Statisztika **1. zárthelyi dolgozat** 2009. március 18.

1. Ismeretlen m várható értékű, 1 szórású normális eloszlásból a következő hatelemű mintát kaptuk:

$$2,48 \quad 3,3 \quad 1,83 \quad 0,1 \quad 1,32 \quad 2,97$$

   a) Számítsuk ki a mintaközepet és a tapasztalati szórásnégyzetet! (8 pont)
   b) Tegyük fel, hogy X normális eloszlású valószínűségi változó, melynek várható értéke az *a)* feladatban meghatározott mintaközép, szórásnégyzete pedig 1. Határozzuk meg annak valószínűségét, hogy $X < 3$! (6 pont)

2. Ismeretlen $b > 0$ paraméterű Poisson-eloszlásból származó független $X_1, X_2, \ldots, X_n$ minta esetén jelölje $\overline{X}$ a mintaközepet, s_n^{*2} pedig a korrigált tapasztalati szórásnégyzetet. Milyen $0 < r < 1$ számokra igaz, hogy

$$T(X_1, X_2, \ldots, X_n) = r \cdot \overline{X} + (1-r) \cdot s_n^{*2}$$

   torzítatlan becslése a b paraméternek? (12 pont)

3. Legyen $X_1, X_2, \ldots, X_n$ független minta a következő diszkrét eloszlásból:

$$P(X_i = 1) = b,\ P(X_i = 2) = \frac{1}{2} - 3b,\ P(X_i = 4) = \frac{1}{2} + 2b,\ i = 1, 2, \ldots, n,$$

   ahol $0 \le b \le \frac{1}{6}$ az ismeretlen paraméter. Adjunk becslést a b paraméterre momentum-módszerrel! (12 pont)

4. $a > 0$ esetén legyen az $f_a(x)$ sűrűségfüggvény a következő:

$$f_a(x) = \begin{cases} a^2 \cdot x \cdot e^{-ax} & \text{ha } x > 0, \\ 0 & \text{különben.} \end{cases}$$

   Legyen $X_1, X_2, \ldots, X_n$ független minta az $f_a(x)$ sűrűségfüggvényhez tartozó eloszlásból, ahol $a > 0$ ismeretlen paraméter. Határozzuk meg az a paraméter maximum likelihood becslését! (12 pont)

Beadható feladat: Legyen $X_1, X_2, \ldots$ független minta az $\left[\frac{a}{2}, \frac{3a}{2}\right]$ intervallumon egyenletes eloszlásból, ahol $a > 0$ ismeretlen paraméter. Igaz-e, hogy a mintaközép konzisztens becslése az a paraméternek?

A megoldásokat indokolni kell, a teljes pontszámhoz jó végeredmény és helyes indoklás szükséges. Összesen 50 pontot lehet elérni, az egyes feladatokért kapható pontszámok a feladat szövege mellett szerepelnek. A várható ponthatárok: 40, 61, 72, 83.

Az elégséges határa 20 pont, aki ezt nem éri el vagy a dolgozatot nem írja meg, szorgalmi időszak utolsó hetében vagy a vizsgaidőszak első hetében pótolhatja a dolgozatot.

Az eredmények az ETR infosheet rovatában lesznek elérhetők, a megoldásokat pedig a http://www.cs.elte.hu/~agnes/gyak címen lehet majd megtalálni.

Mintaközép/tapasztalati közép: $\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$
Tapasztalati szórásnégyzet:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

Tapasztalati szórás:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2}$$

Korrigált tapasztalati szórásnégyzet:

$$s_n^{*2} = \frac{n}{n-1}\left[\frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2\right]$$

Szórási együttható:

$$c = \frac{s}{\overline{X}}$$

Tapasztalati eloszlásfüggvény:

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i < t) = \frac{1}{n}\left|\{i : X_i < t\}\right|$$

Ha $X_1, X_2, \ldots, X_n$ független minta valamely véges szórású eloszlásból, akkor: $E\left(\overline{X}\right) = EX_1$, $D^2\left(\overline{X}\right) = \frac{1}{n}D^2(X_1)$, és $E\left(s_n^{*2}\right) = D^2(X_1)$.

