Statisztika
Hipotézisvizsgálat

Székely Balázs

2010. december 2.

Előadás vázlat

Intervallumbecslések

- Egy érték helyett egy intervallumot adunk a paraméter becslésére.
- Az intervallumbecslés a hipotézisvizsgálat alapja.

Intervallumbecslések

Adott $\mathbf{P}_\theta$, $\theta \in \Theta$ eloszlás család. Adott egy $\underline{X} = (X_1, X_2, \ldots, X_n)$ független minta a $\mathbf{P}_\theta$ eloszlásból. Adott a paraméter egy függvénye $\psi(\theta)$.

Definíció

A $(T_1(\underline{X}), T_2(\underline{X}))$ statisztika párral definiált intervallum legalább $1-\varepsilon$ szintű konfidencia intervvalum a $\psi(\theta)$ paraméterre, ha

$$\mathbf{P}_\theta\left(T_1(\underline{X}) < \psi(\theta) < T_2(\underline{X})\right) \geq 1-\varepsilon \qquad \forall \theta \in \Theta,$$

ahol $\varepsilon > 0$ kicsi szám.
$1-\varepsilon$ neve konfidenciaszint.

Ha a $\mathbf{P}_\theta$-k folytonos eloszások, akkor lehet pontosan $1-\varepsilon$ szintű konfidenciaintervallumról beszélni:

$$\mathbf{P}_\theta\left(T_1(\underline{X}) < \psi(\theta) < T_2(\underline{X})\right) = 1-\varepsilon \qquad \forall \theta \in \Theta,$$

Konfidenciaintervallum jelentése

Legyen most $\psi(\theta) = \theta$, tehát a paraméterre konstruálunk egy 95% szintű konfidenciaintervallumot:

$$\mathbf{P}_\theta\left(T_1(\underline{X}) < \theta < T_2(\underline{X})\right) = 0,95 \qquad \forall\theta \in \Theta,$$

Vegyünk nagyon sok, mondjuk M darab, n elemű mintát. Mindegyikhez készítsük el a $(T_1(\underline{x}), T_2(\underline{x}))$ intervallumot. Ez M darab itervallum.

Fontos tény

- *95%-os konfidenciaszint jelentése: az adatsorok 95%-ban ($0,95 \cdot M$ esetben) $\theta \in (T_1(\underline{x}), T_2(\underline{x}))$*
- *viszont az adatsorok 5%-ban $\theta \notin (T_1(\underline{x}), T_2(\underline{x}))$.*

Konfidenciaintervallum

Példa

Adott egy 30 elemű minta $\mathcal{N}(m, 20)$ eloszlásból: 281, 308, 300, 285, 294, 272, 302, 301, 306, 297, 279, 286, 302, 316, 286, 296, 293, 291, 297, 301, 311, 293, 286, 286, 297, 305, 278, 300, 301, 300.
Szerkesszünk az ismeretlen várható értékre 95% szintű konfidenciaintervallumot.

Konfidenciaintervallum

A normális eloszlás várható értékére ismert szórás esetén

Legyen $X_1, \ldots, X_n \sim \mathcal{N}(m, \sigma_0)$ független minta. σ_0 ismert, m ismeretlen.

Az ismeretlen várható értékre szeretnénk $1-\varepsilon$ szintű konfidenciaintervallumot szerkeszteni.

Tudjuk, hogy $\overline{X}_n$ torzítatlan, konzisztens becslése a várható értéknek, m-nek. Mivel a normális eloszlás szimmetrikus a várható értékre, ezért az intervallumot $\left(\overline{X}_n - r_\varepsilon, \overline{X}_n + r_\varepsilon\right)$ alakban keressük. Az $1-\varepsilon$ szint azt jelenti, hogy

$\mathbf{P}_m(\overline{X}_n - r_\varepsilon < m < \overline{X}_n + r_\varepsilon) = 1 - \varepsilon$. Alakítsjuk az eseményt:

$$\mathbf{P}_m(\overline{X}_n - r_\varepsilon < m < \overline{X}_n + r_\varepsilon) = \mathbf{P}_m(-r_\varepsilon < \overline{X}_n - m < r_\varepsilon) =$$

$$\mathbf{P}_m\left(\frac{-r_\varepsilon}{\sigma_0}\sqrt{n} < \frac{\overline{X}_n - m}{\sigma_0}\sqrt{n} < \frac{r_\varepsilon}{\sigma_0}\sqrt{n}\right) =$$

$$\frac{\overline{X}_n - m}{\sigma_0}\sqrt{n} = \frac{X_1 + \cdots + X_n - mn}{n\sigma_0}\sqrt{n} = \frac{(X_1 + \cdots + X_n) - nm}{\sqrt{n}\sigma_0} \sim \mathcal{N}(0,1)$$

$$= \Phi\left(\frac{r_\varepsilon}{\sigma_0}\sqrt{n}\right) - \Phi\left(\frac{-r_\varepsilon}{\sigma_0}\sqrt{n}\right) = 2\Phi\left(\frac{r_\varepsilon}{\sigma_0}\sqrt{n}\right) - 1 = 1 - \varepsilon$$

$\Phi\left(\frac{r_\varepsilon}{\sigma_0}\sqrt{n}\right) = 1 - \frac{\varepsilon}{2}$. Így $u_{\varepsilon/2} := \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right)$. $r_\varepsilon = \frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}}$, tehát az

$1-\varepsilon$ szintű konfidencia itervallum: $\left(\overline{X}_n - \frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}}, \overline{X}_n + \frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}}\right)$

Konfidenciaintervallum

Példa

Adott egy 30 elemű minta $\mathcal{N}(m, 20)$ eloszlásból: 281, 308, 300, 285, 294, 272, 302, 301, 306, 297, 279, 286, 302, 316, 286, 296, 293, 291, 297, 301, 311, 293, 286, 286, 297, 305, 278, 300, 301, 300.

Szerkesszünk az ismeretlen várható értékre 95% szintű konfidenciaintervallumot. $\left(\overline{X}_n - \frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}}, \overline{X}_n + \frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}}\right)$

$n = 30$, $\varepsilon = 0,05$, $\sigma_0 = 20$

$\overline{X} = 295$,

$r_\varepsilon = \frac{\sigma_0}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right) = \frac{20}{\sqrt{30}}\Phi^{-1}(0,975) \approx 7,16$ Tehát:

$(\overline{x}_{30} - r_{0,05}, \overline{x}_{30} + r_{0,05}) = (295 - 7,16, 295 + 7,16) = (287,84, 302,16)$.

A konfidenciaintervallum jelentése: a 30 hosszú adatsorok 95%-ában az igazi paraméter beleesik a konfidenciaintervallumba, 5%-ában pedig nem.

Előadás vázlat

Feladat kitűzése, hipotézisek konstruálása

Kiindulási példa

Egy fejlesztő azt állítja, hogy új, energiatakarékos, akkumulátoros fűnyírót konstruált, amelyek 5 óráig (300 perc) képesek futni. Ennek ellenőrzésére 30 tesztet végeztünk a futási idők hosszúságára (percben): 281, 308, 300, 285, 294, 272, 302, 301, 306, 297, 279, 286, 302, 316, 286, 296, 293, 291, 297, 301, 311, 293, 286, 286, 297, 305, 278, 300, 301, 300.

Azt feltételezzük, hogy a futási idők normális eloszlásúak valamilyen, ismeretlen várható értékkel. Az egyszerűség kedvéért tegyük fel, hogy a szórás ismert, $\sigma_0 = 20$ perc.

A fejlesztő állítása: a normális eloszlás várható értéke 300.
Azt a hipotézist szeretnénk tesztelni, hogy az igazi várható érték, m 300-zal egyenlő:
$H_0 : m = m_0$, ahol $m_0 := 300$.
Ha H_0 nem igaz, akkor az alternatív hipotézis (vagy más néven ellenhipotézis) igaz:
$H_1 : m \neq m_0$

H_0 és H_1 lefedik az összes lehetséges esetet.

