

Programtervező informatikus BSc, C szakirány

Valószínűségszámítás és statisztika gyakorlat

1. (1-2 hét) Valószínűségek kiszámítása (ismétlés: kombinatorika); feltételes valószínűség és Bayes-tétel, diszkrét valószínűségi változók

Elmélet

Definíció (Ismétlés nélküli permutáció). n (különböző) elem egy sorrendje. A lehetőségek száma:

$$n!.$$

Definíció (Ismétléses permutáció). n (nem feltétlen különböző) elem egy sorrendje, ahol az egyforma elemeket nem különböztetjük meg (tfh. az n elem közül k_1, \dots, k_r darab megegyezik). A lehetőségek száma:

$$\frac{n!}{k_1! \cdots k_r!} = \binom{n}{k_1, \dots, k_r}.$$

Definíció (Ismétlés nélküli kombináció). n (különböző) elem közül k számú ($k \leq n$) elem egyszerre történő kiválasztása (sorrend nem számít, nincs visszatérés). A lehetőségek száma:

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}.$$

Definíció (Ismétléses kombináció). n (különböző) elemből visszatérési eljárással k számú ($k \leq n$) elem kiválasztása (sorrend nem számít). A lehetőségek száma:

$$\binom{n+k-1}{k}.$$

Definíció (Ismétlés nélküli variáció). n (különböző) elem közül kiválasztott valamely k számú ($k \leq n$) elem egy sorrendje (nincs visszatérés). A lehetőségek száma:

$$\frac{n!}{(n-k)!}.$$

Definíció (Ismétléses variáció). n (különböző) elem közül visszatérési eljárással kiválasztott valamely k számú ($k \leq n$) elem egy sorrendje. A lehetőségek száma:

$$n^k.$$

A valószínűség a matematikai fogalma annak, hogy mekkora esély van valamire, például egy ászt kihúzni egy kártyapakliból, vagy egy piros golyót kihúzni egy színes golyókkal teli zsákból. A klasszikus valószínűség azokat az eseteket nézi, amikor minden egyes lehetséges kimenetelhez ugyanakkora esély tartozik. Például, ha egy szabályos dobókockával dobunk, akkor ugyanakkora az esély arra, hogy 1, 2, 3, 4, 5, vagy 6-ost dobjunk. Ekkor egy tetszőleges esemény valószínűsége megadható a kedvező kimenetek és az összes lehetséges kimenetek számának hányadosával:

Klasszikus valószínűség: Az A esemény valószínűsége megadható úgy, mint $P(A) = \frac{\text{kedvező kimenetek száma}}{\text{összes kimenetel száma}}.$

Természetesen a későbbiekben ennél bonyolultabb esetekkel is fogunk találkozni, de az alapfogalmak megértéséhez ez a véges sok lehetőséget tartalmazó egyszerű modell is elegendő.

Definíció (Feltételes valószínűség).

Ha B bekövetkezett, mi a valószínűsége, hogy A bekövetkezik? $P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{ha } P(B) \neq 0$

Definíció (Események függetlensége).

A és B események függetlenek, ha

$$P(A \cap B) = P(A) \cdot P(B) \quad (A \text{ esemény bekövetkezése nem befolyásolja } B \text{ esemény bekövetkezését, és fordítva}).$$

Definíció (Teljes eseményrendszer).

B_1, B_2, \dots események teljes eseményrendszert alkotnak, ha **1)** $B_i \cap B_j = \emptyset \quad \forall i \neq j$ re **2)** $\bigcup_{i=1}^{\infty} B_i = \Omega$

Teljes valószínűség tétele:

Legyen B_1, B_2, \dots teljes eseményrendszer, A tetszőleges esemény, $P(B_j) > 0$ minden j -re. Ekkor

$$P(A) = \sum_{j=1}^{\infty} P(A|B_j)P(B_j).$$

Bayes-tétel:

Legyen B_1, \dots, B_n, \dots teljes eseményrendszer, A tetszőleges esemény, $P(B_j) > 0$ minden j -re. Ekkor

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{j=1}^{\infty} P(A|B_j)P(B_j)}.$$

Valószínűségi változók:

A kísérletek megfigyelése során bekövetkező különböző elemi eseményekhez különböző számértékeket rendelünk, például a mérőeszköz által mutatott értéket vagy két kocka dobásakor azoknak az összegét. Ekkor előfordulhat, hogy két különböző elemi eseményhez is rendelhetünk azonos számértéket, pl. két kocka dobásakor a $\{2, 3\}$ és $\{3, 2\}$ dobásokhoz is 8-ast rendelünk. Ezt a hozzárendelést nevezzük valószínűségi változónak, ami tehát egy függvény az elemi események halmazán, és így számszerűsíti a kísérlet eredményét.

Definíció (X valószínűségi változó eloszlásfüggvénye). $F_X(x) = P(X < x)$.

Az eloszlásfüggvény tulajdonságai:

- $0 \leq F_X(x) \leq 1$;
- monoton növekvő;
- balról folytonos;
- $\lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow \infty} F(x) = 1$.

Állítás Tetszőleges X valószínűségi változó esetén $P(a \leq X < b) = F(b) - F(a)$; $P(a < X \leq b) = F(b) - F(a)$; $P(X = b) = F(b) - F(b-)$.

Diszkrét eloszlások:

Definíció (Diszkrét valószínűségi változó). Értékkészlete legfeljebb megszámlálhatóan végtelen, azaz $\{x_1, \dots, x_n, \dots\}$ elemekből áll. Eloszlása: $p_i := P(X = x_i) = P(\omega : X(\omega) = x_i) \ (i = 1, \dots)$

Feladatok

1.1. Feladat. Hányféleképpen lehet 8 bástyát letenni egy sakktáblára, hogy ne üssék egymást?

1.2. Feladat. Mi a valószínűsége, hogy egy véletlenszerűen kiválasztott 6 jegyű szám jegyei mind különbözőek?

1.3. Feladat. Ha egy magyarkártya-csomagból (32 lap: piros, zöld, makk, tők) visszatevéssel húzunk három lapot, akkor mi annak a valószínűsége, hogy

- a) pontosan egy piros színű lapot húztunk?
- b) legalább egy piros színű lapot húztunk?

1.4. Feladat. Egy zsákban 10 pár cipő van. 4 db-ot kiválasztva, mi a valószínűsége, hogy van közöttük pár, ha

- a) egyformák a párok?
- b) különbözőek a párok?

1.5. Feladat. n dobozba véletlenszerűen helyezünk el n golyót úgy, hogy bármennyi golyó kerülhet az egyes dobozokba.

- a) Mi a valószínűsége, hogy minden dobozba kerül golyó?
- b) Annak mi a valószínűsége, hogy pontosan egy doboz marad üresen?

1.6. Feladat. Egy boltban 10 látszólag egyforma számítógép közül 3 felújított, a többi új. Mi a valószínűsége, hogy ha veszünk 5 gépet a laborba, akkor pontosan 2 felújított lesz közöttük?

1.7. Feladat. Ha a 6 karakteres jelszavunkat véletlenszerűen választjuk a 10 számjegy és a 26 karakter közül, akkor mi a valószínűsége, hogy pontosan 3 szám lesz benne?

1.8. Feladat. Az ötös lottónál adjuk meg annak a valószínűségét, hogy egy szelvényrel játszva öt találatosunk lesz, illetve hogy legalább négyesünk lesz. Mi a valószínűsége, hogy minden kihúzott szám páros? (Hogy viszonylik ez a visszatevéses esethez?)

1.9. Feladat. Mennyi a valószínűsége, hogy két kockadobásnál mind a két dobás 6-os, feltéve, hogy tudjuk, hogy legalább az egyik dobás 6-os?