$a > 0$ paraméterű *exponenciális* eloszlás eloszlásfüggvénye: $F(t) = 1 - e^{-at}$, ha $t \geq 0$, különben 0; sűrűségfüggvénye: $f(t) = ae^{-at}$, ha $t \geq 0$, különben 0;
várható értéke: $\frac{1}{a}$; szórásnégyzete: $\frac{1}{a^2}$.

$b > 0$ paraméterű *Poisson-eloszlás*: $p_k = \frac{b^k}{k!}e^{-b}$; várható értéke: b; szórásnégyzete: b.

$[a, b]$ intervallumon *egyenletes* eloszlás eloszlásfüggvénye:
$F(t) = \frac{t-a}{b-a}$, ha $a \leq t \leq b$, 0, ha $t \leq a$, 1 különben;
sűrűségfüggvénye: $f(t) = \frac{1}{b-a}$, ha $a \leq t \leq b$, 0 különben; várható értéke: $\frac{a+b}{2}$; szórásnégyzete: $\frac{(b-a)^2}{12}$.

Binomiális eloszlás $n \geq 1$ és $0 < p < 1$ paraméterekkel: $p_k = \binom{n}{k}p^k(1-p)^{n-k}$;
várható értéke: $n \cdot p$; szórásnégyzete: $n \cdot p \cdot (1-p)$.

$N(m, \sigma^2)$ *normális* eloszlás sűrűségfüggvénye: $f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$;
várható értéke: m, szórásnégyzete: σ^2. Eloszlásfüggvény táblázata a lap másik oldalán.

$0 < p < 1$ paraméterű *geometriai* eloszlás: $p_k = (1-p)^{k-1} \cdot p$;
várható értéke: $\frac{1}{p}$, szórásnégyzete: $\frac{1-p}{p^2}$.

Megoldások

1. *a)* Legyen a minta $X_1, X_2, \ldots, X_6$, ekkor a mintaközép:

$$\overline{X} = \frac{1}{6}\left(X_1 + X_2 + \ldots + X_6\right) = \frac{1}{6}\left(2,48 + 3,3 + 1,83 + 0,1 + 1,32 + 2,97\right) = 2.$$

A tapasztalati szórásnégyzet:

$$s^2 = \frac{1}{6}\sum_{i=1}^{6} X_i^2 - \overline{X}^2 = \frac{1}{6}\left(2,48^2 + 3,3^2 + 1,83^2 + 0,1^2 + 1,32^2 + 2,97^2\right) - 2^2 \approx 1,1604.$$

b) Az előző szerint tehát $X \sim N\left(2,1\right)$, azaz X normális eloszlású 2 várható értékkel és 1 szórással. Ekkor az $X - 2$ valószínűségi változó normális eloszlású 0 várható értékkel és 1 szórással, azaz eloszlása standard normális, így

$$P\left(X < 3\right) = P\left(X - 2 < 1\right) = \Phi\left(1\right) \approx 0,8413$$

a standard normális eloszlásfüggvény táblázata alapján.

2. Gyakorlaton szerepelt, hogy tetszőleges véges szórású eloszlásból származó $X_1, \ldots, X_n$ független minta esetén a mintaközép várható értéke $E\left(X_1\right)$, a korrigált tapasztalati szórásnégyzet várható értéke pedig $D^2\left(X_1\right)$. A Poisson-eloszlás tetszőleges $b > 0$ paraméter mellett véges szórású, így alkalmazhatjuk ezeket az összefüggéseket. Felhasználva még a várható érték linearitását, azt kapjuk, hogy minden $b > 0$-ra