Döntések, döntések értelmezése

Kiindulási példa, konkrétan

A 30 elemes minta alapján kell döntenünk, hogy H_0 igaz, vagy H_1.

A döntésnél a cél: Ha H_0-t elutasítjuk, akkor annak jó oka legyen.

Következésképpen, annak van bizonyító ereje, ha H_0-t elutasítjuk, és H_1-et fogadjuk el.

Kiindulási példa

Legyen $X_1, X_2, \ldots, X_n$ egy független minta $\mathcal{N}(m, \sigma_0)$ eloszlásból, ahol σ_0 ismert, és m ismeretlen paraméter. Minket az m paraméter igazi értéke érdekel.

Legyen $x_1, x_2, \ldots, x_n$ a minta megvalósulása.
Jelölés: $\underline{X} = (X_1, X_2, \ldots, X_n)$, $\underline{x} = (x_1, x_2, \ldots, x_n)$.

Kiválasztunk egy m_0 értéket, és veszünk egy n elemű minta megvalósulását, $\underline{x}$-et az $\mathcal{N}(m, \sigma_0)$ eloszlásból.

Azt szeretnénk eldönteni az $\underline{x}$ megvalósulás alapján, hogy az **igazi paraméter,** m amelyikből a minta generálódott **megegyezik-e** m_0**-al az általunk adott értékkel**.

Döntések, döntések értelmezése

Kiindulási példa, konkrétan

A döntésnél a cél: Ha H_0-t elutasítjuk, akkor annak jó oka legyen.

Következésképpen, annak van bizonyító ereje, ha H_0-t elutasítjuk, és H_1-et fogadjuk el.

Döntés

Módszer: **1. lépés:** Vegyük egy torzítatlan, és konzisztens becslését az m paraméternek. Az $\overline{x}$ mintaátlag egy ilyen becslés.
2. lépés: Konstruáljunk eköré egy 95% szintű konfidenciaintervallumot: $(\overline{x} - u_{95\%}, \overline{x} + u_{95\%})$.

Ha $m_0 \in (\overline{x} - u_{95\%}, \overline{x} + u_{95\%})$, akkor elfogadjuk H_0-t 95%-os szinten.
Ha $m_0 \notin (\overline{x} - u_{95\%}, \overline{x} + u_{95\%})$, akkor mit döntsünk? Tudjuk, hogy a konfidenciaintervallum olyan, hogy az $\underline{x}$ minták 95%-ban az **igazi paraméter** beleesik, és 5%-ban, pedig nem. Ezért két opciónk van $m_0 \notin \ldots$ értelmezésére:

1. Az $\underline{x}$ minták 5%-ban azért az igazi paraméter nem esik bele a konf. intervallumba $\Rightarrow$ elképzelhető, hogy m_0 az igazi paraméter, de véletlenül nem esett bele.
2. m_0 nem az igazi paraméter, következésképpen H_0 nem igaz.

Mivel 1. igen valószínűtlen, csak az esetek 5%-ban fordul elő, ezért a 2. pontot vesszük következtetésnek. Tehát:

Ha $m_0 \notin (\overline{x} - u_{95\%}, \overline{x} + u_{95\%})$, akkor elvetjük H_0-t 95%-os szinten.

Döntések értelmezése, döntési hibák

Ezen döntési eljárás alapján, ha H_0-t elutasítjuk, akkor annak jó oka van.

Következésképpen, annak van bizonyító ereje, ha H_0-t elutasítjuk, azaz H_1-et fogadjuk el.

Így ha valamit bizonyítani szeretnénk, akkor azt az állítást H_1-be kell tenni.

H_0 elfogadásának nincs bizonyító ereje. Lehet, hogy csak azért fogadtuk el H_0-t, mert nincs elég adat.

Két hibázás lehet:

1. H_0 teljesül, de elvetjük, ez az elsőfajú hiba, ennek valószínűsége $\varepsilon = 1-$(szignifikancia szint). 95%-os szignifikancia szint esetén 5%.
2. H_0 nem teljesül, de elfogadjuk, ez a másodfajú hiba.

Feladat kitűzése, hipotézisek konstruálása

Kiindulási példa megoldása

Egy fejlesztő azt állítja, hogy új, energiatakarékos, akkumulátoros fűnyírót konstruált, amelyek 5 óráig (300 perc) képesek futni. Ennek ellenőrzésére 30 tesztet végeztünk a futási idők hosszúságára (percben): 281, 308, 300, 285, 294, 272, 302, 301, 306, 297, 279, 286, 302, 316, 286, 296, 293, 291, 297, 301, 311, 293, 286, 286, 297, 305, 278, 300, 301, 300.

Azt feltételezzük, hogy a futási idők normális eloszlásúak valamilyen, ismeretlen várható értékkel. Az egyszerűség kedvéért tegyük fel, hogy a szórás ismert, $\sigma_0 = 20$ perc.

A fejlesztő állítása: a normális eloszlás várható értéke 300.
Azt a hipotézist szeretnénk tesztelni, hogy az igazi várható érték, m 300-zal egyenlő:
$H_0 : m = m_0$, ahol $m_0 := 300$.
Ha H_0 nem igaz, akkor az alternatív hipotézis (vagy más néven ellenhipotézis) igaz:
$H_1 : m \neq m_0$

H_0 és H_1 lefedik az összes lehetséges esetet.

A példa megoldása

Hipotézisek
$H_0 : m = m_0$, ahol $m_0 := 300$.
$H_1 : m \neq m_0$

Konfidenciaintervallum szerkesztése. **Tegyük fel, hogy H_0 teljesül**, ekkor $X_1, \dots, X_n$ eloszlása $\mathcal{N}(m_0, \sigma_0)$. A konfidenciaszint $1 - \varepsilon$.
$n = 30$, $\sigma_0 = 20$, $1 - \varepsilon = 0,95 \Rightarrow \varepsilon = 0,05$.
Korábbról tudjuk, hogy a normális eloszlásra szerkesztett konfidenciaintervallum sugara

$$r := \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right) = \frac{20}{\sqrt{30}} \Phi^{-1}(0,975) \approx 7,16.$$

A megadott adatok átlaga 295.
Tehát a konfidenciaintervallum: $(\overline{x}_{30} - r, \overline{x}_{30} + r)$
$= (295 - 7,16, 295 + 7,16) = (287,84, 302,16)$. Ennek eleme $m_0 = 300$, így H_0-t nem tudjuk elutasítani, azaz el kell fogadni a fejlesztő állítását.

Formalizálva, teszt statisztikával

Legyen $X_1, X_2, \dots, X_n$ egy független minta $\mathcal{N}(m, \sigma_0)$ eloszlásból, ahol σ_0 ismert, és m ismeretlen paraméter. m_0 adott paraméter.
A hipotézisek:
$H_0 : m = m_0$
$H_1 : m \neq m_0$
Ha H_0 teljesül, azaz $X_1, X_2, \dots, X_n$ egy független minta $\mathcal{N}(m_0, \sigma_0)$ eloszlásból. Ekkor

H_0-t elfogadjuk $1 - \varepsilon$ szinten $\Leftrightarrow$
$m_0 \in \left(\overline{X}_n - \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right), \overline{X}_n + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right)\right)$
$\Leftrightarrow -\Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right) < \sqrt{n}\frac{\overline{X}_n - m_0}{\sigma_0} < \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right)$

$\sqrt{n}\frac{\overline{X}_n - m_0}{\sigma_0}$ neve: teszt statisztika.
Ha H_0 teljesül, akkor $\overline{X}_n \approx m_0$, tehát $\sqrt{n}\frac{\overline{X}_n - m_0}{\sigma_0}$ értéke 0-hoz közeli.
Ha H_0 nem teljesül, azaz $m \neq m_0$, akkor $\overline{X}_n \approx m \neq m_0$ tehát, $\sqrt{n}\frac{\overline{X}_n - m_0}{\sigma_0}$ értéke ∞-ehz, vagy $-\infty$-hez tart, ha n nő.
Így egy fix n-re akkor utasítjuk el H_0-t, ha $\sqrt{n}\frac{\overline{X}_n - m_0}{\sigma_0}$ a teszt statisztika értéke túl nagy abszolút értékben. A határ pont $\Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right)$.