1.10. Feladat. 41 millió ötöslottó-szelvényt töltenek ki egymástól függetlenül. Mennyi a valószínűsége, hogy lesz legalább egy 5-ös találat?

1.11. Feladat. 100 érme közül az egyik hamis (ennek mindkét oldalán fej található). Egy érmét véletlenszerűen kiválasztva és azzal 10-szer dobva, 10 fejet kaptunk. Ezen feltétellel mi a valószínűsége, hogy a hamis érmével dobtunk?

1.12. Feladat. Egy diák a vizsgán p valószínűséggel tudja a helyes választ. Amennyiben nem tudja, akkor tippel (az esélye, hogy ekkor eltalálja a helyes választ $\frac{1}{3}$). Ha helyesen válaszolt, mennyi a valószínűsége, hogy tudta a helyes választ?

1.13. Feladat. Egy számítógépes program két független részből áll. Az egyikben 0,2, a másikban 0,3 a hiba valószínűsége. Ha a program hibát jelez, akkor mi a valószínűsége, hogy mindkét rész hibás?

1.14. Feladat. Egy számítógép processzorát 3 üzemben készítik. 20% eséllyel az elsőben, 30% eséllyel a másodikban és 50% eséllyel a harmadikban. A garanciális hibák valószínűsége az egyes üzemekben rendre 10%, 4%, illetve 1%. Ha a gépünk processzora elromlott, akkor mi a valószínűsége, hogy az első üzemben készült?

1.15. Feladat. Tegyük fel, hogy az új internet-előfizetők véletlenszerűen választott 20%-a speciális kedvezményt kap. Mi a valószínűsége, hogy 10 ismerősünk közül, akik most fizettek elő, legalább négyen részesülnek a kedvezményben?

1.16. Feladat. Egy tétel áru 1% selejtet tartalmaz. Hány darabot kell taláalomra kivennünk és megvizsgálnunk, hogy a megvizsgált darabok között legalább 0,95 valószínűséggel selejtes is legyen, ha az egyes kiválasztott darabokat vizsgálatuk után visszatesszük?

1.17. Feladat. Dobjunk egy kockával annyiszor, ahány fejet dobtunk két szabályos érmével. Jelölje X a kapott számok összegét. Adjuk meg X eloszlását!

1.18. Feladat. Jelölje X az ötöslottón kihúzott lottószámok legkisebbikét. Adjuk meg X eloszlását!

1.19. Feladat. Egy érmével dobva (tfh. p a fej valószínűsége), jelölje X az első azonosakból álló sorozat hosszát. (Azaz pl., ha a sorozat FFI..., akkor $X = 2$.) Adjuk meg X eloszlását!

1.20. Feladat. Legyenek az X diszkrét valószínűségi változó értékei $-2, 1, 3$, a következő valószínűségekkel:

$$P(-2) = 1/2, \quad P(1) = 1/3, \quad P(3) = 1/6.$$

Számoljuk ki az $F(x)$ eloszlásfüggvényt!

2. (3-4 hét) Várható érték, szórás, abszolút folytonos eloszlások, függetlenség, aszimptotikus tulajdonságok

Elmélet

Definíció (Diszkrét valószínűségi változó várható értéke). Jelölése: EX .

Legyen X diszkrét valószínűségi változó, amely az x_1, x_2, \dots értékeket veszi fel, p_1, p_2, \dots valószínűségekkel, ekkor

$$EX = \sum_{k=1}^{\infty} x_k p_k, \text{ ha a végtelen összeg abszolút konvergens.}$$

Legyen X diszkrét valószínűségi változó, melyre $p_i = P(X = x_i)$, ekkor $EX^2 = \sum x_i^2 p_i$.

Állítás Ha EX és EY véges, $a, b \in \mathbb{R}$, akkor

$$E(aX + b) = aEX + b \text{ és}$$

$$E(X + Y) = EX + EY.$$

Definíció (X szórásnégyzete). $D^2X = E[(X - EX)]^2 = EX^2 - E^2X$

Definíció (X szórása). $DX = \sqrt{D^2X}$

Nevezetes diszkrét eloszlások:

Név (paraméterek)	Értékek (k)	$P(X = k)$	EX	D^2X
Indikátor (p) (= Binomiális ($1, p$))	0, 1	$p^k(1-p)^{1-k}$	p	$p(1-p)$
Binomiális (n, p)	0, 1, ..., n	$\binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Poisson (λ)	0, 1, ...	$\frac{\lambda^k}{k!} e^{-\lambda}$	λ	λ
Geometriai vagy Pascal (p) (= Negatív binomiális ($1, p$))	1, 2, ...	$p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negatív binomiális (n, p)	$n, n+1, \dots$	$\binom{k-1}{n-1} p^n (1-p)^{k-n}$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$
Hipergeometriai (N, M, n)	0, 1, ..., n	$\frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}$	$n \frac{M}{N}$	$n \frac{M}{N} \left(1 - \frac{M}{N}\right) \left(1 - \frac{n-1}{N-1}\right)$

Abszolút folytonos eloszlások:

Definíció (Abszolút folytonos valószínűségi változó). Ha létezik olyan $f(x)$ függvény, amelyre $F(x) = \int_{-\infty}^x f(t) dt$.

Ilyenkor $f(x)$ -et sűrűségfüggvénynek hívjuk. (Megjegyzés: Az f sűrűségfüggvény létezéséhez szükséges (de nem elégséges), hogy F folytonos legyen (azaz $P(X = x) = 0 \quad \forall x$ -re).)

Tétel. Legyen X abszolút folytonos eloszlású. Ekkor $f(x) = F'(x)$; $f(x) \geq 0$; $\int_{-\infty}^{\infty} f(x) dx = 1$; $P(X = x) = 0$

$\forall x$ -re;

$$P(a < X \leq b) = P(a \leq X < b) = F(b) - F(a).$$

Definíció (Várható érték). Legyen X abszolút folytonos valószínűségi változó $f(x)$ sűrűségfüggvénnyel, ekkor

$$EX = \int_{-\infty}^{\infty} x f(x) dx, \text{ ha az integrál létezik.}$$

Nevezetes abszolút folytonos eloszlások:

Név (paraméterek)	Értékek	Eloszlásfüggvény (F)	Sűrűségfüggvény (f)	EX	D^2X
Standard normális $N(0, 1)$	$(-\infty, \infty)$	$\Phi(x) = \text{táblázatban}$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad x \in \mathbb{R}$	0	1
Normális $N(m, \sigma^2)$	$(-\infty, \infty)$	visszavezethető $\Phi(x)$ -re	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} \quad x \in \mathbb{R}$	m	σ^2
Egyenletes $E[a, b]$	$[a, b]$	$\begin{cases} 0 & \text{ha } x \leq a \\ \frac{x-a}{b-a} & \text{ha } a < x \leq b \\ 1 & \text{ha } b < x \end{cases}$	$\begin{cases} \frac{1}{b-a} & \text{ha } a < x \leq b \\ 0 & \text{különben} \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponenciális $\text{Exp}(\lambda)$	$(0, \infty)$	$\begin{cases} 1 - e^{-\lambda x} & \text{ha } x \geq 0 \\ 0 & \text{különben} \end{cases}$	$\begin{cases} \lambda e^{-\lambda x} & \text{ha } x \geq 0 \\ 0 & \text{különben} \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma (k, λ)	$(0, \infty)$	nincs zárt elemi képlet	$\begin{cases} \frac{1}{\Gamma(k)} \lambda^k x^{k-1} e^{-\lambda x} & \text{ha } x \geq 0 \\ 0 & \text{különben} \end{cases}$	$\frac{k}{\lambda}$	$\frac{k}{\lambda^2}$

Normális eloszlás standardizálása: Legyen $X \sim N(m, \sigma^2)$, ekkor $\frac{X-m}{\sigma} \sim N(0, 1)$.

