Valószínűségszámítás és Számelmélet összefoglaló

1. Valószínűségi változó fogalma
   Formálisan az X : Ω → R függvényt nevezzük valószínűségi változónak.
   5.5 Definíció $\xi : \Omega \to \mathbb{R}$ **valószínűségi változó**, ha minden $x \in \mathbb{R}$ számra $\{w : \xi(w) < x\} \in \mathcal{A}$.
   5.6 Definíció A ξ valószínűségi változó **diszkrét**, ha értékkészlete véges vagy megszámlálható, azaz léteznek olyan x_k valós számok és A_k teljes eseményrendszer, hogy $\xi = \sum_k x_k \cdot \chi_{A_k}$.
2. Teljes valószínűség tétele
   Legyenek A1, ..., An események. Akkor mondjuk, hogy teljes eseményrendszert alkotnak, ha:
   1. páronként egymást kizárják;
   2. egyesítésük az Ω (biztos esemény)
3. Mikor független két esemény?
   Az A és a B esemény független, ha P(A ∩ B) = P(A) · P(B), ami éppen azt jelenti, hogy P(A|B) = P(A) (ha a feltételes valószínűség értelmes, azaz P(A) > 0).
4. Binomiális tétel
   Ha a és b tetszőleges valós számok és n pozitív egész szám, akkor:
   $$(a+b)^n = \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b^1 + \ldots + \binom{n}{i}a^{n-i}b^i + \ldots + \binom{n}{n-1}ab^{n-1} + \binom{n}{n}b^n.$$
   A binomiális tétel rövidebb alakja:
   $$(a+b)^n = \sum_{i=0}^{n}\binom{n}{i}a^{n-i}b^i$$
5. Valószínűségi változók függvényének várható értékét diszkrét és folytonos esetre is
   A pi =P (X=xi) eloszlással megadott valószínőségi változó várható értéke E(X):= p1x1+ p2x2 +…, ha a sor abszolút konvergens.
   5.11 Definíció $E\xi = \int_{\mathbb{R}} x \, dF_\xi(x)$ és $Eg(\xi) = \int_{\mathbb{R}} g(x) \, dF_\xi(x)$, ahol a $dF_\xi(x)$ szerinti integrálás a Lebesgue-Stieltjes-integrálást jelöli, ha $\int_{\mathbb{R}} |x| \, dF_\xi(x)$ illetve $\int_{\mathbb{R}} |g(x)| \, dF_\xi(x)$ véges.

   Abszolút folytonos esetben a várható érték a sűrűségfüggvény segítségével határozható meg. $E\xi = \int_{\mathbb{R}} x \cdot f_\xi(x) \, dx$ és $Eg(\xi) = \int_{\mathbb{R}} g(x) \cdot f_\xi(x) \, dx$.
   A következő példában a legnevezetesebb abszolút folytonos eloszlású valószínűségi változók várható értékét határozzuk meg.
6. Maximum-likelihood módszer
   A maximum likelihood módszer célja, hogy adott mérési értékekhez, az ismeretlen paramétereknek olyan becslését adja meg, amely mellett az adott érték a legnagyobb valószínűséggel következik be. Az eljárás a *likelihood függvény* maximalizálásával történik.

   A maximum likelihood becslés azokban az esetekben használatos amikor az egyes mérési eredmények olyan véletlen eseményekként interpretálhatóak, amelyek egy vagy több ismeretlen paramétertől függenek. Mivel a vizsgált értékek kizárólagosan az ismeretlen paraméter(ek)től függenek, előállíthatók ezen paraméter vagy paraméterek függvényeként. A mérést, becslést végző kutató ezt a paramétert határozza meg, így maximalizálja a mért minta által követett valószínűséget.
   A maximum likelihood módszer egy **X** valószínűségi változóból indul ki, amelynek a sűrűség- vagy tömegfüggvénye **f** és **q** paramétertől függ.
7. Likelihood függvény folytonos és diszkrét eloszlásra is

8. Egy valószínűségi változó X^2 – eloszlású

22.2. definíció: Legyenek $X_1, X_2, \ldots, X_n$ független, standard normális eloszlású valószínűségi változók. Ekkor a belőlük képzett

$$\chi_n^2 = X_1^2 + X_2^2 + \ldots + X_n^2$$

valószínűségi változó eloszlását n szabadsági fokú χ^2 (khí négyzet)-eloszlásnak nevezzük.