Várható érték tesztelése, ha a szórás ismeretlen

Legyen $X_1, X_2, \ldots, X_n$ egy független minta $\mathcal{N}(m, \sigma)$ eloszlásból, ahol σ és m ismeretlen paraméterek.
Minket az m paraméter igazi értéke érdekel. Azt szeretnénk tesztelni, hogy m az megegyezik-e egy adott m_0 értékkel. A hipotézisek:
$H_0 : m = m_0$
$H_1 : m \neq m_0$
Ismert szórás esetén $\sqrt{n}\frac{\overline{X}_n - m_0}{\sigma_0}$ a teszt statisztika. Most a szórás ismeretlen. Becsüljük a szórást, vegyük egy torzítatlan, konzisztens becslését a szórásnak: $s_n^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2}$.

Tekintsük a $\sqrt{n}\frac{\overline{X}_n - m_0}{s_n^*}$ statisztikát. Ez lesz a teszt statisztika.

Ha H_0 teljesül, akkor $\overline{X}_n \approx m_0$, tehát $\sqrt{n}\frac{\overline{X}_n - m_0}{s_n^*}$ értéke 0-hoz közeli.

Ha H_0 nem teljesül, azaz $m \neq m_0$, akkor $\overline{X}_n \approx m \neq m_0$, tehát $\sqrt{n}\frac{\overline{X}_n - m_0}{s_n^*}$ értéke végtelenhez tart, ha n nő.

Így egy fix n-re akkor utasítjuk el H_0-t, ha $\sqrt{n}\frac{\overline{X}_n - m_0}{s_n^*}$ a teszt statisztika értéke túl nagy abszolút értékben.

Mi az elutasítás határa, a kritikus érték?

Várható érték tesztelése, ha a szórás ismeretlen

t eloszlások

Tekintsük a $\sqrt{n}\frac{\overline{X}_n - m_0}{s_n^*}$ statisztikát. Ez lesz a teszt statisztika.
Mi az elutasítás határa, a kritikus érték?

Állítás

Ha $X_1, \ldots, X_n$ független, azonos, $\mathcal{N}(m_0, \sigma)$ eloszlású (m_0!), akkor

$$\sqrt{n}\frac{\overline{X}_n - m_0}{s_n^*}$$ *eloszlásának neve:*

t eloszlás $n-1$ szabadsági fokkal.

Minden t eloszlás szimmetrikus. Az eloszlásfüggvényére,
$T_n : x \mapsto \mathbf{P}\left(\sqrt{n}\frac{\overline{X}_n - m_0}{s_n^*} < x\right) =: T_n(x)$ függvényre *táblázat van*.
$1-\varepsilon$ szignifikancia szintre a kritikus érték, κ kiszámolása:
$\mathbf{P}(-\kappa < \sqrt{n}\frac{\overline{X}_n - m_0}{s_n^*} < \kappa) = 1-\varepsilon$. Tehát a b.o.
$= T_{n-1}(\kappa) - T_{n-1}(-\kappa) =$, mivel a t eloszlás szimmetrikus,
$= T_{n-1}(\kappa) - (1 - T_{n-1}(\kappa)) = 2T_{n-1}(\kappa) - 1 = 1-\varepsilon$. Átrendezve:
$T_{n-1}(\kappa) = 1 - \varepsilon/2$. $\kappa = T_{n-1}^{-1}(1-\varepsilon/2)$. ÁBRA!
Így H_0-t elfogadjuk $1-\varepsilon$ szinten $\Leftrightarrow -\kappa < \sqrt{n}\frac{\overline{X}_n - m_0}{s_n^*} < \kappa$

Táblázat *t* eloszlások farok valószínűségére

PERCENTAGE POINTS OF THE T DISTRIBUTION

Tail Probabilities

	One Tail	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
D	1	3.078	6.314	12.71	31.82	63.66	318.3	637
E	2	1.886	2.920	4.303	6.965	9.925	22.330	31.6
G	3	1.638	2.353	3.182	4.541	5.841	10.210	12.92
R	4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
E	5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
E	6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
S	7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
	8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
O	9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
F	10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
	11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
F	12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
R	13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
E	14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
E	15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
D	16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
O	17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
M	18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
	19	1.328	1.729	2.093	2.539	2.861	3.579	3.883

Példa

6 darab csapágy belső gyűrűjének átmérőjét mérjük az A és a B mérőműszeren. A következő mérési eredményeket kapjuk:

csapágy	1.	2.	3.	4.	5.	6.
A műszer	6,0	10,1	8,0	13,0	12,0	9,2
B műszer	6,2	9,9	8,0	12,9	11,7	9,0

(Az adatokat normális eloszlásból származónak feltételezzük.) Teszteljük, mutat-e a két műszeren mért érték 95%-os szinten szignifikáns eltérést.

A különbségek: -0,2; 0,2; 0; 0,1; 0,3; 0,2.
A feladat szerint az szignifikáns, ha eltérnek. Tehát az eltérést kell a H_1-be rakni.
Tehát tegyük fel, hogy a különbségek $\mathcal{N}(m, \sigma)$ eloszlásból vett független minták, ahol σ ismeretlen. Így a hipotézisek:
$H_0 : m = 0$
$H_1 : m \neq 0$

Példa

$(x_1; \dots; x_6) =$ (-0,2; 0,2; 0; 0,1; 0,3; 0,2).

A teszt statisztika $t_5 = \sqrt{6}\dfrac{\overline{X}_6 - 0}{s_6^*}$ $(m_0 = 0)$

$\overline{x}_6 = 0,1$, $s_6^* = 0,1789$. Tehát $t_5 = 1,369$. Ezt kell összehasonlítani a 95%-os szignifikancia szinthez tartozó κ kritikus értékkel:

H_0-t elfogadjuk $\Leftrightarrow |t_5| < \kappa$.

95% szignifikancia szint esetén $\varepsilon = 0,05$.

Korábbi dia: a kritikus érték $\kappa = T_5^{-1}(1 - \varepsilon/2) = T_5^{-1}(0,975)$.

Így az 5 szabadságfokú t eloszlás 0,025 valószínűségű farkát keressük. A táblázatból $T_5^{-1}(0,975) =$

Táblázat t eloszlások farok valószínűségére

PERCENTAGE POINTS OF THE T DISTRIBUTION

Tail Probabilities

One Tail		0.10	0.05	0.025	0.01	0.005	0.001	0.0005
D	1	3.078	6.314	12.71	31.82	63.66	318.3	637
E	2	1.886	2.920	4.303	6.965	9.925	22.330	31.6
G	3	1.638	2.353	3.182	4.541	5.841	10.210	12.92
R	4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
E	5	1.476	2.015	$\boxed{2.571}$	3.365	4.032	5.893	6.869
E	6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
S	7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
	8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
O	9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
F	10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
	11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
F	12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
R	13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
E	14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
E	15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
D	16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
O	17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
M	18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
	19	1.328	1.729	2.093	2.539	2.861	3.579	3.883

Példa

$(x_1; \ldots; x_6) =$ (-0,2; 0,2; 0; 0,1; 0,3; 0,2).

A teszt statisztika $t_5 = \sqrt{6}\dfrac{\overline{X}_6 - 0}{s_6^*}$ $\quad (m_0 = 0)$

$\overline{x}_6 = 0,1$, $s_6^* = 0,1789$. Tehát $t_5 = 1,369$. Ezt kell összehasonlítani a 95%-os szignifikancia szinthez tartozó κ kritikus értékkel:

H_0-t elfogadjuk $\Leftrightarrow |t_5| < \kappa$.

95% szignifikancia szint esetén $\varepsilon = 0,05$.

Korábbi dia: a kritikus érték $T_5^{-1}(1 - \varepsilon/2) = T_5^{-1}(0,975)$.