Függetlenség:

Definíció (Valószínűségi változók függetlensége). Az X_1, X_2, \dots, X_n valószínűségi változók függetlenek, ha bármely I_1, I_2, \dots, I_n intervallumra $P(X_1 \in I_1, \dots, X_n \in I_n) = \prod_{i=1}^n P(X_i \in I_i)$

Megjegyzés: Független valószínűségi változók függvényei is függetlenek lesznek.

Tétel (Valószínűségi változók függetlensége). (i) Az X_1, X_2, \dots, X_n valószínűségi változók pontosan akkor függetlenek, ha együttes eloszlásfüggvényük megegyezik eloszlásfüggvényeik szorzatával, azaz $F_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^n F_{X_i}(x_i) \quad \forall \mathbf{x}$ -re.

(ii) Az X_1, X_2, \dots, X_n diszkrét valószínűségi változók pontosan akkor függetlenek, ha

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i) \quad \forall x_i\text{-re.}$$

(iii) Az X_1, X_2, \dots, X_n abszolút folytonos valószínűségi változók pontosan akkor függetlenek ha

$$f(\mathbf{x}) = \prod_{i=1}^n f_{X_i}(x_i) \quad \forall \mathbf{x}\text{-re.}$$

Definíció (X és Y kovarianciája). $\text{cov}(X, Y) = E(XY) - EXEY$

Definíció (X és Y korrelációja). $R(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{DXDY}}$

Ha X és Y függetlenek $\Rightarrow \text{cov}(X, Y) = 0$, de fordítva nem igaz.

$$D^2(aX + b) = a^2 D^2X, \quad D^2(X + Y) = D^2(X) + D^2(Y) + 2\text{cov}(X, Y)$$

Egyenlőtlenségek:

Tétel (Markov-egyenlőtlenség). Legyen $g : \mathbb{R} \rightarrow \mathbb{R}$ monoton növekvő pozitív függvény, $X \geq 0$ valószínűségi változó, melyre $EX < \infty$ és $\varepsilon > 0$ tetszőleges. Ekkor

$$P(X \geq \varepsilon) \leq \frac{E(g(X))}{g(\varepsilon)}$$

Spec., ha $g(x) = x$, akkor

$$P(X \geq \varepsilon) \leq \frac{EX}{\varepsilon}$$

Tétel (Csebisev-egyenlőtlenség). Legyen X tetszőleges valószínűségi változó, melyre $D^2X < \infty$ és $\varepsilon > 0$ tetszőleges. Ekkor

$$P(|X - EX| \geq \varepsilon) \leq \frac{D^2X}{\varepsilon^2}$$

Aszimptotikus tulajdonságok:

Tétel (Nagy számok törvénye (NSZT)). Legyenek X_1, X_2, \dots független azonos eloszlású valószínűségi változók, $EX_1 = m < \infty$. Ekkor

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{n \rightarrow \infty} m \quad \text{1 valószínűséggel.}$$

Tétel (Centrális határeloszlás tétel (CHT)). Legyenek X_1, X_2, \dots független azonos eloszlású valószínűségi változók, $EX_1 = m$, $D^2X_1 = \sigma^2 < \infty$. Ekkor

$$\frac{X_1 + \dots + X_n - nm}{\sqrt{n}\sigma} \xrightarrow{n \rightarrow \infty} N(0, 1) \quad \text{gyengén,}$$

azaz

$$P\left(\frac{X_1 + \dots + X_n}{\sqrt{nm}} < x\right) \xrightarrow{n \rightarrow \infty} \Phi(x)$$

ahol Φ a standard normális eloszlás eloszlásfüggvénye.

Feladatok

2.1. Feladat. Tegyük fel, hogy a 3 valószínűségszámítás gyakorlatra rendre 15, 20, illetve 25 diák jár. Várhatóan mekkora egy véletlenszerűen kiválasztott diák csoportja?

2.2. Feladat. Két kockával dobunk. Egy ilyen dobást sikeresnek nevezünk, ha van 6-os a kapott számok között. Várhatóan hány sikeres dobásunk lesz n próbálkozásból?

2.3. Feladat. Tegyük fel, hogy egy dobozban van $2N$ kártyalap, melyek közül kettőn 1-es, kettőn 2-es szám van és így tovább. Válasszunk ki véletlenszerűen m lapot. Várhatóan hány pár marad a dobozban?

2.4. Feladat. Mennyi az ötöslottón kihúzott

- a) számok összegének várható értéke?
- b) páros számok számának várható értéke?

2.5. Feladat. Egy adott területről származó talajmintákban a spórák száma Poisson-eloszlású. A minták harmadában egyáltalán nincs spóra. Mi a valószínűsége annak, hogy egy mintában a spórák száma egyenél több? Mekkora a spórák számának várható értéke és szórása?

2.6. Feladat. Tegyük fel, hogy egy számítógép meghibásodási időpontja 0 és 10 év között van és itt geometriai modellel írható le. Határozzuk meg a jelenség eloszlásfüggvényét!

2.7. Feladat. Legyen $0 < Y < 3$ valószínűségi változó. Eloszlásfüggvénye ezen az intervallumon $F(x) = cx^3$. Mennyi c és $P(1 < Y < 1)$?

2.8. Feladat. Legyen X egy abszolút folytonos valószínűségi változó a $[0, c]$ intervallumon, sűrűségfüggvénye:

$$f(x) = \begin{cases} \frac{1}{9}x^2, & \text{ha } 0 \leq x < c \\ 0, & \text{ha } x < 0 \text{ vagy } x \geq c. \end{cases}$$

Határozza meg c -t és X eloszlásfüggvényét!

2.9. Feladat. Véletlenszerűen választunk egy pontot az $x^2 + y^2 < 1$ kör belsejében. Jelölje Z a távolságát a középponttól. Adjuk meg Z eloszlás- és sűrűségfüggvényét valamint várható értékét!

2.10. Feladat. Tapasztalatok szerint az út hossza, amit egy bizonyos típusú robogó megtesz az első meghibásodásáig exponenciális eloszlású valószínűségi változó. Ez a távolság átlagosan 6000 km. Mi a valószínűsége annak, hogy egy véletlenszerűen kiválasztott robogó

- a) kevesebb, mint 4000 km megtétele után meghibásodik?
- b) több, mint 6500 km megtétele után hibásodik meg?
- c) 4000 km-nél több, de 6000 km-nél kevesebb út megtétele után hibásodik meg?
- d) Legfeljebb mekkora utat tesz meg az első meghibásodásig a robogók leghamarabb meghibásodó 20%-a?

2.11. Feladat. Egy tehén napi tejhozamát normális eloszlású valószínűségi változóval, $m = 22,1$ liter várható értékkel és $\sigma = 1,5$ liter szórással, modellezzük.

- a) Mi annak a valószínűsége, hogy egy adott napon a tejhozam 23 és 25 liter közé esik?
- b) Mekkora valószínűséggel esik a napi tejhozam $m - \sigma$ és $m + \sigma$ közé?

$$(\Phi(0,6) = 0,7257, \Phi(1,93) = 0,9732, \Phi(1) = 0,8413)$$

2.12. Feladat. Mennyi garanciát adjunk, ha azt szeretnénk, hogy termékeink legfeljebb 10%-át kelljen garanciaidőn belül javítani, ha a készülék élettartama 10 év várható értékű és 2 év szórású normális eloszlással közelíthető.

Standard normális eloszlás eloszlásfüggvényének értékei: https://zemleni.elte.hu/st_normelo.p f

2.13. Feladat. Legyen X egyenletes eloszlású az $[1, 4]$ intervallumon Számítsuk ki $(X - 1)^2$ várható értékét!

2.14. Feladat. Legyen X és Y független valószínűségi változók mindkettő 0 várható értékkel és 1 szórással. Legyen $W = X - Y$. Számítsa ki W várható értékét és szórását!

