen egy n elemű mintát, akkor a mintaátlag eloszlása: $\overline{X} \sim N(\mu, \sigma^2/n)$. A szórása: $\sigma/\sqrt{n}$.

Esetünkben 4 elemű mintáról van szó, így a mintaátlag

szórásnégyzete: $41.55/4 = 10.38$, szórása **3.22**.

Most már látható számszerűen is, hogy a mintaátlagok közötti különbség (4) nem nagyon meggyőző, hiszen a mintaátlag szórása 3.22. Ha a **mintaátlagok szórásnégyzetét** becsüljük a mintákból, akkor

$$\frac{(54-53.6)^2+(53-53.6)^2+(52-53.6)^2}{3-1} = \frac{3.08}{2} = 1.54.$$

Itt 53.6 a mintaátlagok átlaga, a szabadsági fok $k-1$, ahol k most a kezelések, vagy minták számát

jelenti.)

Ahogy az F-próbánál tettük, most is meg kell néznünk, hogy a két szórásnégyzet hányadosa mennyire van közel az 1-hez. Esetünkben:

$$F = 1.54/10.38 = 0.148.$$

A szabadsági fok: $(3-1, 3\cdot(4-1))$, azaz $(2, 9)$.

Az 5%-os szignifikanciaszinthez és ehhez a szabadsági fokhoz tartozó táblázatbeli F érték: **4.26**.

A 0.148 jóval kisebb ennél az értéknél, így elfogadjuk a nullhipotézist, hogy a minták ugyanabból a populációból származnak.

Fogalmazzuk meg általánosan az eddig elmondottakat:

1. Van egy valószínűségi változónk és több csoportunk.
2. Feltételezzük, hogy a változó eloszlása minden csoportban normális, és a szórása azonos. A várható érték nem feltétlenül ugyanaz minden csoportban.
3. Dönteni szeretnénk a következő nullhipotézisről (H_0): a várható érték minden csoportban azonos.
4. Elvégezzük a varianciaanalízist.

Tehát a null- és az alternatív hipotézis:

H_0 : a populációátlagok **megegyeznek**,

H_1 : a populációátlagok **nem egyeznek meg**.

A próbastatisztika:

$$\mathbf{F} = \frac{\text{a mintaátlagok szórásnégyzete}}{\text{a minta szórásnégyzetek átlaga}}.$$

A szabadsági fok: $(k-1, k \cdot (n-1))$,

ahol k a kezelések vagy csoportok száma, n pedig az egyes minták elemszáma.

2.2. Az ANOVA táblázat

A varianciaanalízist nemcsak az előbb elmondottak alapján lehet elvégezni. A következő formalizmust (amely lényegileg ugyanaz, mint az előző fejezetben leírtak), azért tekintjük át, mert a számítógépes programcsomagok általában ilyen formában adják a varianciaanalízis eredményét.

Az egyszerűség kedvéért tételezzük fel, hogy az egyes

kezeléseket **ugyanakkora mintákon** végeztük el. Legyen a kezelések száma k, az egyes minták elemszáma pedig n. Jelöljük $Y_{i,j}$-vel a mért eredményeket, ahol i a kezelés sorszáma, j pedig a mintán belüli sorszám. Az adatokat táblázatba rendezve:

minta-elem \ kezelés	1	2	...	k
1	$Y_{1,1}$	$Y_{2,1}$	...	$Y_{k,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$Y_{1,n}$	$Y_{2,n}$	...	$Y_{k,n}$
átlag	$\overline{Y_1}$	$\overline{Y_2}$	...	$\overline{Y_k}$
variancia	s_1^2	s_2^2	...	s_k^2

$\overline{Y}_i$ az i-edik kezelést kapott minta átlaga, s_i^2 pedig a szórásnégyzete, azaz

$$\overline{Y}_i = \frac{\sum_{j=1}^{n} Y_{ij}}{n},$$

$$s_i^2 = \frac{\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_i\right)^2}{n-1}.$$

Az ún. fő átlag:

$$\overline{Y} = \frac{\overline{Y_1} + \overline{Y_2} + \ldots + \overline{Y_k}}{k}.$$

A teljes eltérés négyzetösszeg:

$$SQ_T = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(Y_{i,j} - \overline{Y}\right)^2 .$$

$n\cdot$(a mintaátlagok négyzetes eltérése a főátlagtól):

$$SQ_K = n \cdot \sum_{i=1}^{k} \left(\overline{Y}_i - \overline{Y}\right)^2 .$$

A mintaelemeknek a mintaátlagtól való négyzetes

eltérés összege:

$$SQ_B = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(Y_{i,j} - \overline{Y}_i\right)^2 .$$

Bebizonyítható, hogy igaz a következő egyenlőség:

$$\boxed{SQ_T = SQ_K + SQ_B.}$$

Az SQ_T az összes mért érték közötti különbözőséget méri. Az SQ_K azt, hogy mennyire változik a mért mennyiség kezelésről kezelésre. Ezért ezt a **kezelések közötti eltérés négyzetösszegnek** nevezzük.

Végül SQ_B az egyes kezeléseken belüli változatosságot méri, így ezt **kezelésen belüli eltérés négyzetösszegnek** nevezzük. Az eddig elmondottakat és a varianciaanalízis további részeit általában egy ún. **ANOVA táblázatban** szokták összefoglalni:

A variancia eredete	Eltérés-négyzetösszeg	szabadsági fok	átlagos négyzetes eltérés (variancia)	F
Kezelések közötti	SQ_K	k-1	$MQ_K = \frac{SQ_K}{k-1}$	
Kezelésen belüli	SQ_B	$k(n$-1)	$MQ_B = \frac{SQ_B}{k(n-1)}$	$F = \frac{MQ_K}{MQ_B}$
Teljes	SQ_T	nk-1		

Az első oszlopban az szerepel, hogy milyen eltérésnégyzetösszegről, illetve varianciáról van szó. Ennek megfelelően a második oszlopban az előbb vázolt eltérésnégyzetösszegek szerepelnek, a harmadikban pedig a hozzájuk tartozó szabadsági fokok. A negyedikben szereplő értékek a már ismert mintaátlagok szórásnégyzetei szorozva n-nel, illetve a minta varianciák átlaga. A továbbiakban a már említett módon kell eljárni.

Példánkban:

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.