Így az 5 szabadságfokú t eloszlás 0,025 valószínűségű farkát keressük.

A táblázatból $T_5^{-1}(0,975) = 2,571$.

Mivel $t_5 = 1,369 < 2,571 = \kappa$, ezért nem tudjuk elutasítani H_0-t, azaz nem tudjuk bizonyítani, hogy az A és a B gépek különböző pontossággal működnének.

Előadás vázlat

1. Intervallumbecslések
   - Általánosan
   - Példa
2. Paraméteres próbák
   - Kiindulási példa
   - Várható érték tesztelése, ha a szórás ismert
   - Várható érték tesztelése, ha a szórás ismeretlen
3. Hipotézisek fajtái
   - Egy- és kétoldali ellenhipotézis
   - Egymintás és kétmintás próbák
4. Nemparaméteres próbák
   - Khi négyzet eloszlás
   - Illeszkedés vizsgálat
   - Homogenitás vizsgálat
   - Függetlenség vizsgálat

Egyoldali és kétoldali ellenhipotézis

1. példa. 6 darab csapágy belső gyűrűjének átmérőjét mérjük az A és a B mérőműszeren. A következő mérési eredményeket kapjuk:

csapágy	1.	2.	3.	4.	5.	6.
A műszer	6,0	10,1	8,0	13,0	12,0	9,2
B műszer	6,2	9,9	8,0	12,9	11,7	9,0

(Az adatokat normális eloszlásból származónak feltételezzük.) Teszteljük, mutat-e a két műszeren mért érték 95%-os szinten szignifikáns eltérést.

2. példa. Az alábbi két minta 5 autó fogyasztási adatait tartalmazza. Az első sorban a szerviz előtti, a második sorban a szerviz utáni értékek találhatók.

szerviz előtt	7,9	8,1	8,8	7,2	6,0
szerviz után	7,5	7,5	8,1	7,2	5,7

Döntsünk 95%-os szignifikancia szinten, hogy a szerviz csökkentette-e a fogyasztást.

Kétoldali ellenhipotézis

1. példa. 6 darab csapágy belső gyűrűjének átmérőjét mérjük az A és a B mérőműszeren. A következő mérési eredményeket kapjuk:

csapágy	1.	2.	3.	4.	5.	6.
A műszer	6,0	10,1	8,0	13,0	12,0	9,2
B műszer	6,2	9,9	8,0	12,9	11,7	9,0

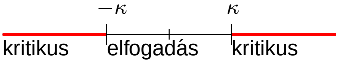
(Az adatokat normális eloszlásból származónak feltételezzük.) Teszteljük, mutat-e a két műszeren mért érték 95%-os szinten szignifikáns eltérést.

Mérések különbségéről feltételezzük, hogy $\mathcal{N}(m, \sigma)$ eloszlású, ahol m és σ ismeretlenek. A hipotézisek:

$H_0 : m = 0$

$H_1 : m \neq 0.$

Döntés: H_0-t elfogadjuk $\Leftrightarrow |t_5| < \kappa$,



azaz

H_1-t elfogadjuk $\Leftrightarrow \kappa < t_5$ vagy $t_5 < -\kappa$.

Egyoldali ellenhipotézis

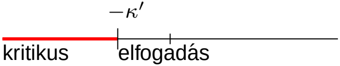
2. példa. Az alábbi két minta 5 autó fogyasztási adatait tartalmazza. Az első sorban a szerviz előtti, a második sorban a szerviz utáni értékek találhatók.

szerviz előtt	7,9	8,1	8,8	7,2	6,0
szerviz után	7,5	7,5	8,1	7,2	5,7

Döntsünk 95%-os szignifikancia szinten, hogy a szerviz csökkentette-e a fogyasztást.

Mérések különbségéről (után–előtt), -0,4 -0,6 -0,7 0 -0,3
feltételezzük, hogy $\mathcal{N}(m, \sigma)$ eloszlású, ahol m és σ ismeretlenek. A hipotézisek:
$H_0 : m \geq 0$
$H_1 : m < 0.$ $\leftarrow$Szignifikáns: a fogyasztás csökkent.

Döntés: H_0-t elfogadjuk $\Leftrightarrow -\kappa' < t_4$,



azaz H_1-t elfogadjuk $\Leftrightarrow t_4 < -\kappa$.

Egyoldali ellenhipotézis

A 2. példa megoldása

Mivel a szórás ismeretlen, t-próbát kell alkalmazni. A próbastatisztika értéke
$t_4 = \sqrt{5}\frac{\overline{x}_5}{s_5^*} = \sqrt{5}\frac{-0,4}{0,075} = -3,27$.

A 4 szabdságfokú t eloszlás 95%-os szignifikancia értékhez ($=$ $\varepsilon = 0,05$ farok valószínűséghez) tartozó kritikus értéke $\kappa' = 2,132$.

Egyoldali ellenhipotézis

A 2. példa megoldása

PERCENTAGE POINTS OF THE T DISTRIBUTION

Tail Probabilities								
One Tail		0.10	0.05	0.025	0.01	0.005	0.001	0.0005
D	1	3.078	6.314	12.71	31.82	63.66	318.3	637
E	2	1.886	2.920	4.303	6.965	9.925	22.330	31.6
G	3	1.638	2.353	3.182	4.541	5.841	10.210	12.92
R	4	1.533	$\boxed{2.132}$	2.776	3.747	4.604	7.173	8.610
E	5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
E	6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
S	7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
	8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
O	9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
F	10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
	11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
F	12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
R	13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
E	14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
E	15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
D	16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
O	17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
M	18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
	19	1.328	1.729	2.093	2.539	2.861	3.579	3.883

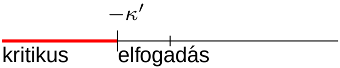
Egyoldali ellenhipotézis

A 2. példa megoldása

Mivel a szórás ismeretlen, t-próbát kell alkalmazni. A próbastatisztika értéke

$$t_4 = \sqrt{5}\frac{\overline{x}_5}{s_5^*} = \sqrt{5}\frac{-0,4}{0,075} = -3,27.$$

A 4 szabadságfokú t eloszlás 95%-os szignifikancia értékhez ($= \varepsilon = 0,05$ farok valószínűséghez) tartozó kritikus értéke $\kappa' = 2,132$.



Tehát
$t_4 = -3,27 < -2,132 = -\kappa$. Így elutasítjuk H_0-t, azaz a szerviz javított a fogyasztáson.
Figyeljük meg, hogy 97,5%-os szinten is szignifikáns a szerviz hatása, viszont 99%-os szinten már nem.

Egyoldali ellenhipotézis

A 2. példa megoldása, 97,5%, 99% szignifikancia szintek

PERCENTAGE POINTS OF THE T DISTRIBUTION

Tail Probabilities

$t_4 = -3,27$

One Tail		0.10	0.05	0.025	0.01	0.005	0.001	0.0005
D	1	3.078	6.314	12.71	31.82	63.66	318.3	637
E	2	1.886	2.920	4.303	6.965	9.925	22.330	31.6
G	3	1.638	2.353	3.182	4.541	5.841	10.210	12.92
R	4	1.533	2.132	$\boxed{2.776}$	$\boxed{3.747}$	4.604	7.173	8.610
E	5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
E	6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
S	7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
	8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
O	9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
F	10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
	11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
F	12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
R	13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
E	14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
E	15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
D	16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
O	17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
M	18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
	19	1.328	1.729	2.093	2.539	2.861	3.579	3.883

Egymintás és kétmintás próbák

A példa. Az alábbi két minta 5 autó fogyasztási adatait tartalmazza. Az első sorban a szerviz előtti, a második sorban a szervíz utáni értékek találhatók.

szerviz előtt	7,9	8,1	8,8	7,2	6,0
szerviz után	7,5	7,5	8,1	7,2	5,7

Döntsünk 95%-os szignifikancia szinten, hogy a szerviz csökkentette-e a fogyasztást.