2.15. Feladat. Adjon meg véges sok értéket felvehető (X) ill. végtelen sok értéket felvehető (Y) diszkrét valószínűségi változókat melyeknek szórása 1!

2.16. Feladat. Legyen $X \sim N(2, \sqrt{5}^2)$ és $Y \sim N(5, 3^2)$ függetlenek és legyen $W = 3X - 2Y + 1$. Számítsa ki

a) EW -t és D^2W -t, ill.

b) $P(W \leq 6)$ -ot!

($\Phi(1) = 0,8413$)

2.17. Feladat. Legyen X és Y független valószínűségi változók, melyre $D^2X < \infty$ és $D^2Y < \infty$.

a) Mutassa meg, hogy $X + Y$ és X kovarianciája egyenlő X szórásnégyzetével!

b) Számolja ki $X + Y$ és X korrelációját!

2.18. Feladat. Tegyük fel, hogy egy tábla csokoládé tömege normális eloszlású 100g várható értékkel és 3g szórással. Legalább hány csokoládét csomagoljunk egy dobozba, hogy a dobozban levő táblák átlagos tömege legalább 0,9 valószínűséggel nagyobb legyen 99,5 g-nál, ha feltételezzük, hogy az egyes táblák tömege egymástól független? ($\Phi(1,28) = 0,8997$)

2.19. Feladat. Egy scannelt kép átlagos mérete 600KB, 100KB szórással. Mi a valószínűsége, hogy 80 ilyen kép együttesen 47 és 48MB közötti tárhelyet foglal el, ha feltételezzük, hogy a képek mérete egymástól független?

($\Phi(1,12) = 0,8686$)

2.20. Feladat. Egy szoftver frissítéséhez 68 file-t kell installálni, amik egymástól függetlenül 10mp várható értékű és 2mp szórással ideig töltnének.

a) Mi a valószínűsége, hogy a teljes frissítés lezajlik 12 percen belül?

b) A cég a következő frissítésnél azt ígéri, hogy az már 95% valószínűséggel 10 percen belül betöltődik. Hány file-ból állhat ez a frissítés?

($\Phi(2,42) = 0,992$, $\Phi(1,645) = 0,95$)

2.21. Feladat. Legyen egy X pozitív valószínűségi változó várható értéke $EX = 3$ és szórása $DX = 3$. Számítsuk ki, hogy legfeljebb mekkora valószínűséggel vesz fel a változó 13-at vagy annál nagyobb értéket! Mennyi a valószínűség pontos értéke, ha feltesszük, hogy az eloszlás exponenciális?

2.22. Feladat. Egy elektromos vezetékgyártó cég 40 m-es vezetékeket gyárt 0,2 m szórással. Legfeljebb mennyi annak a valószínűsége, hogy a vezeték hossza legalább 1 m-rel eltér a várható 40 m-es értéktől?

3. (6-7 hét) Leíró statisztikák, statisztikai alapfogalmak, becslések (maximum likelihood, momentum)

Elmélet

Definíció (Minta). X_1, \dots, X_n valószínűségi változó sorozat. A továbbiakban feltesszük, hogy függetlenek és azonos eloszlásúak. Realizációja: x_1, \dots, x_n

Definíció (Statisztika). A minta valamely függvénye, pl.:

Mintaátlag v. átlag: $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

Tapasztalati szórás: $S_n = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$

Korrigált tapasztalati szórás: $S_n^* = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$

Szórási együttható (vagy relatív szórás): $V = \frac{S_n}{\bar{X}} = \frac{S_n^*}{\bar{X}} 100\%$ (az átlagtól való átlagos eltérés százalékban)
/megjegyzés: lehet a korrigált tapasztalati szórással is számolni/

k-adik tapasztalati momentum ($k \geq 1, k \in \mathbb{Z}$): $m_k = \frac{1}{n} \sum_{i=1}^n X_i^k$

Tapasztalati módusz: a legtöbbször előforduló érték

Rendezett minta: $X_1^* \leq \dots \leq X_n^*$ a mintaelemek nem csökkenő sorrendben

Tapasztalati medián: $X_{\frac{n+1}{2}}^*$, ha n páratlan és $\frac{X_{\frac{n}{2}}^* + X_{\frac{n}{2}+1}^*}{2}$, ha n páros

Terjedelem: $R = X_n^* - X_1^*$ (legnagyobb legkisebb mintaelem)

z-kvantilis: $q_z = \inf\{x : F(x) \geq z\}$. Ha F invertálható, akkor $q_z = F^{-1}(z)$.

Tapasztalati z-kvantilis: q_z értelmezése: a mintaelemek z -ed része legfeljebb a q_z , $(1-z)$ -ed része pedig legalább a q_z értéket veszi fel ($0 < z < 1$); sokféleképpen számolható, pl. interpolációs módszerrel: először megállapítjuk a sorszámot: $(n+1)z = e + t$ (e : egészrész, t : törtrész), majd kiszámoljuk a z -kvantilist: $q_z = X_e^* + t(X_{e+1}^* - X_e^*)$.

Kvartilisek: Speciális kvantilisok, alsó (vagy első) kvartilis: $Q_1 = q_{\frac{1}{4}}$,
medián: $Q_2 = q_{\frac{1}{2}}$,
felső (vagy harmadik) kvartilis: $Q_3 = q_{\frac{3}{4}}$

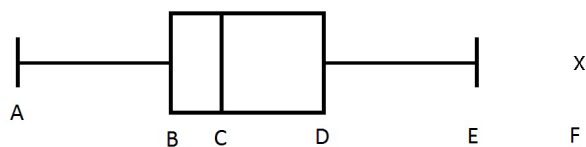
Interkvartilis terjedelem: $IQR = q_{\frac{3}{4}} - q_{\frac{1}{4}} = Q_3 - Q_1$

Tapasztalati eloszlásfüggvény: $F_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i < x)$

ahol $I(X_i < x) = \begin{cases} 1 & \text{ha } X_i < x \\ 0 & \text{ha } X_i \geq x \end{cases}$ indikátor függvény

Tétel (Glivenko-Cantelli). Az $F_n(x)$ tapasztalati eloszlásfüggvény és az $F(x)$ elméleti eloszlásfüggvény közötti eltérés maximuma 1 valószínűséggel 0-hoz konvergál, ami azt jelenti, hogy elég nagy minta esetén $F_n(x)$ értéke minden x -re tetszőleges közel van $F(x)$ értékéhez és n -et növelve mindenütt annak közelében marad.

Definíció (Boxplot).



$$A = \max\{x_1^*, Q_1 - 1.5 \cdot IQR\}, \quad B = Q_1, \quad C = Q_2, \quad D = Q_3, \quad E = \min\{x_n^*, Q_3 + 1.5 \cdot IQR\}$$

F: kieső értékek, azokat tüntetjük fel pontokként, amik A-n vagy E-n kívülre esnek

Legyenek X_1, X_2, \dots, X_n független, azonos eloszlású valószínűségi változók (minta) egy ϑ paraméterrel és legyen $\mathbf{x} = (X_1, X_2, \dots, X_n)$. A becslés a minta eloszlásának ismeretlen paraméterét közelíti a minta segítségével.

Definíció (Torzítatlan becslés). A ϑ valós paraméter $T(\cdot)$ becslése torzítatlan, ha $E(T(\cdot)) = \vartheta$ minden ϑ paraméterértékre.