$$E_b\left(r \cdot \overline{X} + (1-r) \cdot s_n^{*2}\right) = r \cdot E_b\left(\overline{X}\right) + (1-r)\, E_b\left(s_n^{*2}\right) = r \cdot E_b\left(X_1\right) + (1-r)\, D_b^2\left(X_1\right).$$

b paraméterű Poisson-eloszlás várható értéke és szórásnégyzete is b, tehát a b paraméter mellett X_1 várható értéke és szórásnégyzete is b. Azaz

$$E_b\left(T\left(X_1, \ldots, X_n\right)\right) = r \cdot E_b\left(X_1\right) + (1-r)\, D_b^2\left(X_1\right) = r \cdot b + (1-r) \cdot b = b.$$

Tehát minden $0 < r < 1$ számra minden $b > 0$ paraméterre $E_b\left(r \cdot \overline{X} + (1-r) \cdot s_n^{*2}\right) = b$, azaz minden $0 < r < 1$ számra $r \cdot \overline{X} + (1-r) \cdot s_n^{*2}$ torzítatlan becslése a b paraméternek.

3. Egyetlen paramétert kell becsülnünk, így használjuk az első momentumot, a várható értéket. A várható érték definíciója szerint tetszőleges $i = 1, 2, \ldots, n$-re

$$\begin{aligned} E\left(X_i\right) &= 1 \cdot P\left(X_i = 1\right) + 2 \cdot P\left(X_i = 2\right) + 4 \cdot P\left(X_i = 4\right) = \\ &= b + 2 \cdot \left(\frac{1}{2} - 3b\right) + 4 \cdot \left(\frac{1}{2} + 2b\right) = b + 1 - 6b + 2 + 8b = 3 + 3b. \end{aligned}$$

A tapasztalati eloszlás várható értéke a mintaközép. Tehát a b paraméter momentum-módszerrel kapott becslése az a b szám lesz, melyre

$$\overline{X} = E\left(X_1\right) = 3 + 3b.$$

Átrendezve adódik a b becslése:

$$\hat{b} = \frac{\overline{X} - 3}{3}.$$

4. Mivel adott a sűrűségfüggvény, folytonos eloszlások családjáról van szó, így a likelihood-függvényt úgy kapjuk, hogy az egyes mintaelemeket behelyettesítjük az adott paraméterhez tartozó sűrűségfüggvénybe, és ezeket összeszorozzuk. A paraméter minden értékére a negatív számokon a sűrűségfüggvény nulla, így 1 valószínűséggel minden mintaelem pozitív, vagyis:

$$f\left(a\right) = \prod_{i=1}^{n} f_a\left(X_i\right) = \prod_{i=1}^{n}\left(a^2 \cdot X_i \cdot e^{-aX_i}\right).$$

A maximum likelihood becslés az az $a > 0$ szám lesz, melyre $f(a)$ maximális. $f(a)$ minden a-ra pozitív, és a logaritmusfüggvény monoton növő, így elég a log-likelihood függvényt maximalizálni. A log-likelihood függvény a likelihood függvény logaritmusa, szorzat logaritmusa pedig a logaritmusok összege:

$$\begin{aligned} l(a) &= \sum_{i=1}^{n} \log\left(a^2 \cdot X_i \cdot e^{-aX_i}\right) = \\ &= \sum_{i=1}^{n} \left(\log a^2 + \log X_i - aX_i\right) = \\ &= 2n \log a + \sum_{i=1}^{n} \log X_i - a \sum_{i=1}^{n} X_i. \end{aligned}$$