B példa. Az alábbi két minta két különböző gyáregységben tapasztalt selejt-arányra vonatkozik (ezrelékben). Állítható-e 95%-os szignifikancia szinten, hogy az “A” gyáregység jobban dolgozott?

A	11,9	12,1	12,8	12,2	12,5	11,9	12,5	11,8	12,4	12,9
B	12,1	12,0	12,9	12,2	12,7	12,6	12,6	12,8	12,0	13,1

Egymintás próbák

A példa. Az alábbi két minta 5 autó fogyasztási adatait tartalmazza. Az első sorban a szerviz előtti, a második sorban a szervíz utáni értékek találhatók.

szerviz előtt	7,9	8,1	8,8	7,2	6,0
szerviz után	7,5	7,5	8,1	7,2	5,7

Döntsünk 95%-os szignifikancia szinten, hogy a szerviz csökkentette-e a fogyasztást.

Egymintás, hiszen **egy mintánk van a különbségekre**,
-0,4 -0,6 -0,7 0 -0,3 $\mathcal{N}(m, \sigma)$ eloszlásból.

Kétmintás próbák

B példa. Az alábbi két minta két különböző gyáregységben tapasztalt selejt-arányra vonatkozik (ezrelékben). Állítható-e 95%-os szignifikancia szinten, hogy az “A” gyáregység jobban dolgozott?

A	11,9	12,1	12,8	12,2	12,5	11,9	12,5	11,8	12,4	12,9
B	12,1	12,0	12,9	12,2	12,7	12,6	12,6	12,8	12,0	13,1

Az első minta: $X_1, X_2, \dots, X_{n_1}$ az $\mathcal{N}(m_1, \sigma_1)$ eloszlásból.
A második minta: $Y_1, Y_2, \dots, Y_{n_2}$ az $\mathcal{N}(m_2, \sigma_2)$ eloszlásból.
m_1, m_2, σ_1, σ_2 ismeretlen paraméterek. Tegyük fel, hogy $\sigma_1 = \sigma_2$ és X_i-k függetlenek Y_j-től.
A hipotézisek kétoldali ellenhipotézis esetén:
$H_0 : m_1 = m_2$
$H_1 : m_1 \neq m_2$
A hipotézisek egyoldali ellenhipotézis esetén (ez kell a B-ben):
$H_0 : m_1 \geq m_2$
$H_1 : m_1 < m_2$ (A-ban kevesebb a selejt)

Kétmintás próbák

Feltettük, hogy a két minta független és a szórásuk ugyanaz.
A teszt statisztika

$$t_{n_1+n_2-2} = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{(n_1-1)s_X^{*2} + (n_2-1)s_Y^{*2}}} \sqrt{\frac{n_1 n_2 (n_1+n_2-2)}{n_1+n_2}}$$

Állítás

Ha H_0 teljesül, azaz X_i-k és Y_j-k ugyanabból az $\mathcal{N}(m,\sigma)$ eloszlásból valók ($m = m_1 = m_2$), akkor $t_{n_1+n_2-2}$ t eloszlású n_1+n_2+2 szabadsági fokkal.

Kétmintás próbák, egy- és kétoldali ellenhipotézissel

$$t_{n_1+n_2-2} = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{(n_1-1)s_X^{*2} + (n_2-1)s_Y^{*2}}} \sqrt{\frac{n_1 n_2 (n_1+n_2-2)}{n_1+n_2}}$$

Kétoldali ellenhipotézis esetén: Ha H_0 teljesül ($m_1 = m_2$), akkor $\overline{X}_{n_1} \approx \overline{Y}_{n_2}$, tehát $t_{n_1+n_2-2}$ értéke 0-hoz közeli.
Ha H_0 nem teljesül, azaz $m_1 \neq m_2$, akkor $\overline{X}_{n_1} \not\approx \overline{Y}_{n_2}$ tehát, $t_{n_1+n_2-2}$ értéke végtelenhez tart, ha n nő.
Így egy fix n-re akkor utasítjuk el H_0-t, ha $t_{n_1+n_2-2}$ a teszt statisztika értéke túl nagy abszolút értékben.
Tehát: H_0-t elfogadjuk $\Leftrightarrow |t_{n_1+n_2-2}| < \kappa$

Egyoldali ellenhipotézis esetén: Ha H_0 teljesül ($m_1 \geq m_2$), akkor $\overline{X}_{n_1} \gtrapprox \overline{Y}_{n_2}$, tehát $t_{n_1+n_2-2}$ értéke $> -\kappa'$.
Ha H_0 nem teljesül, azaz $m_1 < m_2$, akkor $\overline{X}_{n_1} \lnapprox \overline{Y}_{n_2}$ tehát, $t_{n_1+n_2-2}$ értéke $-\infty$-hez tart, ha n nő.
Így egy fix n-re akkor utasítjuk el H_0-t, ha $t_{n_1+n_2-2}$ a teszt statisztika értéke túl kicsi, azaz $< -\kappa'$.
Tehát H_0-t elfogadjuk $\Leftrightarrow -\kappa' < t_{n_1+n_2-2}$

Kétmintás próba egyoldali ellenhipotézissel

A **B példa** megoldása

B példa. Az alábbi két minta két különböző gyáregységben tapasztalt selejt-arányra vonatkozik (ezrelékben). Állítható-e 95%-os szignifikancia szinten, hogy az "A" gyáregység jobban dolgozott?

A	11,9	12,1	12,8	12,2	12,5	11,9	12,5	11,8	12,4	12,9
B	12,1	12,0	12,9	12,2	12,7	12,6	12,6	12,8	12,0	13,1

Az első minta: $X_1, X_2, \ldots, X_{10}$ az $\mathcal{N}(m_1, \sigma_1)$ eloszlásból.
A második minta: $Y_1, Y_2, \ldots, Y_{10}$ az $\mathcal{N}(m_2, \sigma_2)$ eloszlásból.
A hipotézisek:
$H_0 : m_1 \geq m_2$
$H_1 : m_1 < m_2$ (A-ban kevesebb a selejt)
Kétmintás t próba kell. Feltehető-e, hogy a szórások egyezőek? Ugyanis ez feltétele a t-próbának. Erre F-próbát végzünk, az jön ki, elfogadható, hogy a szórások megegyeznek.

$$t_{18} = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{(n_1-1)s_X^{*2} + (n_2-1)s_Y^{*2}}} \sqrt{\frac{n_1 n_2 (n_1+n_2-2)}{n_1+n_2}} = \frac{\overline{X}_{10} - \overline{Y}_{10}}{\sqrt{9(s_X^{*2} + s_Y^{*2})}} \sqrt{90} = -1,13$$

A kritikus érték 95%-os szignifikancia szinthez(= 0,05 elsőfajú hibavalószínűséghez) 18 szabadságfokhoz $\kappa' = 1,734$.

Egyoldali ellenhipotézis

A 2. példa megoldása, 97,5%, 99% szignifikancia szintek

PERCENTAGE POINTS OF THE T DISTRIBUTION

Tail Probabilities

One Tail		0.10	0.05	0.025	0.01	0.005	0.001	0.0005
D	1	3.078	6.314	12.71	31.82	63.66	318.3	637
E	2	1.886	2.920	4.303	6.965	9.925	22.330	31.6
G	3	1.638	2.353	3.182	4.541	5.841	10.210	12.92
R	4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
E	5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
E	6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
S	7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
	8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
O	9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
F	10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
	11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
F	12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
R	13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
E	14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
E	15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
D	16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
O	17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
M	18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
	19	1.328	1.729	2.093	2.539	2.861	3.579	3.883

Kétmintás próbák egyoldali ellenhipotézissel

A **B példa** megoldása

B példa. Az alábbi két minta két különböző gyáregységben tapasztalt selejt-arányra vonatkozik (ezrelékben). Állítható-e 95%-os szignifikancia szinten, hogy az “A” gyáregység jobban dolgozott?