Definíció (Likelihood függvény). $L(\vartheta; \mathbf{x}) = f_{\vartheta}(\mathbf{x}) = \prod_{i=1}^n f_{\vartheta}(x_i)$, ha az eloszlás abszolút folytonos

$$L(\vartheta; \mathbf{x}) = P_{\vartheta}(\mathbf{x} = \mathbf{x}) = \prod_{i=1}^n P_{\vartheta}(X_i = x_i), \text{ ha az eloszlás diszkrét}$$

Definíció (Log-likelihood függvény). $l(\vartheta; \mathbf{x}) = \ln(L(\vartheta; \mathbf{x}))$

Paraméterbecslési módszerek:

Maximum likelihood módszer (ML-módszer):

Azt a paraméterértéket keressük, ahol a likelihood függvény a legnagyobb értéket veszi fel (azaz diszkrét esetben az ismeretlen paraméter azon értéket keressük, amely mellett a bekövetkezett eredmény maximális valószínűségű): $\max_{\vartheta} L(\vartheta; \mathbf{x})$. Ez nyilván megegyezik azzal a paraméterértékkel, ahol a log-likelihood függvény veszi fel a legnagyobb értéket, azaz: $\max_{\vartheta} l(\vartheta; \mathbf{x})$.

Amennyiben a függvény deriválható ϑ szerint, akkor a maximumot kereshetjük a szokásos módon, a deriváltak segítségével, azonban a feladatunkat jelentősen megnehezíti, hogy olyan n -szere szorzatot kellene deriválni, amelynek minden tagjában ott van az a változó, ami szerint deriválnunk kellene. Ezért likelihood függvény helyett a log-likelihood függvény maximumhelyét keressük.

Ha ϑ 1 dimenziós, akkor $\partial_{\vartheta} l(\vartheta, \mathbf{x}) = 0$, míg ha $\vartheta = (\vartheta_1, \dots, \vartheta_p)$ p dimenziós, akkor $\partial_{\vartheta_i} l(\vartheta, \mathbf{x}) = 0$ megoldásából kapjuk a becslést. (A második deriváltak segítségével ellenőrizzük, hogy valóban maximum.)

Tétel (ML-becslés invariáns tulajdonsága). Ha ϑ ML-becslése $\hat{\vartheta}$, akkor tetszőleges g függvény esetén $g(\vartheta)$ ML-becslése $g(\hat{\vartheta})$.

Momentum módszer:

A mintából számítható tapasztalati momentumokat ($m_i := \frac{1}{n} \sum_j x_j^i$) egyenlővé tesszük az elméleti momentumokkal ($M_i(\vartheta) := E_{\vartheta} X^i$), mégpedig annyit, amennyiből a paramétereket meg tudjuk határozni. p darab ismeretlen paraméter esetén tipikusan p ismeretlenes egyenletrendszert oldunk meg ϑ -ra: $M_1(\vartheta) = m_1, \dots, M_p(\vartheta) = m_p$ (megjegyzés: $m_1 = \bar{x}$)

Feladatok

3.1. Feladat. Legyen X_1, \dots, X_n független, azonos eloszlású valószínűségi változók m várható értékkel. Célunk az ismeretlen m paraméter becslése. Tekintsük az alábbi statisztikákat és állapítsuk meg, hogy melyek torzítatlanok! Amelyik nem torzítatlan, hogyan tudnánk torzítatlanná tenni?

$$T_1(\cdot) = X_8, \quad T_2(\cdot) = \frac{X_9 + X_{19}}{9}, \quad T_3(\cdot) = \bar{X}$$

3.2. Feladat. Adjon torzítatlan becslést a független, azonos $E[0, \vartheta]$ eloszlású X_1, \dots, X_n minta ϑ paraméterére a mintaátlag segítségével!

3.3. Feladat. Legyen az alábbi gyakorisági tábla egy 20 elemű minta, a következő diszkrét eloszlásból:

$P(X_i = -1) = c, P(X_i = 1) = 3c, P(X_i = 2) = 1 - 4c \quad i = 1, \dots, 20$ és c az ismeretlen paraméter, $0 < c < \frac{1}{4}$.

érték	1	1	2
gyakoriság	4	10	6

Határozza meg c ML-becslését és c becslését a momentum módszerrel!

3.4. Feladat. Legyenek X_1, X_2, \dots, X_n független azonos eloszlású valószínűségi változók az alábbi eloszlásokból. Számolja ki az ismeretlen paraméter ML-becslését!

- $Bin(m, p)$ binomiális eloszlás, ahol $m \in \mathbb{N}$ adott és p a paraméter
- $Exp(\lambda)$ exponenciális eloszlás
- $N(\mu, \sigma^2)$ normális eloszlás, ahol $\sigma \in \mathbb{R}$ adott és μ a paraméter

3.5. Feladat. Határozza meg az ismeretlen paraméter ML-becslését, ha a minta $E[a, 1]$ eloszlású!

3.6. Feladat. Legyenek X_1, X_2, \dots, X_n független azonos $E[a, b]$ eloszlású valószínűségi változók. Számolja ki az ismeretlen paraméterek becslését a momentum módszerrel!

4. (8-9 hét) Konfidenciaintervallumok, paraméteres próbák

Elmélet

Definíció (Konfidenciaintervallum a normális eloszlás várható értékére). Legyenek $X_1, X_2, \dots, X_n \sim N(m, \sigma^2)$ független azonos eloszlású valószínűségi változók (tfh. σ ismert). Ekkor az $(1 - \frac{\alpha}{2})100\%$ -os konfidenciaintervallum m -re: $\bar{X} \pm u_{1 - \frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$, ahol $u_{1 - \frac{\alpha}{2}}$ a standard normális $1 - \frac{\alpha}{2}$ -kvantilisét jelöli.

Hipotézisvizsgálat

Hipotézis: állítás, aminek igazságát vizsgálni szeretnénk

Statisztikai próba: eljárás aminek a segítségével döntést hozhatunk a hipotézisről

Legyen $\mathbf{X} = (X_1, \dots, X_n)$ független, azonos eloszlású minta. Jelölje \mathcal{X} a mintateret, azaz a minta lehetséges értékeinek halmazát.

Nullhipotézis: $H_0 : \vartheta \in \Theta_0$

Ellenhipotézis: $H_1 : \vartheta \in \Theta_1$

Paraméterter: $\Theta = \Theta_0 \cup \Theta_1$

Döntés: $T(\mathbf{X})$ statisztika ($T : \mathcal{X} \rightarrow \mathbb{R}$ próbastatisztika) segítségével, melynek ismerjük az eloszlását a nullhipotézis fennállása esetén

Mintateret két részre bontjuk: $\mathcal{X} = \mathcal{X}_e \cup \mathcal{X}_k$ és $\mathcal{X}_e \cap \mathcal{X}_k = \emptyset$

\mathcal{X}_k : kritikus tartomány azon \mathbf{X} megfigyelések halmaza, amikre elutasítjuk a nullhipotézist

\mathcal{X}_e : elfogadási tartomány azon \mathbf{X} megfigyelések halmaza, amikre elfogadjuk a nullhipotézist

Kritikus érték: c (függ ϑ -tól, ld. alább)

$\mathcal{X}_k = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) \geq c\}$ vagy $\mathcal{X}_k = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) \leq c\}$ vagy $\mathcal{X}_k = \{\mathbf{x} \in \mathcal{X} : |T(\mathbf{x})| \geq c\}$

$\mathcal{X}_e = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) < c\}$ $\mathcal{X}_e = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) > c\}$ $\mathcal{X}_e = \{\mathbf{x} \in \mathcal{X} : |T(\mathbf{x})| < c\}$

Valós állapot	Döntés	
	H_0 -t elfogadjuk (\mathcal{X}_e)	H_0 -t elvetjük (\mathcal{X}_k)
H_0 igaz ($\vartheta \in \Theta_0$)	helyes döntés ($1 - \alpha$)	elsőfajú hiba (α)
H_0 hamis ($\vartheta \in \Theta_1$)	másodfajú hiba (β)	helyes döntés ($1 - \beta$)