Ez a-nak folytonosan differenciálható függvénye. Ha $a \to 0+$, akkor az első tag $-\infty$-hez tart, a második konstans, a harmadik nullához tart, azaz ilyenkor $l(a) \to -\infty$. Ha $a \to \infty$, akkor az első tag végtelenhez tart logaritmikusan, a második tag konstans, a harmadik tag $-\infty$-hez tart lineárisan, így ilyenkor $l(a) \to -\infty$. Mindezekből következik, hogy $l(a)$-nak van maximuma, és ott a deriváltja nulla. Tekintsük tehát a log-likelihood egyenletet: $l'(a) = 0$, azaz

$$\frac{2n}{a} - \sum_{i=1}^{n} X_i = 0.$$

Ennek egyetlen megoldása van, amit átrendezéssel kaphatunk meg: $a = \frac{2n}{\sum_{i=1}^{n} X_i} = \frac{2}{\overline{X}}$. Tehát ez az a maximalizálja a log-likelihood és likelihood függvényeket, azaz a maximum likelihood becslése:

$$\hat{a} = \frac{2n}{\sum_{i=1}^{n} X_i} = \frac{2}{\overline{X}}.$$

1. A strandon figyeljük, hogy ki mennyi időt tölt a tóban fürdéssel, és feltételezzük, hogy az egyes emberek vízben töltött ideje egymástól független, eloszlásuk pedig a következő sűrűségfüggvénnyel adható meg:

$$f(x) = \frac{1}{2} \cdot a^3 \cdot x^2 \cdot e^{-ax}, \text{ ha } x > 0, \text{ és 0 különben,}$$

ahol $a > 0$ ismeretlen paraméter. Számítsuk ki n ember vízben töltött idejének megfigyeléséből adódó Fisher-információt. Ehhez felhasználhatjuk, hogy a fent megadott eloszlás várható értéke $\frac{3}{a}$, szórásnégyzete $\frac{3}{a^2}$. (9 pont)

2. Az Esterházy-kastélyban pihenő vendégek szokásait szeretnénk feltérképezni. Feltételezésünk, hogy a vendégek egymástól függetlenül $p = 1/4$ valószínűséggel töltik alvással a délutánt. Megkérdezünk 10 véletlenszerűen kiválasztott vendéget, és ha közülük legalább 9-en szoktak délután aludni, akkor elvetjük a nullhipotézisünket, különben elfogadjuk. Számítsuk ki az elsőfajú hiba valószínűségét! (9 pont)

3. A közeli település híres pincéinek egyikében szürkebarát és olaszrizling borokat vizsgáltak. Mindkét fajtából 7-7 pohárban megmérték a cukortartalmat, g/l-ben kifejezve a következő értékek adódtak:

szürkebarát	13,1	12,5	13,0	14,0	13,5	12,6	13,4
olaszrizling	13,6	12,7	12,3	11,9	13,4	12,8	12,0

Legyen nullhipotézisünk az, hogy a kétféle bor átlagos cukortartalma megegyezik, ellenhipotézisünk az, hogy a szürkebarátés nagyobb. Ezt a feladatot vizsgálva $\alpha = 0,03$ terjedelem mellett döntsünk arról a feltételezésről, hogy az olaszrizling nem édesebb a szürkebarátnál! (9 pont)

4. A parton a nádas átlagos magasságát vizsgálták. 5 szálat véletlenszerűen kiválasztottak, ezek vízszint feletti magassága méterben:

1,43	1,55	1,58	1,64	1,44

$\alpha = 0,05$ terjedelem mellett elfogadható-e az a feltételezés, hogy az átlagos magasság legfeljebb másfél méter? (9 pont)

5. A XIII. században alapított vár köveit vizsgálták színük és állapotuk szerint. Világos kőből 105-t, sötétből 91-t találtak. A világosak közül 61, a sötétek közül 48 volt ép, a többi sérült.
a) $\alpha = 0,01$ terjedelem mellett elfogadható-e, hogy a szín és az állapot egymástól független?
b) $\alpha = 0,01$ terjedelem mellett elfogadható-e a következő feltételezés: a szín és az állapot egymástól független, minden kő $1/2 - 1/2$ valószínűséggel sötét vagy világos, és szintén $1/2 - 1/2$ valószínűséggel ép vagy sérült? (14 pont)

A megoldásokat indokolni kell, a teljes pontszámhoz jó végeredmény és helyes indoklás szükséges. Összesen 50 pontot lehet elérni, az egyes feladatokért kapható pontszámok a feladat szövege mellett szerepelnek. A várható ponthatárok: 40, 61, 72, 83.