A	11,9	12,1	12,8	12,2	12,5	11,9	12,5	11,8	12,4	12,9
B	12,1	12,0	12,9	12,2	12,7	12,6	12,6	12,8	12,0	13,1

Az első minta: $X_1, X_2, \ldots, X_{10}$ az $\mathcal{N}(m_1, \sigma_1)$ eloszlásból.
A második minta: $Y_1, Y_2, \ldots, Y_{10}$ az $\mathcal{N}(m_2, \sigma_2)$ eloszlásból.
A hipotézisek:
$H_0 : m_1 \geq m_2$
$H_1 : m_1 < m_2$ (A-ban kevesebb a selejt)
Kétmintás t próba kell. Feltehető-e, hogy a szórások egyezőek? Ugyanis ez feltétele a t-próbának. Erre F-próbát végzünk, az jön ki, elfogadható, hogy a szórások megegyeznek.

$$t_{18} = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{(n_1-1)s_X^{*2} + (n_2-1)s_Y^{*2}}} \sqrt{\frac{n_1 n_2 (n_1+n_2-2)}{n_1+n_2}} = \frac{\overline{X}_{10} - \overline{Y}_{10}}{\sqrt{9(s_X^{*2} + s_Y^{*2})}} \sqrt{90} = -1,13$$

A kritikus érték 95%-os szignifikancia szinthez(= 0,05 elsőfajú hibavalószínűséghez) 18 szabadságfokhoz $\kappa' = 1,734$. Mivel $-\kappa' = -1,734 < t_{18} = -1,13$, ezért H_0-t nem tudjuk elvetni.

Előadás vázlat

1. Intervallumbecslések
   - Általánosan
   - Példa
2. Paraméteres próbák
   - Kiindulási példa
   - Várható érték tesztelése, ha a szórás ismert
   - Várható érték tesztelése, ha a szórás ismeretlen
3. Hipotézisek fajtái
   - Egy- és kétoldali ellenhipotézis
   - Egymintás és kétmintás próbák
4. Nemparaméteres próbák
   - Khi négyzet eloszlás
   - Illeszkedés vizsgálat
   - Homogenitás vizsgálat
   - Függetlenség vizsgálat

Nemparaméteres próbák

3 fajta feladatot oldunk meg:

- Illeszkedés vizsgálat: egy $X_1, \ldots, X_n$ mintát veszünk egy eloszlásból, amely nem ismert számunkra. Igaz-e, hogy a minta egy általunk adott eloszlásból generálódott? (Mondjuk Exponenciális 2,3 paraméterrel.)
- Homogenitás vizsgálat: két mintánk van egy-egy eloszlásból: $X_1, \ldots, X_n$ és $Y_1, \ldots, Y_m$. Igaz-e, hogy a két eloszlás megegyezik?
- Függetlenség vizsgálat: adott egy kétdimenziós minta $(X_1, Y_1), \ldots, (X_n, Y_n)$ egy kétdimenziós véges értékű valószínűségi változóból. Igaz-e, hogy a marginálisok, X és Y függetlenek?

Khi négyzet eloszlás

Adott egy véges diszkrét eloszlás $p_1, p_2 \ldots, p_r$ az $\{1, 2, \ldots, r\}$ értékeken.
Adott egy $X_1, X_2, \ldots$ független minta ebből az eloszlásból.
Mindennek az alapja a következő:

Tétel

Legyen $N_i^{(n)} = \#\{j : 1 \leq j \leq n,\ X_j = i\} =$ "hányszor fordult elő az i érték az első n kísérletben". (Ugye $\sum_{i=1}^r N_i^{(n)} = n$.) Ekkor

$$\sum_{i=1}^{r} \frac{\left(N_i^{(n)} - np_i\right)^2}{np_i} \longrightarrow \chi^2_{r-1}$$

amint $n \to \infty$. χ^2_k k szabadságfokú "khi-négyzet" eloszlás.

χ^2_{r-1} eloszlása megegyezik $X_1^2 + X_2^2 + \cdots + X_{r-1}^2$ eloszlásával, ahol $X_1, X_2, \ldots, X_{r-1}$ független standard normális, $\mathcal{N}(0, 1)$, eloszlásúak.

Khi négyzet eloszlás, táblázat

df	0.1	0.05	0.025	0.01	0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718

Khi négyzet eloszlás

$N_i^{(n)} \sim Binom(n; p_i)$, így $\mathbf{E}N_i^{(n)} = np$, $\mathbf{Var}(N_i^{(n)}) = np_i(1-p_i)$.
$\left(N_1^{(n)}, N_2^{(n)}, \ldots, N_r^{(n)}\right) \sim Polinomialis(n; p_1, p_2 \ldots, p_r)$

A tétel bizonyítása $r = 2$-re:
Legyen $r = 2$, azaz az eloszlásunk $p_1 = p$ és $p_2 = 1-p$.
Ekkor $\frac{\left(N_1^{(n)}-np\right)^2}{np} + \frac{\left(N_2^{(n)}-n(1-p)\right)^2}{n(1-p)}$ összeget vizsgáljuk. Hozzunk közös nevezőre, és használjuk, hogy $N_2^{(n)} = n - N_1^{(n)}$:

$$\frac{(1-p)\left(N_1^{(n)}-np\right)^2+p\left(n-N_1^{(n)}-n(1-p)\right)^2}{np(1-p)} = \frac{(1-p)\left(N_1^{(n)}-np\right)^2+p\left(-N_1^{(n)}+np\right)^2}{np(1-p)}$$

$$= \frac{\left(N_1^{(n)}-np\right)^2}{np(1-p)} = \left(\frac{N_1^{(n)}-np}{\sqrt{np(1-p)}}\right)^2$$

A de Moivre-Laplace tétel szerint (= centrális-határeloszlás tétel spec. esete): $\frac{N_1^{(n)}-np}{\sqrt{np(1-p)}} \longrightarrow X, \quad n \to \infty,$
ahol $X \sim \mathcal{N}(0,1)$. Ezért a négyzet eloszlása $X^2 = \chi^2_{2-1}$.

Illeszkedésvizsgálat

Példa

$\varepsilon = 0,05$ elsőfajú hibavalószínűség mellet döntsünk arról a hipotézisről, hogy az alábbi megfigyelés-sorozat szabályos kockával dobva adódott.

értékek	1	2	3	4	5	6
gyakoriságok	24	21	19	12	13	11

H_0 : minden oldal valószínűsége $\frac{1}{6}$
H_1 : nem minden oldal valószínűsége $\frac{1}{6}$

Illeszkedésvizsgálat

Adott egy $r \geq 2$ pozitív egész. A háttérben adott egy $\underline{\pi} = (\pi_1, \pi_2, \dots, \pi_r)$ eloszlás, amit mi nem ismerünk. Nevezzük ezt igazi eloszlásnak. Ekkor veszünk egy független mintát ebből az eloszlásból: $X_1, X_2, \dots, X_n$.

Megadunk egy $\underline{p} = (p_1, p_2, \dots, p_r)$ eloszlást. Azt szeretnénk *tesztelni*, hogy az igazi eloszlás megegyezik-e az általunk adottal. Így a hipotézisek:

$H_0 : \underline{\pi} = \underline{p}$
$H_1 : \underline{\pi} \neq \underline{p}$

Tesztelés: a teszt statisztika $\chi^2_{r-1} := \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i}$. Miért? ↻

Ha H_0 teljesül, akkor $N_i \approx np_i$ minden i-re. Ezért a számlálók 0-hoz közeliek, így χ^2_{r-1} 0-hoz közeli.

Ha H_0 nem teljesül, akkor valamelyik i-re $N_i \approx n\pi_i \neq np_i$, így $\frac{(N_i - np_i)^2}{np_i}$ tart ∞-hez, tehát χ^2_{r-1} tart ∞-hez, ha $n \to \infty$

Tehát fix n-re: H_0-t elfogadjuk $\Leftrightarrow \chi^2_{r-1} < \kappa$.