Elsőfajú hiba valószínűsége:

Egyszerű hipotézis (Θ_0 halmaz egyelemű) esetén: $\mathbb{P}_{\vartheta_0}(\mathbf{X} \in \mathcal{X}_k) = \alpha$ $\vartheta_0 \in \Theta_0$ $\alpha = \mathbb{P}(\text{elvetjük } H_0\text{-t} \mid H_0 \text{ igaz})$

Összetett hipotézis (Θ_0 halmaz több elemű) esetén: $\mathbb{P}_{\vartheta}(\mathbf{X} \in \mathcal{X}_k) \leq \alpha$ $\forall \vartheta \in \Theta_0$

Próba (pontos) terjedelme vagy szignifikanciaszintje: $\alpha = \sup\{\mathbb{P}_{\vartheta}(\mathbf{X} \in \mathcal{X}_k) : \vartheta \in \Theta_0\}$

Megbízhatósági (konfidencia-) szint: $1 - \alpha$ $\alpha = \mathbb{P}(\text{elfogadjuk } H_0\text{-t} \mid H_0 \text{ igaz})$

A próba meghatározása: előre rögzített α terjedelemhez azt a c értéket keressük, amire a próba pontos terjedelme éppen α .

Másodfajú hiba valószínűsége:

$\beta(\vartheta) = \mathbb{P}_{\vartheta}(\mathbf{X} \in \mathcal{X}_e) = 1 - \mathbb{P}_{\vartheta}(\mathbf{X} \in \mathcal{X}_k)$ $\vartheta \in \Theta_1$ $\beta = \mathbb{P}_{\vartheta}(\text{elfogadjuk } H_0\text{-t} \mid H_0 \text{ hamis})$

Erőfüggvény: $\psi(\vartheta) = 1 - \beta(\vartheta)$ $\beta = \mathbb{P}(\text{elvetjük } H_0\text{-t} \mid H_0 \text{ hamis})$

Minél erősebb a próba, annál nagyobb valószínűséggel veti el a hamis nullhipotézist. Vagyis a próba ereje annak a valószínűsége, hogy egy adott eltérést adott mintanagyság és terjedelem mellett egy statisztikai próba kimutat. (Kísérletek tervezésekor az erő nagyságának előre meghatározott értékéből határozható meg a szükséges mintanelemszám.) A próba erejét addig nem tudjuk kiszámolni, ameddig az ellenhipotézis egy értékét nem rögzítjük ill. nem mondjuk meg a különbség nagyságát, amit ki szeretnénk mutatni.

p-érték: annak a valószínűsége, hogy igaz H_0 esetén a próbastatisztikára a tapasztalt vagy annál nagyobb eltérést kapunk. Ha egy próbát számítógép segítségével végzünk el, rendszerint a p-érték révén tudunk dönteni: ha $p\text{-érték} < \alpha$, akkor elvetjük H_0 -t.

A hipotézisek nem egyenrangúak. H_0 -t csak indokolt esetben szeretnénk elutasítani, így az elsőfajú hiba súlyosabbnak számít, mint a másodfajú hiba (különösen, ha csak kicsi az eltérés a H_0 -hoz képest). Itt általában az elsőfajú hiba legnagyobb valószínűségét adjuk meg, de a másodfajú hiba csökkentésére is törekszünk (pl. mintanagyság növelésével).

elfogad sa: statisztikailag nem találtunk komoly bizonyítékot arra, hogy H_0 nem lenne igaz; vagyis H_0 elfogadása esetén sem lehet állítani, hogy H_0 teljesül

elvetése: statisztikailag komoly bizonyítékot találtunk arra, hogy a H_0 nem igaz, azaz H_1 igaz

Prób k norm lis eloszl s v rható értékére

Egymint s prób k

$$X_1, \dots, X_n \sim N(m, \sigma^2)$$

$$H_0 : m = m_0$$

$$H_1 : m \neq m_0$$

$$H_0 : m \leq m_0$$

$$H_1 : m > m_0$$

$$H_0 : m \geq m_0$$

$$H_1 : m < m_0$$

Egymintás u-próba (σ ismert)

Próbastatisztika: $T(\mathbf{X}) = u = \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}} \stackrel{H_0 \text{ esetén}}{\sim} N(0, 1)$

Kritikus tartományok:

$$\mathcal{X}_k = \{ : |u| > u_{1-\frac{\alpha}{2}} \} \quad \mathcal{X}_k = \{ : u > u_{1-\alpha} \} \quad \mathcal{X}_k = \{ : u < -u_{1-\frac{\alpha}{2}} \}$$

Kapcsolat a konfidenciaintervallummal (az alábbi lépések ekvivalensek):

$$\boxed{|u| > u_{1-\frac{\alpha}{2}}} \Leftrightarrow u > u_{1-\frac{\alpha}{2}} \text{ vagy } u < -u_{1-\frac{\alpha}{2}} \Leftrightarrow \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}} > u_{1-\frac{\alpha}{2}} \text{ vagy } \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}} < -u_{1-\frac{\alpha}{2}} \Leftrightarrow$$

$$\bar{X} - m_0 > u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \text{ vagy } \bar{X} - m_0 < -u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \Leftrightarrow \boxed{m_0 \notin \left(\bar{X} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)}$$

Vagyis a null hipotézist pontosan akkor utasítjuk el, ha a $(1 - \alpha)100\%$ -os konfidenciaintervallum nem tartalmazza m_0 -t.

Egymintás t-próba (σ ismeretlen)

Próbastatisztika: $T(\mathbf{X}) = t = \frac{\bar{X} - m_0}{\frac{s_n^*}{\sqrt{n}}} \stackrel{H_0 \text{ esetén}}{\sim} t_{n-1}$

Kritikus tartományok:

$$\mathcal{X}_k = \{ : |t| > t_{n-1, 1-\frac{\alpha}{2}} \} \quad \mathcal{X}_k = \{ : t > t_{n-1, 1-\alpha} \} \quad \mathcal{X}_k = \{ : t < -t_{n-1, 1-\frac{\alpha}{2}} \}$$

Kétmint s prób k

$$X_1, \dots, X_n \sim N(m_1, \sigma_1^2), \quad Y_1, \dots, Y_m \sim N(m_2, \sigma_2^2) \quad \text{függetlenek}$$

$$H_0 : m_1 = m_2$$

$$H_1 : m_1 \neq m_2$$

$$H_0 : m_1 \leq m_2$$

$$H_1 : m_1 > m_2$$

$$H_0 : m_1 \geq m_2$$

$$H_1 : m_1 < m_2$$

	a két minta független	a két minta páronként összetartozó, nem független
σ_1 és σ_2 ismert	Kétmintás u-próba	Egymintás u-próba a különbségekre
σ_1 és σ_2 ismeretlen	előzetes F-próba	
	$\sigma_1 = \sigma_2$ Kétmintás t-próba	$\sigma_1 \neq \sigma_2$ Welch-próba
		Egymintás t-próba a különbségekre

előzetes F-próba (σ_1, σ_2 ismeretlen)

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

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Próbastatisztika:

$$F = \begin{cases} \frac{s_1^{*2}}{s_2^{*2}} & H_0 \text{ esetén } F_{n-1, m-1} & \text{ha } s_1^* > s_2^* \\ \frac{s_2^{*2}}{s_1^{*2}} & H_0 \text{ esetén } F_{m-1, n-1} & \text{ha } s_2^* > s_1^* \end{cases}$$

Kétmintás u-próba (σ_1, σ_2 ismert)

Próbastatisztika: $u = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \stackrel{H_0 \text{ esetén}}{\sim} N(0, 1)$

Kétmintás t-próba ($\sigma_1 = \sigma_2$ ismeretlen)