Az n szabadsági fokú χ^2-eloszlás sűrűségfüggvénye:

$$f_n(x) = \begin{cases} \dfrac{x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} & \text{ha } x > 0, \\ 0 & \text{ha } x \le 0. \end{cases}$$

Az n szabadsági fokú χ^2-eloszlású valószínűségi változó várható értéke és szórásnégyzete:

$$E(\chi_n^2) = n \qquad \text{és} \qquad D^2(\chi_n^2) = 2n\,.$$

9. Valószínűségi változó t-eloszlású
A Student-eloszlás (vagy másnéven t-eloszlás) adja a matematikai statisztikában használatos t-próba alapját.

22.3. definíció: Legyenek Y és $X_1, X_2, \ldots, X_n$ független, standard normális eloszlású valószínűségi változók. Ekkor a belőlük képzett

$$t_n = \frac{Y}{\sqrt{\dfrac{X_1^2 + X_2^2 + \ldots + X_n^2}{n}}}$$

valószínűségi változó eloszlását n szabadsági fokú Student-eloszlásnak (t-eloszlásnak) nevezzük.

Az n szabadsági fokú t-eloszlás sűrűségfüggvénye:

$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \frac{1}{\left(1 + \frac{x^2}{n}\right)^{\frac{n+1}{2}}}\,.$$

Az n szabadsági fokú t-eloszlású valószínűségi változó várható értéke:

$$E(t_n) = \begin{cases} 0 & \text{ha } n \ge 2, \\ \text{nem létezik} & \text{ha } n = 1. \end{cases}$$

Szórásnégyzete:

$$D^2(t_n) = \begin{cases} \dfrac{n}{n-2} & \text{ha } n \ge 3, \\ \text{nem létezik} & \text{ha } n = 1, 2. \end{cases}$$

10. Fisher-tétel
11. Konfidencia-intervallum

23.12. definíció: A (c_1, c_2) intervallumot az X valószínűségi változó α paraméterére vonatkozó $(1-\varepsilon) \cdot 100\%$ megbízhatósági szintű konfidencia-intervalumnak nevezzük, ha

$$P(c_1 < \alpha < c_2) = 1 - \varepsilon$$

teljesül.

12. Normális eloszlású minta konfidenciaintervalluma a várható értékre és a szórásnégyzetre.
13. Valószínűségi vektorváltozó és legfontosabb tulajdonságai

- Emlékeztető: $X=(X_1, X_2, \ldots, X_d): \Omega \to \boldsymbol{R}^d$ függvény valószínűségi vektorváltozó, ha $\{\omega: X(\omega) \in B\} \in \mathcal{A}$ minden B d-dimenziós Borel halmazra. (Pontosan akkor teljesül, ha X_i valószínűségi változó minden $1 \le i \le d$-re.)
- $Q_X(B) := P\{\omega: X(\omega) \in B\}$ az X *eloszlása* $\boldsymbol{R}^d$ Borel-halmazain.
  - Az $F_X(z) := P(X<z)$ $\boldsymbol{R}^d \to \boldsymbol{R}$ függvény az X valószínűségi vektorváltozó együttes eloszlásfüggvénye.
  - Az egydimenziós esettel analóg tulajdonságai:
    - $0 \le F_X(z) \le 1$
    - $F_X(z)$ minden koordinátájában monoton növő
    - $\lim F_X(z)=1$, ha z minden koordinátájára $z_i \to \infty$
    - $\lim F_X(z)=0$ ha z legalább egy koordinátájára $z_i \to -\infty$
    - $F_X(z)$ minden koordinátájában balról folytonos.

14. Valószínűségi mező fogalma

3.1. definíció: Ha egy kísérlettel kapcsolatban az elemi események száma véges (n), és minden elemi esemény valószínűsége egyenlő $\left(\frac{1}{n}\right)$, akkor a k féleképpen bekövetkező A esemény valószínűsége

$$P(A) = \frac{\text{kedvező esetek száma}}{\text{összes eset száma}} = \frac{k}{n}.$$

Ebben az esetben azt mondjuk, hogy az események és ezek valószínűségei klasszikus valószínűségi mezőt alkotnak.

15. Becslések aszimptotikus torzítatlansága
16. t-próba egy-és kétmintás esetre is

Legyen X $\mathcal{N}(m, \sigma^2)$ eloszlású valószínűségi változó, ahol az m várható érték és a σ szórás ismeretlenek. Az X valószínűségi változóra vonatkozó n elemű minta $X_1, X_2, \ldots, X_n$. A próba szintje $1-\alpha$. A hipotézis a várható értékre vonatkozik:

$$H_0 \;:\; m = m_0$$
$$H_1 \;:\; m \neq m_0 \text{ (kétoldali eset).}$$

Ismert, hogy $\frac{\overline{X}-m}{s_n^*}\sqrt{n}$ valószínűségi változó $(n-1)$ paraméterű t (Student)-eloszlású, ahol