Illeszkedésvizsgálat

Tehát fix n-re: H_0-t elfogadjuk $\Leftrightarrow \chi^2_{r-1} < \kappa$.
Pontosabban:
$1-\varepsilon$ szinten elfogadjuk H_0-t, ha κ-ra teljesül:
$\mathbf{P}\left(\chi^2_{r-1} < \kappa\right) = 1-\varepsilon$.
Ehhez szükség van $\chi^2_{r-1} = \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i}$ eloszlására, ha H_0 teljesül, azaz $X_1, \dots, X_n$ a $p_1, \dots, p_r$ eloszlásból vett független minta.
A fenti tétel szerint, ha H_0 teljesül, azaz $\underline{\pi} = \underline{p}$, azaz $X_1, \dots, X_n$ a $p_1, \dots, p_r$ eloszlásból vett független minta,
akkor $\sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i}$ *aszimptotikusan* χ^2_{r-1} eloszlású.
Így ahogy CHT-t alkalmazó közelítéseknél, fix n-re is
tekintsük a $\sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i}$ valváltozót χ^2_{r-1} eloszlásúnak.

Illeszkedés vizsgálat, a példa megoldása

$\varepsilon = 0,05$ elsőfajú hibavalószínűség mellet döntsünk arról a hipotézisről, hogy az alábbi megfigyelés-sorozat szabályos kockával dobva adódott.

értékek	1	2	3	4	5	6
gyakoriságok	24	21	19	12	13	11

H_0 : minden oldal valószínűsége $\frac{1}{6} \Leftrightarrow (\pi_1, \ldots, \pi_6) = (\frac{1}{6}, \ldots, \frac{1}{6})$
H_1 : H_0 nem teljesül $\Leftrightarrow (\pi_1, \ldots, \pi_6) \neq (\frac{1}{6}, \ldots, \frac{1}{6})$
A teszt statisztika értéke
$\chi_5^2 = \chi_{6-1}^2 = \sum_{i=1}^{6} \frac{(N_i - np_i)^2}{np_i} = \frac{44^2+26^2+14^2+28^2+22^2+34^2}{600} = 8,72.$
Az 5 szabadságfokú χ^2 eloszláshoz és $0,95$ szignifikancia szinthez tartozó kritikus érték $\kappa = 11,07$.

Khi négyzet eloszlás, táblázat

df	0.1	0.05	0.025	0.01	0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718

Illeszkedés vizsgálat, a példa megoldása

$\varepsilon = 0,05$ elsőfajú hibavalószínűség mellet döntsünk arról a hipotézisről, hogy az alábbi megfigyelés-sorozat szabályos kockával dobva adódott.

értékek	1	2	3	4	5	6
gyakoriságok	24	21	19	12	13	11

H_0 : minden oldal valószínűsége $\frac{1}{6} \Leftrightarrow (\pi_1, \ldots, \pi_6) = (\frac{1}{6}, \ldots, \frac{1}{6})$
H_1 : H_0 nem teljesül $\Leftrightarrow (\pi_1, \ldots, \pi_6) \neq (\frac{1}{6}, \ldots, \frac{1}{6})$
A teszt statisztika értéke
$\chi^2_5 = \chi^2_{6-1} = \sum_{i=1}^{6} \frac{(N_i - np_i)^2}{np_i} = \frac{44^2+26^2+14^2+28^2+22^2+34^2}{600} = 8,72.$
Az 5 szabadságfokú χ^2 eloszláshoz és $0,95$ szignifikancia szinthez tartozó kritikus érték $\kappa = 11,07$.
Mivel $\chi^2_5 < \kappa$ ezért nem tudjuk elutasítani H_0-t, így a minta tekinthető szabályos kockából származónak.

Illeszkedés vizsgálat, másik példa

Megoldás nélkül

Amikor az embereket megkérdezik, hogy mekkora a tömegük, gyakran mondanak a valóságosnál kisebb értékeket. Szeretnénk eldönteni az alábbi adathalmazról, hogy igazi mérésből származik, vagy az emberek megkérdezéséből nyerték. Azt a tényt fogjuk használni, hogy mérés esetén az utolsó számjegyek eloszlásának egyenletesnek kell lennie a $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ halmazon. Döntsünk $0,95$ szinten arról a hipotézisről, hogy mérésből származnak az adatok.

utosló számjegy	0	1	2	3	4	5	6	7	8	9
mérések száma	35	4	4	3	4	24	2	4	8	2

Homogenitás vizsgálat, példa

Két dobókockával dobva az alábbi gyakoriságokat figyeltük meg:

értékek	1	2	3	4	5	6
I. kocka	27	24	26	23	18	32
II. kocka	18	12	15	21	14	20

$\alpha = 0,05$ elsőfajú hibavalószínűség mellet döntsünk arról, hogy tekinthető-e a két eloszlás azonosnak.

H_0 : a két eloszlás azonos

H_1 : a két eloszlás különböző

Homogenitás vizsgálat

Adott két háttér eloszlás az $\{1, 2, \ldots, r\}$ halmazon:
$\underline{p} = (p_1, p_2, \ldots, p_r)$ és $\underline{q} = (q_1, q_2, \ldots, q_r)$. Ezek nem ismertek.

A $\underline{p}$ eloszlásból veszünk egy mintát: $X_1, X_2, \ldots, X_n$
A $\underline{q}$ eloszlásból veszünk egy mintát: $Y_1, Y_2, \ldots, Y_m$

$H_0 : \underline{q} = \underline{p}$
$H_1 : \underline{q} \neq \underline{p}$

Legyen $N_i = \#\{j : 1 \leq j \leq n,\ X_j = i\} = n$ kísérletből az i értékek száma az első mintában.
Legyen $M_i = \#\{j : 1 \leq j \leq n,\ Y_j = i\} = n$ kísérletből az i értékek száma az első mintában.

A teszt statisztika $T = nm \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i}$. ($H_0$ melletti eloszlása ismert.)

Ha H_0 teljesül, akkor minden j-re $\frac{N_j}{n} \approx \frac{np_j}{n} = p_j$ és $\frac{M_j}{m} \approx \frac{mp_j}{m} = p_j$, tehát a számlálók 0-hoz közeliek, így az egész szumma 0-hoz közeli.
Ha H_1 teljesül, akkor valamely j-re az egyik tört ∞-hez tart.

Így fix n-re akkor fogadjuk el H_1-t, ha T elég nagy, $\kappa < T$.

Homogenitás vizsgálat

Állítás

Ha $X_1, X_2, \dots, X_n$ és $Y_1, Y_2, \dots, Y_m$ két független minta ugyanabból az eloszlásból az $\{1, 2, \dots, r\}$ értékeken, akkor

$$nm \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i} \longrightarrow \chi^2_{r-1}, \text{ ha } n \to \infty.$$

Fix n esetén úgy vesszük, hogy $nm \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i}$ eloszlása $\approx \chi^2_{r-1}$.

Tehát a döntés:

H_0-t $1 - \varepsilon$ szignifikancia szinten elfogadjuk $\Leftrightarrow$

$nm \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i} < \kappa$, ahol $\mathbf{P}(\chi^2_{r-1} < \kappa) = 1 - \varepsilon$.

Homogenitás vizsgálat, példa megoldása

Két dobókockával dobva az alábbi gyakoriságokat figyeltük meg:

értékek	1	2	3	4	5	6
I. kocka	27	24	26	23	18	32
II. kocka	18	12	15	21	14	20

$\alpha = 0,05$ elsőfajú hibavalószínűség mellet döntsünk arról, hogy tekinthető-e a két eloszlás azonosnak.

H_0 : a két eloszlás azonos

H_1 : a két eloszlás különböző

$\chi_5^2 = nm \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i} = 2,2$

0,95 szignifikancia szinthez és 5 szabadsági fokhoz tartozó kritikus érték $\kappa = 11,07$

Khi négyzet eloszlás, táblázat

df	0.1	0.05	0.025	0.01	0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718

Homogenitás vizsgálat, példa megoldása

Két dobókockával dobva az alábbi gyakoriságokat figyeltük meg:

értékek	1	2	3	4	5	6
I. kocka	27	24	26	23	18	32
II. kocka	18	12	15	21	14	20

$\alpha = 0,05$ elsőfajú hibavalószínűség mellet döntsünk arról, hogy tekinthető-e a két eloszlás azonosnak.