Próbastatisztika: $t = \sqrt{\frac{nm}{n+m}} \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{n-1}{n+m} s_1^{*2} + \frac{m-1}{n+m} s_2^{*2}}} \stackrel{H_0 \text{ esetén}}{\sim} t_{n+m-2}$

Welch-próba ($\sigma_1 \neq \sigma_2$ ismeretlen)

Próbastatisztika: $t' = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^{*2}}{n} + \frac{s_2^{*2}}{m}}} \stackrel{H_0 \text{ esetén}}{\sim} t_f$, ahol $f \approx \frac{\left(\frac{s_1^{*2}}{n} + \frac{s_2^{*2}}{m} \right)^2}{\left(\frac{s_1^{*2}}{n} \right)^2 + \left(\frac{s_2^{*2}}{m} \right)^2}$

Feladatok

4.1. Feladat. Legyen X_1, X_2, X_3, X_4 független azonos $N(\mu, 2^2)$ eloszlású minta. A megfigyelt értékek a következők:

14,8; 12,2; 16,8; 11,1

a) Adjon 95%-os megbízhatóságú konfidenciaintervallumot μ -re!

b) Hány elemű mintára van szükség, ha azt szeretnénk, hogy a konfidenciaintervallum legfeljebb 1,6 hosszúságú legyen?
($u_{0,975} = 1,96$)

4.2. Feladat. Azt szeretnénk vizsgálni, hogy a napi középhőmérséklet október 18-án Budapesten 15°C alatt van-e? Az elmúlt 4 év napi középhőmérsékletei a következők voltak: 14, 8; 12, 2; 16, 8; 11, 1 $^{\circ}\text{C}$, valamint tegyük fel, hogy az adatok normális eloszlásból származnak.

a) Írjuk fel a null- és ellenhipotézist!

b) Tegyük fel, hogy a napi középhőmérséklet szórása $\sigma = 2$. Tesztelje a fenti hipotézist $\alpha = 0.05$ terjedelem mellett!

Adja meg a kritikus tartományt és p-értéket! Mi a döntés?

c) Teszteljük a hipotézist úgy is, hogy nem használjuk a szórásra vonatkozó előzetes információt!

d) Milyen hipotézist írunk fel, ha azt szeretnénk vizsgálni, hogy a napi középhőmérséklet október 18-án

Budapesten 15°C -tól különbözik-e? Teszteljük a fenti adatok segítségével!

($u_{0,05} = 1,645$, $\Phi(1,275) = 0,899$, $t_{3;0,05} = 2,353$, $u_{0,975} = 1,96$)

4.3. Feladat. Az alábbi két minta két különböző gyáregységben tapasztalt selejtarányra vonatkozik (ezrelékben). Ílítható-e, hogy az „A” gyáregység jobban dolgozott? (Feltételezhetjük, hogy a minták normális eloszlásúak, függetlenek.)

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

($F_{9,9;0,975} = 4,026$, $t_{18;0,05} = 1,734$)

4.4. Feladat. Két szerveret hasonlítottunk össze. Az elsőn 30 futás átlagos ideje 6,7 mp volt, míg ettől függetlenül a másodikon 20 futásé 7,2 mp. Vizsgáljuk meg, hogy van-e szignifikáns különbség a két szerver sebessége közt, ha a futási idők szórása mindkét gépen 0,5 volt?

($u_{0,975} = 1,96$)

4.5. Feladat. Az alábbi két minta 10 forgalmas csomópont levegőjében található szennyezőanyag koncentrációra vonatkozó két adatsort tartalmaz. Az első sorban a november 15-i, a másodikban a november 29-i számok szerepelnek. Szignifikánsan változott-e a légszennyezettség?

november 15.	20,9	17,1	15,8	18,8	20,1	15,6	14,8	24,1	18,9	12,5
november 29.	21,4	16,7	16,4	19,2	19,9	16,6	15,0	24,0	19,2	13,2

($t_{9;0,975} = 2,262$)

5. (10-11 hét) Nemparaméteres próbák, egyszerű lineáris regresszió

Elmélet

Nemparaméteres próbák:

Diszkrét illeszkedésvizsgálat

Legyen X_1, \dots, X_n egy n elemű minta és tegyük fel, hogy a mintaelemek r különböző x_j ($j = 1, \dots, r$) értéket vehetnek fel. Továbbá jelölje ν_j ($j = 1, \dots, r$) az egyes értékek megfigyelt gyakoriságát, azaz n független megfigyelést osztályozunk valamilyen szempont szerint, r páronként diszjunkt osztályba. Az egyes osztályok feltételezett valószínűségei rendre p_1, \dots, p_r .

Osztályok	1	2	...	r	Összesen
Értékek	x_1	x_2	...	x_r	
Gyakoriságok	ν_1	ν_2	...	ν_r	n
Valószínűségek	p_1	p_2	...	p_r	1

Azt vizsgáljuk, hogy a minta eloszlása megegyezik-e a feltételezett eloszlással. Ismert eloszlás esetén tiszta illeszkedésvizsgálatot végzünk. Ha viszont az eloszlás paraméteres és csak az eloszláscsaládot ismerjük, a paraméter(ek)e)t viszont nem (pl. az a kérdés, hogy származhatnak-e az adatok p paraméterű binomiális eloszlásból), akkor becsléses illeszkedésvizsgálatot végzünk.

Tiszta illeszkedésvizsgálat:

$$H_0 : P(X_i = x_j) = p_j \quad j = 1, \dots, r$$

$$H_1 : \exists \text{ legalább egy } j \text{ melyre } P(X_i = x_j) \neq p_j$$

$$\text{Próbastatisztika: } T_n = \sum_{j=1}^r \frac{(\nu_j - np_j)^2}{np_j} \stackrel{H_0 \text{ esetén}}{\sim} \chi_{r-1}^2 \quad \text{Kritikus tartomány: } \mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi_{r-1, 1}^2\}$$

Becsléses illeszkedésvizsgálat:

Legyen ϑ egy s dimenziós paramétervektor, valamint legyen $\hat{\vartheta}$ a ϑ paramétervektor ML-becslése, és legyen $\hat{p}_j = p_j(\hat{\vartheta})$.

$$H_0 : P(X_i = x_j) = \hat{p}_j \quad j = 1, \dots, r$$

$$H_1 : \exists \text{ legalább egy } j \text{ melyre } P(X_i = x_j) \neq \hat{p}_j$$

$$\text{Próbastatisztika: } T_n = \sum_{j=1}^r \frac{(\nu_j - n\hat{p}_j)^2}{n\hat{p}_j} \stackrel{H_0 \text{ esetén}}{\sim} \chi_{r-s-1}^2 \quad \text{Kritikus tartomány: } \mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi_{r-s-1, 1}^2\}$$

Megjegyzés: Mivel a próba aszimptotikus, vigyáznunk kell arra, hogy a minta elemszáma elég nagy legyen. Ökölszabályként meg szokás követelni, hogy az ún. elméleti gyakoriság (np_j) legalább 5 legyen minden i -re. Ha ez nem teljesül, akkor a kis várt gyakoriságokkal rendelkező eseményeket összevonjuk.