Az elégséges határa 20 pont, aki ezt nem éri el vagy a dolgozatot nem írja meg, szorgalmi időszak utolsó hetében vagy a vizsgaidőszak első hetében pótolhatja a dolgozatot, ennek időpontját később egyeztetjük, erről a kurzusfórumban lehet elolvasni.

Az eredmények és a megajánlott jegyek az ETR infosheet rovatában lesznek elérhetők, a megoldásokat pedig a http://www.cs.elte.hu/~agnes/gyak címen lehet majd megtalálni.

1. A megfigyeléseket jelölje $X_1, \ldots, X_n$, ezek tehát független, azonos eloszlású valószínűségi változók a megadott sűrűségfüggvénnyel. Folytonos eloszlásokról van szó, hiszen létezik sűrűségfüggvény, így n elemű mintából számolva a likelihood-függvény:

$$f(a) = \prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} \left[\frac{1}{2} \cdot a^3 \cdot X_i^2 \cdot e^{-aX_i}\right] = \left(\frac{1}{2}\right)^n \cdot a^{3n} \cdot \left(\prod_{i=1}^{n} X_i^2\right) \cdot e^{-\sum_{i=1}^{n} aX_i}.$$

Ennek logaritmusa a log-likelihood függvény:

$$l(a) = n \log\left(\frac{1}{2}\right) + 3n \log a + \log\left(\prod_{i=1}^{n} X_i^2\right) - a \sum_{i=1}^{n} X_i.$$

Az a paraméter szerint deriválva:

$$\frac{d}{da} l(a) = \frac{3n}{a} - \sum_{i=1}^{n} X_i.$$

A Fisher-információt megkaphatjuk úgy, hogy ennek a mennyiségnek kiszámítjuk a szórásnégyzetét az a paraméterhez tartozó sűrűségfüggvénnyel kapható mintából. Az első tag konstans, véletlentől nem függ, így levonva a szórásnégyzet nem változik. -1-gyel szorozva sem változik a szórásnégyzet. Tehát független valószínűségi változók összegének szórásnégyzete jelenik meg, ez a szórásnégyzetek összege, most $D^2(X_1)$ n-szerese, hiszen azonos eloszlású valószínűségi változókról van szó. A feladatban szerepel, hogy az a-hoz tartozó eloszlás szórásnégyzete $3/a^2$, így ez lesz X_1 szórásnégyzete is. Összefoglalva:

$$I(a) = D_a^2\left(\frac{d}{da} l(a)\right) = D_a^2\left(\frac{3n}{a} - \sum_{i=1}^{n} X_i\right) = D_a^2\left(\sum_{i=1}^{n} X_i\right) =$$
$$= \sum_{i=1}^{n} D_a^2(X_i) = nD_a^2(X_1) = n\frac{3}{a^2}.$$

2. Nullhipotézisünk, hogy minden vendég egymástól függetlenül p valószínűséggel alszik délután. Az elsőfajú hiba a nullhipotézis melletti valószínűsége annak, hogy hibás döntést hozunk, azaz elvetjük a nullhipotézist. A nullhipotézist akkor vetjük el, ha a tíz megkérdezett ember közül leglább kilencen alszanak délután. Az elsőfajú hiba valószínűsége tehát annak valószínűsége, hogy tíz ember közül legalább kilencen alszanak délután, feltételezve, hogy mindenki egymástól függetlenül p valószínűséggel alszik délután. A binomiális eloszlás megfelelő tagjaiból adódik ennek értéke:

$$p^{10} + 10 \cdot p^9 \cdot (1-p).$$

Az első tag ugyanis annak valószínűsége, hogy mind a tízen alszanak, a második annak, hogy pontosan kilencen alszanak, tehát az összeg adja az elsőfajú hiba valószínűségét. $p = 1/3$-ra ez körülbelül $3,56 \cdot 10^{-4}$, $p = 1/4$-re körülbelül $2,96 \cdot 10^{-5}$.