H_0 : a két eloszlás azonos

H_1 : a két eloszlás különböző

$\chi_5^2 = nm \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i} = 2,2$

0,95 szignifikancia szinthez és 5 szabadsági fokhoz tartozó kritikus érték $\kappa = 11,07$

Mivel $\chi_5^2 < \kappa$, H_0-t nem tudjuk elvetni, tehát elfogadjuk, hogy a két dobókocka azonosnak tekinthető.

Függetlenség vizsgálat, Példa

Az alábbi három táblázat három – a TV-ben különböző intenzitással reklámozott – fogkrém fogyasztására vonatkozó adatokat tartalmaz a TV-nézés idejének függvényében:

Fogkrém fajtája→	A	B	C
TV nézés hetente↓	1 óra reklám	5 perc reklám	0 perc reklám
$<$ 5 óra	80	60	60
5–15 óra	70	70	60
$>$ 15 óra	90	65	45

Van-e összefüggés a kedvelt fogkrém márkája és a TV nézés időtartama között?
A cellákban az eladott mennyiség áll az adott kategóriából.

H_0 : a reklám időtartama és a TV nézés függetlenek

H_1 : van közöttük összefüggés

Függetlenség vizsgálat

Adott egy kétdimenziós véges értékű valváltozó, (X, Y).
X értékei $\{1, 2, \ldots, r\}$, Y értékei $\{1, 2, \ldots, s\}$.
Tesztelni szeretnénk, hogy X független-e Y-tól. Ehhez veszünk egy n elemű mintát: $(X_1, Y_1), (X_2, Y_2) \ldots, (X_n, Y_n)$.

	1	2	$\cdots$	j	$\cdots$	s	sor$\sum$
1	N_{11}	N_{12}		N_{1j}		N_{1s}	$N_{1\bullet}$
2	N_{21}	N_{22}		N_{2j}		N_{2s}	$N_{2\bullet}$
$\vdots$							
i	N_{i1}	N_{i2}		N_{ij}		N_{is}	$N_{i\bullet}$
$\vdots$							
r	N_{r1}	N_{r2}		N_{rj}		N_{rs}	$N_{r\bullet}$
oszlop$\sum$	$N_{\bullet 1}$	$N_{\bullet 2}$		$N_{\bullet j}$		$N_{\bullet s}$	

$\sum_{i=1}^{r} \sum_{j=1}^{s} N_{ij} = n$.

A teszt statisztika: $\chi^2_{(r-1)(s-1)} = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(\frac{N_{ij}}{n} - \frac{N_{i\bullet}}{n} \frac{N_{\bullet j}}{n} \right)^2}{\frac{N_{i\bullet}}{n} \frac{N_{\bullet j}}{n}}$

Függetlenség vizsgálat

A teszt statisztika: $\chi^2_{(r-1)(s-1)} = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(\frac{N_{ij}}{n} - \frac{N_{i\bullet}}{n}\frac{N_{\bullet j}}{n}\right)^2}{\frac{N_{i\bullet}}{n}\frac{N_{\bullet j}}{n}}$

Ha H_0 teljesül, azaz X és Y függetlenek, akkor $\mathbf{P}(X = i, Y = j) = \mathbf{P}(X = i)\mathbf{P}(Y = j)$. Továbbá tudjuk, hogy ekkor $\frac{N_{ij}}{n} \to \mathbf{P}(X = i, Y = j)$, $\frac{N_{i\bullet}}{n} \to \mathbf{P}(X = i)$, és $\frac{N_{\bullet j}}{n} \to \mathbf{P}(Y = j)$. Tehát a szummában felsorolt törtek 0-hoz tartanak. Így H_0-t elfogadjuk, ha $\chi^2_{(r-1)(s-1)}$ 0-hoz közeli.

Ha H_1 teljesül, akkor valamilyen (i, j) párra $\frac{N_{ij}}{n} - \frac{N_{i\bullet}}{n}\frac{N_{\bullet j}}{n} \to c \neq 0$, így n-es szorzó miatt $\chi^2_{(r-1)(s-1)} \to \infty$, ha $n \to \infty$.

Fix n-re, H_0-t elfogadjuk, ha $\chi^2_{(r-1)(s-1)}$ elég kicsi, azaz ha $\chi^2_{(r-1)(s-1)} < \kappa$, valamilyen kappára.

Függetlenség vizsgálat

$$\chi^2_{(r-1)(s-1)} = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(\frac{N_{ij}}{n}-\frac{N_{i\bullet}}{n}\frac{N_{\bullet j}}{n}\right)^2}{\frac{N_{i\bullet}}{n}\frac{N_{\bullet j}}{n}},$$ ha n tart végtelenhez,

akkor χ^2 eloszlású $(r-1)(s-1)$ szabadságfokkal.
Azaz ha H_0 teljesül, a határeloszlás ismert. (aszimptotikusan...)

Döntés:
H_0-t elfogadjuk $1-\varepsilon$ szinten $\Leftrightarrow \chi^2_{(r-1)(s-1)} < \kappa$, ahol κ az $(r-1)(s-1)$ szabadságfokú χ^2 eloszlás ε valószínűséghez tartozó farok eloszlása, azaz $\mathbf{P}(\chi^2_{(r-1)(s-1)} > \kappa) = \varepsilon$
(ez ugye azt jelenti, hogy $\mathbf{P}(\chi^2_{(r-1)(s-1)} < \kappa) = 1-\varepsilon$)

Függetlenség vizsgálat, a példa megoldása

Az alábbi három táblázat három – a TV-ben különböző intenzitással reklámozott – fogkrém fogyasztására vonatkozó adatokat tartalmaz a TV-nézés idejének függvényében:

Fogkrém fajtája$\rightarrow$	A	B	C
TV nézés hetente$\downarrow$	1 óra reklám	5 perc reklám	0 perc reklám
$<$ 5 óra	80	60	60
5–15 óra	70	70	60
$>$ 15 óra	90	65	45

Van-e összefüggés a kedvelt fogkrém márkája és a TV nézés időtartama között? Döntsünk 0,95%-os szinten.

$r = s = 3$

$$\chi_4^2 = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(\frac{N_{ij}}{n} - \frac{N_{i\bullet}}{n}\frac{N_{\bullet j}}{n}\right)^2}{\frac{N_{i\bullet}}{n}\frac{N_{\bullet j}}{n}} = 6$$

A 4 szabadságfokú 0,05 farokvalószínűséghez tartozó érték a χ^2 eloszlásban $\kappa = 9.488$.

Khi négyzet eloszlás, táblázat

df	0.1	0.05	0.025	0.01	0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	$\boxed{9.488}$	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718

Függetlenség vizsgálat, a példa megoldása

Az alábbi három táblázat három – a TV-ben különböző intenzitással reklámozott – fogkrém fogyasztására vonatkozó adatokat tartalmaz a TV-nézés idejének függvényében:

Fogkrém fajtája$\rightarrow$	A	B	C
TV nézés hetente$\downarrow$	1 óra reklám	5 perc reklám	0 perc reklám
$<$ 5 óra	80	60	60
5–15 óra	70	70	60
$>$ 15 óra	90	65	45

Van-e összefüggés a kedvelt fogkrém márkája és a TV nézés időtartama között? Döntsünk 0,95%-os szinten.

$r = s = 3$

$$\chi_4^2 = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(\frac{N_{ij}}{n} - \frac{N_{i\bullet}}{n}\frac{N_{\bullet j}}{n}\right)^2}{\frac{N_{i\bullet}}{n}\frac{N_{\bullet j}}{n}} = 6$$

A 4 szabadságfokú 0,05 farokvalószínűséghez tartozó érték a χ^2 eloszlásban $\kappa = 9.488$.

Mivel $\chi_4^2 < \kappa$, elfogadjuk H_0-t.