Függetlenségvizsgálat

n független megfigyelést két szempont szerint osztályozunk, az 1. szempont szerint r osztály, míg a 2. szempont szerint s osztály van. Annak a valószínűsége, hogy egy megfigyelést az 1. szempont szerint az i -edik, a második szerint pedig a j -edik osztályba sorolunk, p_{ij} . Az ilyen tulajdonságú megfigyelések számát pedig ν_{ij} -vel jelöljük. Az osztályozási eljárás eredményét ún. kontingenciátábla formájában szokás megadni:

		2. szempont					Sorösszegek
		1	...	j	...	s	
1. szempont	1	ν_{11}	...	ν_{1j}	...	ν_{1s}	$\nu_{1\bullet}$
	\vdots	\vdots		\vdots		\vdots	\vdots
	i	ν_{i1}	...	ν_{ij}	...	ν_{is}	$\nu_{i\bullet}$
	\vdots	\vdots		\vdots		\vdots	\vdots
	r	ν_{r1}	...	ν_{rj}	...	ν_{rs}	$\nu_{r\bullet}$
Oszlopösszegek		$\nu_{\bullet 1}$...	$\nu_{\bullet j}$...	$\nu_{\bullet s}$	n

ν_{ij} = megfigyelések gyakorisága az (i, j) osztályban

$$\nu_{i\bullet} = \sum_{j=1}^s \nu_{ij} \quad \nu_{\bullet j} = \sum_{i=1}^r \nu_{ij}$$

Hasonlóan $p_{i\bullet}$ ill. $p_{\bullet j}$ a marginális eloszlást jelölik, tehát a $[p_{ij}]$ mátrix sor-, illetve oszlopösszegei: $p_{i\bullet} = \sum_{j=1}^s p_{ij} \quad p_{\bullet j} = \sum_{i=1}^r p_{ij}$

H_0 : a két szempont független egymástól, azaz $p_{ij} = p_{i\bullet} \cdot p_{\bullet j} \quad 1 \leq i \leq r, 1 \leq j \leq s$

H_1 : a két szempont nem független, azaz $p_{ij} \neq p_{i\bullet} \cdot p_{\bullet j}$ legalább egy (i, j) párra

$$\text{Próbastatisztika: } T_n = \sum_{i=1}^r \sum_{j=1}^s \frac{\nu_{ij} \frac{\nu_{i\bullet} \nu_{\bullet j}}{n}}{\frac{\nu_{i\bullet} \nu_{\bullet j}}{n}} \stackrel{H_0 \text{ esetén}}{\sim} \chi^2_{r-1, s-1}$$

$$\text{Kritikus tartomány: } \mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-1, s-1, 1}\}$$

Megjegyzés: Ha $r = s = 2$, akkor a próbastatisztika a következőképpen leegyszerűsödik:

$$T_n = n \cdot \frac{(\nu_{11}\nu_{22} - \nu_{12}\nu_{21})^2}{\nu_{1\bullet}\nu_{2\bullet}\nu_{\bullet 1}\nu_{\bullet 2}} \stackrel{H_0 \text{ esetén}}{\sim} \chi^2_1.$$

Homogenitásvizsgálat

Van két független mintánk (adatsorunk) az egyikben n , a másikban m megfigyeléssel. Valamilyen szempont szerint r , páronként diszjunkt osztályba soroljuk a megfigyeléseket. Az i -edik osztály valószínűsége p_i az 1. minta és q_i a 2. minta esetén ($i = 1, 2, \dots, r$). Legyenek az egyes osztályok gyakoriságai ν_1, \dots, ν_r az 1. minta és μ_1, \dots, μ_r a 2. minta esetén.

Osztályok	1	2	...	r	Összesen
1. minta					
Gyakoriságok	ν_1	ν_2	...	ν_r	n
Valószínűségek	p_1	p_2	...	p_r	1
2. minta					
Gyakoriságok	μ_1	μ_2	...	μ_r	m
Valószínűségek	q_1	q_2	...	q_r	1

Azt vizsgáljuk, hogy a két minta ugyanolyan eloszlás szerint sorolódik-e be az egyes osztályokba:

H_0 : a két eloszlás megegyezik, azaz $p_i = q_i$ $i = 1, \dots, r$

H_1 : a két eloszlás nem megegyezik meg, azaz \exists legalább egy i , hogy $p_i \neq q_i$

$$\text{Próbastatisztika: } T_{n,m} = nm \sum_{i=1}^r \frac{\frac{\nu_i}{n} \frac{\mu_i}{m}}{\frac{\nu_i + \mu_i}{n+m}} \stackrel{H_0 \text{ esetén}}{\sim} \chi^2_{r-1} \quad \text{Kritikus tartomány: } \mathcal{X}_k = \{\mathbf{x} : T_{n,m}(\mathbf{x}) > \chi^2_{r-1, 1}\}$$

Korreláció becslése:

Legyenek X_1, \dots, X_n és Y_1, \dots, Y_n n elemű minták. A korreláció becslése a minták alapján:

$$\text{Tapasztalati korrelációs együttható: } r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Egyszerű lineáris regresszió:

Adott $(x_1, y_1), \dots, (x_n, y_n)$ számpárookra szeretnénk egyenest illeszteni.

Modell: $y_i = a + bx_i + \varepsilon_i$, ahol $E\varepsilon_i = 0$ és $D^2\varepsilon_i = \sigma^2 < \infty$ ($i = 1, \dots, n$)

Cél: a és b becslése

Módszer: legkisebb négyzetek: $\min \sum_{i=1}^n (y_i - (a + bx_i))^2$

$$\text{Megoldás: } \hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \text{ ez torzítatlan } b\text{-re és a szórásnégyzete: } D^2(\hat{b}) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}, \text{ ez torzítatlan } a\text{-ra és a szórásnégyzete: } D^2(\hat{a}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)$$

Reziduálisok: $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{a} + \hat{b}x_i)$ ($i = 1, \dots, n$)

$$\text{Reziduális szórásnégyzet becslése: } \hat{\sigma}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

Determinációs együttható: $R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = r^2$ Az R^2 mutatja meg, hogy X változékonysága mennyire magyarázza Y változékonyságát. Értéke 0 és 1 között lehet, minél nagyobb, annál jobban teljesít a model, azaz annál jobban illeszkedik az egyenes.

Feladatok

5.1. Feladat. Egy gyárban egy termék minőségét 4 elemű mintákat véve ellenőrzik, havonta 300 mintavétellel. Megszámolták, hogy a legutóbbi hónapban hányszor volt selejtes a minta, melynek eredményét az alábbi táblázat tartalmazza:

Selejtesek száma	0	1	2	3	4
Darabszám	80	113	77	27	3

Modellezhető a mintákban levő selejtesek száma

a) $(4; 0, 25)$, ill.

b) $(4; p)$ paraméterű binomiális eloszlással ($p = 0, 05$)? $\chi^2_{3;0,95} = 7, 81$, $\chi^2_{2;0,95} = 5, 99$)

5.2. Feladat. Az alábbi kontingencia-táblázat mutatja, hogy egy 100 éves időszakban egy adott hónapban a csapadék mennyisége és az átlaghőmérséklet hogyan alakult:

Hőmérséklet Csapadék	kevés	átlagos	sok
hűvös	15	10	5
átlagos	10	10	20
meleg	5	20	5

A cellákban az egyes esetek gyakoriságai találhatók. $p = 0, 05$ mellett tekinthető-e a csapadékmennyiség és a hőmérséklet függetlennek? ($\chi^2_{4;0,95} = 9, 49$)

5.3. Feladat. Két dobókockával dobva az alábbi gyakoriságokat figyeltük meg:

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

$p = 0, 05$ mellett döntsünk arról, hogy tekinthető-e a két eloszlás azonosnak! $\chi^2_{5;0,95} = 11, 1$)

5.4. Feladat. A következő feladatot csak R-rel kell megoldani, a számolások tájékoztató jellegűek, ha valakit érdekel, nem szerepelnek a számonkérésen. Adottak a következő (x, y) párok:

x	0	1	6	5	3
y	4	3	0	1	2

a) Határozzuk meg és ábrázoljuk is az $a + bx$ alakú regressziós egyenest!

b) Mi az $x = 2$ -höz tartozó előrejelzett y érték?

c) Számoljuk ki a reziduálisokat és becsüljük meg a hiba szórásnégyzetét, valamint a becsléseink szórásnégyzetét!