3. Feltételezzük, hogy az egyes borok cukortartalma normális eloszlású, és a szórások megegyeznek. A szürkebarát cukortartalmának eloszlása $N(m_1, \sigma)$, az olaszrizlingének pedig $N(m_2, \sigma)$. Ezekkel a jelölésekkel:

$$H_0 : m_1 = m_2$$

$$H_1 : m_1 > m_2$$

Tehát a hipotézisvizsgálati feladat normális eloszlások várható értékének összehasonlításáról szól, a szórások nem ismertek, de egyenlőnek feltételezhetők, továbbá feltételezhetjük, hogy két független mintával van dolgunk. Ezért kétmintás t-próbát végzünk. Kiszámítjuk a próbastatisztikát:

$$t = \sqrt{\frac{mn\left(m+n-2\right)}{m+n}} \frac{\overline{X} - \overline{Y}}{\sqrt{\left(n-1\right) s_n^{*2}\left(X\right) + \left(m-1\right) s_m^{*2}\left(Y\right)}}.$$

Ennek értéke a különböző adatsorokra különböző. A próbastatisztika eloszlása H_0 mellett $n+m-2$ szabadságfokú t-eloszlás, ahol n és m az egyes minták elemszáma, azaz $n = m = 7$, a szabadságfok pedig 12. Tehát a kritikus érték 12 szabadságfokú t-eloszláshoz és $\alpha = 0,05$ terjedelemhez tartozó egyoldali kritikus érték:

$$c_{krit} = 1,782.$$

Tehát a próbánk a következő:
Ha $t \leq 1,782$, akkor elfogadjuk H_0-t, és ezt úgy értelmezzük, hogy vagy az olaszrizling legalább olyan édes, mint a szürkebarát, vagy nincs elég adatunk ahhoz, hogy ezt megcáfoljuk.
Ha $t > 1,782$, akkor elvetjük H_0-t, és ezt úgy értelmezzük, hogy statisztikai bizonyítékot nyertünk arra, hogy a szürkebarát édesebb az olaszrizlingnél.

4. Feltételezzük, hogy a nádszálak magassága $N\left(m, \sigma\right)$ eloszlású. A következő feladatot vizsgáljuk:

$$H_0 : m \leq 1,5$$

$$H_1 : m > 1,5$$

Normális eloszlás várható értékről szól a hipotézisvizsgálati feladat, és a szórás nem ismert, ezért egymintás t-próbát végzünk. Kiszámítjuk a próbastatisztikát:

$$t = \sqrt{n} \frac{\overline{X} - m_0}{s_n^*}.$$

H_0 mellett a próbastatisztika $n-1$ szabadságfokú t-eloszlás, ahol n a minta elemszáma, azaz 5. Egyoldali ellenhipotéziünk van, ezért a kritikus érték a 4 szabadságfokú t-eloszláshoz és $\alpha = 0,05$ terjedelmhez tartozó egyoldali kritikus érték:

$$c_{krit} = 2,132.$$

A próbánk a következő:
Ha $t \leq 2,132$, akkor elfogadjuk H_0-t, azaz elfogadható az a feltételezés, hogy a nádas átlagos magassága legfeljebb másfél méter.
Ha $t > 2,132$, akkor elvetjük H_0-t, és statisztikai bizonyítékunk van arra, hogy a nádas átlagos magassága több másfél méternél, a feltételezés nem fogadható el.