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

és

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

Tehát ha a nullhipotézis igaz, a

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

A

próbastatisztika $(n-1)$ paraméterű t-eloszlású. Az $(n-1)$ paraméterű t-eloszlás táblázatából kiolvasható az a $t_{n-1}(\alpha/2)$ kritikus érték, amelyre

$$P(t_{n-1} \geq t_{n-1}(\alpha/2)) = \frac{\alpha}{2}$$

fennáll. Erre az értékre igaz, hogy

$$P(-t_{n-1}(\alpha/2) < t_{n-1} < t_{n-1}(\alpha/2)) = 1 - \alpha.$$

A kritikus tartomány tehát:

$$C_1 = \left\{(x_1, \ldots, x_n) : |t| \geq t_{n-1}\left(\frac{\alpha}{2}\right)\right\}$$

és az elfogadási tartomány:

$$C_0 = \left\{(x_1, \ldots, x_n) : |t| < t_{n-1}\left(\frac{\alpha}{2}\right)\right\}.$$

Egyoldali esetben az ellenhipotézis

$$H_1 : m > m_0 \text{ (vagy } m < m_0\text{) alakú.}$$

Ekkor azt a $t_{n-1}(\alpha)$ értéket kell kikeresnünk a táblázatból, amely a következő összefüggést teljesíti:

$$\begin{aligned} \mathbb{P}(t_{n-1} \geq t_{n-1}(\alpha)) &= \alpha \\ (\text{vagy } \mathbb{P}(t_{n-1} \leq -t_{n-1}(\alpha)) &= \alpha). \end{aligned}$$

Az egyoldali ellenhipotézishez tartozó kritikus tartomány:

$$\begin{aligned} C_1 &= \{(x_1, \ldots, x_n) : t \geq t_{n-1}(\alpha)\} \\ (\text{vagy } C_1 &= \{(x_1, \ldots, x_n) : t \leq -t_{n-1}(\alpha)\}). \end{aligned}$$

A kétmintás t-próba azt vizsgálja, hogy két külön mintában egy-egy valószínűségi változó átlagai egymástól szignifikánsan különböznek-e.
H_0: Az X és Y valószínűségi változók várható értékei megegyeznek (E(X) = E(Y))
H_1: Az X és Y valószínűségi változók várható értékei nem egyeznek meg (E(X)! = E(Y))
Próbastatisztika:

$$t_{n+m-2} = \sqrt{\frac{nm(n+m-2)}{n+m}} \frac{\overline{X} - \overline{Y}}{\sqrt{\sum (X_i - \overline{X})^2 + \sum (Y_i - \overline{Y})^2}}$$

Kritikus tartomány: mint az egymintás esetben

Feltételei: ha ismeretlenek, de azonosak a szórások

- a val.változók függetlenek és normális eloszlásúak

17. Feltételes valószínűség

4.1. definíció: Ha A és B egy kísérlettel kapcsolatos két tetszőleges esemény és $P(B) > 0$, akkor az A esemény B-re vonatkozó feltételes valószínűségét a

$$P(A|B) = \frac{P(A \cdot B)}{P(B)}$$

kifejezéssel definiáljuk.

- Az *A* esemény valószínűségét keressük.
- Tudjuk, hogy *B* esemény bekövetkezett.
- A relatív gyakoriságokkal: csak azokat a kísérleteket nézzük, amelyekben *B* bekövetkezett. Ezen részsorozatban az *A* relatív gyakorisága:

$$r_{A \cap B} / r_B$$

- Megfelelője a valószínűségekre:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

az A esemény B-re vonatkozó feltételes valószínűsége (feltétel: $P(B)>0$).

18. Várható érték fogalma diszkrét valószínűségi változóban

7.1. definíció: Legyen X diszkrét valószínűségi változó, mely az $x_1, x_2, \dots$ értékeket veheti fel. Legyen továbbá $p_i = P(X = x_i)$. Ekkor X várható értékén az $E(X) = \sum_i x_i \cdot p_i$ összeget értjük, amennyiben a $\sum_i |x_i| \cdot p_i$ összeg véges. Egyéb esetben azt mondjuk, hogy a várható érték nem létezik.