5. *a)* Az adatokat az alábbi táblázatban foglalhatjuk össze:

	ép	sérült	összesen
világos	61	44	105
sötét	48	43	91
összesen	109	87	196

Függetlenségvizsgálatot végzünk, a nullhipotézis az, hogy a két szempont szerinti osztályozás egymástól független, ellenhipotézis, hogy nem független. Végezhetünk χ^2-próbát osztályok összevonása nélkül, mert minden osztályba esik legalább 5 megfigyelés. Mindkét szempont szerint két osztály van, $r = s = 2$, és így a próbastatisztika az órán használt jelölésekkel:

$$n\frac{(\nu_{11}\nu_{22} - \nu_{12}\nu_{21})^2}{\nu_{1.}\nu_{2.}\nu_{.1}\nu_{.2}} = 196 \cdot \frac{(61 \cdot 43 - 44 \cdot 48)^2}{105 \cdot 91 \cdot 109 \cdot 87} \approx 0,565.$$

A kritikus érték az 1 szabadságfokú χ^2-próbához tartozó kritikus érték, $\alpha = 0,1$ terjedelem mellett 2,71, $\alpha = 0,05$ terjedelem mellett 3,84, végül $\alpha = 0,01$ terjedelem mellett 6,63. Mindhárom esetben $\chi^2 < c_{krit}$, ezért elfogadjuk a nullhipotézist, azaz elfogadható az a feltételezés, hogy a szín és állapot szerinti osztályozás egymástól független, az adatok ezt nem cáfolják.

b) A köveket négy osztályba sorolták:

	ép és világos	ép és sötét	sérült és világos	sérült és sötét
darab	61	48	44	43

A következő feltételezést vizsgáljuk: a szín és az állapot egymástól független, minden kő $1/2 - 1/2$ valószínűséggel sötét vagy világos, és szintén $1/2 - 1/2$ valószínűséggel ép vagy sérült. Ezek együttesen azt jelentik, hogy minden kő a négy osztály mindegyikébe $1/4 - 1/4$ valószínűséggel esik, tehát illeszkedésvizsgálatot végzünk, a nullhipotézis, hogy minden osztályba $1/4$ valószínűséggel esnek a kövek, az ellenhipotézis, hogy az eloszlás ettől különböző.

Most sem kell osztályokat összevonni. A próbastatisztika:

$$\sum_{i=1}^{r}\frac{(\nu_i - np_i)^2}{np_i} = \frac{(61-49)^2}{49} + \frac{(48-49)^2}{49} + \frac{(44-49)^2}{49} + \frac{(43-49)^2}{49} \approx 1,051.$$

r az osztályok száma, azaz 4, így az $r - 1 = 3$ szabadságfokú χ^2-eloszlás kritikus értékét kell megkeresnünk. $\alpha = 0,1$ terjedelem mellett 6,25, $\alpha = 0,05$ terjedelem mellett 7,81, végül $\alpha = 0,01$ terjedelem mellett 11,3. Mindhárom esetben $\chi^2 < c_{krit}$, ezért elfogadjuk a nullhipotézist, azaz elfogadható a fenti feltételezés, az adatok ezt nem cáfolják.

Mintaközép/tapasztalati közép: $\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$
Tapasztalati szórásnégyzet:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

Tapasztalati szórás:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2}$$

Korrigált tapasztalati szórásnégyzet:

$$s_n^{*2} = \frac{n}{n-1}\left[\frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2\right]$$

$X_1, \ldots, X_n$ független minta esetén a likelihood-függvény:
$f(\theta) = \prod_{i=1}^{n} f_\theta(X_i)$, ha a minta folytonos eloszlásból származik, és $f_\theta(x)$ a θ paraméterhez tartozó sűrűségfüggvény;
$f(\theta) = \prod_{i=1}^{n} P_\theta(\xi = X_i)$, ha a minta diszkrét eloszlásból származik, ahol ξ eloszlása megegyzik X_1 eloszlásával, és független a mintától.

A log-likelihood függvény: $l(\theta) = \log f(\theta)$.

A minta Fisher-információja: $I(\theta) = E_\theta\left[\left(\frac{d}{d\theta} l(\theta)\right)^2\right] = D_\theta^2\left(\frac{d}{d\theta} l(\theta)\right)$ megfelelő feltételek mellett. Megfelelő regularitási feltételek mellett n elemű független minta Fisher-információja az egyelemű minta Fisher-információjának n-szerese.

$N(m, \sigma)$ normális eloszlásból származó $X_1, \ldots, X_n$ független minta esetén legyen u_y olyan, hogy $\Phi(u_y) = y$. Ekkor

$$P\left(\overline{X} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le m \le \overline{X} + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \ge 1 - \alpha.$$

$X_1, X_2, \ldots, X_n$, $Y_1, Y_2, \ldots, Y_m$ független normális eloszlású minták

u-próba

Egymintás esetben $H_0 : m = m_0$ nullhipotézis mellett

$$u = \sqrt{n}\frac{\overline{X} - m_0}{\sigma} \sim N(0,1).$$

Kétmintás esetben $H_0 : m_1 = m_2$ nullhipotézis mellett

$$u = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1).$$

t-próba

Egymintás esetben $H_0 : m = m_0$ nullhipotézis mellett

$$t = \sqrt{n}\frac{\overline{X} - m_0}{s_n^*} \sim t_{n-1}.$$

Kétmintás esetben $H_0 : m_1 = m_2$ nullhipotézis mellett

$$t = \sqrt{\frac{mn(m+n-2)}{m+n}}\frac{\overline{X} - \overline{Y}}{\sqrt{(n-1)s_n^{*2}(X) + (m-1)s_m^{*2}(Y)}} \sim t_{n+m-2}.$$

F-próba

$H_0 : \sigma_1^2 = \sigma_2^2$ nullhipotézis mellett

$$F = \frac{s_n^{*2}(X)}{s_m^{*2}(Y)} \sim F_{n-1,m-1}.$$

χ^2-próbák

Illeszkedésvizsgálat

$A_1, A_2, \ldots, A_r$ teljes eseményrendszer, p_i adott számok, ν_i az A_i gyakorisága n elemű mintából
H_0: minden i-re $P(A_i) = p_i$

Ekkor H_0 mellett

$$\sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} \to \chi^2_{r-1} \quad \text{eloszlásban, ha } n \to \infty.$$

Becsléses illeszkedésvizsgálat

H_0 mellett

$$\sum_{i=1}^{r} \frac{(\nu_i - n\hat{p}_i)^2}{n\hat{p}_i} \to \chi^2_{r-s-1} \quad \text{eloszlásban, ha } n \to \infty,$$

ahol s a becsült paraméterek száma

Homogenitásvizsgálat

ν_i illetve μ_i az i. osztály gyakorisága az egyes mintákban. H_0 mellett

$$nm \sum_{i=1}^{r} \frac{(\nu_i/n - \mu_i/m)^2}{\nu_i + \mu_i} \to \chi^2_{r-1} \quad \text{eloszlásban, ha } n \to \infty.$$

Függetlenségvizsgálat

H_0 mellett

$$n \sum_{i,j} \frac{\left(\nu_{ij} - \frac{\nu_{i\cdot}\nu_{\cdot j}}{n}\right)^2}{\nu_{i\cdot}\nu_{\cdot j}} \to \chi^2_{(r-1)(s-1)} \quad \text{eloszlásban, ha } n \to \infty.$$

$r = s = 2$-re

$$n \frac{(\nu_{11}\nu_{22} - \nu_{12}\nu_{21})^2}{\nu_{1\cdot}\nu_{2\cdot}\nu_{\cdot 1}\nu_{\cdot 2}} \to \chi^2_1$$

A χ^2-próba kritikus értékei

f	0,1	0,05	0,01
1	2,71	3,84	6,63
2	4,61	5,99	9,21
3	6,25	7,81	11,3
4	7,78	9,49	13,3
5	9,24	11,1	15,1
6	10,6	12,6	16,8
7	12,0	14,1	18,5
8	13,4	15,5	20,1
9	14,7	16,9	21,7