19. Példa az exponenciális, egyenletes és normális eloszlás alkalmazására

10.2. definíció: Az X valószínűségi változó $\lambda > 0$ paraméterű exponenciális eloszlású, ha sűrűségfügvénye

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x}, & \text{ha } x > 0; \\ 0 & \text{különben.} \end{cases}$$

A sűrűségfüggvény alapján meghatározható az eloszlásfüggvénye:

$$F(x) = \begin{cases} 1 - e^{-\lambda \cdot x}, & \text{ha } x > 0; \\ 0, & \text{ha } x \leq 0. \end{cases}$$

10.2. tétel: Ha X egy λ paraméterű exponenciális eloszlású valószínűségi változó, akkor X várható értéke és szórása is a λ paraméter reciprokával egyezik meg, azaz

$$E(X) = D(X) = \frac{1}{\lambda}.$$

Exponenciális eloszlással modellezhető például a várakozási-, illetve a sorbanállási idő, továbbá bizonyos típusú berendezések, alkatrészek élettartama is.

10.1. definíció: Az X valószínűségi változó egyenletes eloszlású az $(a; b)$ intervallumon, ha sűrűségfüggvénye

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{ha } a < x < b; \\ 0 & \text{különben.} \end{cases}$$

A sűrűségfüggvény alapján meghatározható az eloszlásfüggvénye

$$F(x) = \begin{cases} 0, & \text{ha } x \leq a; \\ \frac{x-a}{b-a}, & \text{ha } a < x \leq b; \\ 1, & \text{ha } x > b. \end{cases}$$

10.1. tétel: Ha X egyenletes eloszlású valószínűségi változó az $(a; b)$ intervallumon, akkor várható értéke és szórása:

$$E(X) = \frac{a+b}{2}, \qquad D(X) = \frac{b-a}{\sqrt{12}}.$$

Alkalmazható

10.3. definíció: Az X folytonos eloszlású valószínűségi változót m, σ ($\sigma > 0$) paraméterű normális eloszlásúnak nevezünk, ha a sűrűségfüggvénye

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

Eloszlásfüggvénye pedig

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} \, dt.$$

Várható értéke és szórása:

$$E(X) = m, \qquad D(X) = \sigma.$$

Alkalmazható mérési hibákra, méretingadozásokra és olyan élettartamvizsgálatokra, ahol a készülékek, alkatrészek rendszeres kopással mennek tönkre.

20. bootstrap módszer

Újramintavételezési eljárás, a becsléseink szórásának vizsgálatára, modellilleszkedés ellenőrzésére használatos. Előnye: rugalmas a minta/statisztika eloszlására vonatkozó feltételek változására.Adott $X_1, X_2,$…azonos F eloszlású, független val. változók (F ismeretlen): X_n:= { X_1, … , X_n } minta. Ebből az X_n-ből m elemű mintát veszünk visszatevéssel (általában m = n). X^*_m = { X^*_1, … , X^*_m } Ebből az új mintából becsülhetjük a számunkra érdekes statisztika eloszlását.

21. Leíró statisztika feladata
Nem a véletlen hatását vizsgálja, hanem a konkrét minta 1. megjelenítése, 2. jellemzőinek kiszámítása a feladata. Adatok elrendezhetők táblázatban (fontos: forrás feltüntetése), illetve ábrázolhatók grafikusan.

22. Egy-és kétmintás u-próba
Az **egymintás *u*-próba** (más néven egymintás **z-próba**) a statiszitkai hipotézisvizsgálatok közül a paraméteres próbák közé tartozik. A próba azt ellenőrzi, hogy egy adott statisztikai ismérv esetén a mintabeli átlag szignifikánsan eltér-e a populációs átlagtól. Más szavakkal, hogy egy valószínűségi változó átlaga szignifikánsan különbözik-e egy adott *m* értéktől.

a) **Az egymintás *u*-próba.** Tegyük fel, hogy X egy $N(m;\sigma)$ normális eloszlású valószínűségi változó, ismeretlen m és ismert σ értékkel. Ellenőrizni akarjuk, hogy az X valószínűségi változó várható értéke m egyenlő-e egy adott m_0 számmal. Egy n elemű minta $\overline{X}$ átlaga általában nem lesz pontosan m_0. Kérdés: a mintaátlag mekkora eltérése esetén feltételezhetjük, hogy a várható érték m_0? Ha a

$$H_0 : M(X) = m_0$$

nullhipotézis teljesül a

$$H : M(X) = m \neq m_0 \qquad (*)$$

ellenhipotézissel szemben, akkor az

$$u = \sqrt{n}\,\frac{\overline{X} - m_0}{\sigma} \qquad (1)$$

formulával képzett u valószínűségi változó (próbafüggvény) $N(0;1)$ standard normális eloszlású. Ekkor a II. rész 2. fejezet 2.2. pontjának (1) relációja alapján az u konfidencia-intervalluma $1-p$ valószínűséggel megadható, azaz

$$P\left(-u_p \le \sqrt{n}\,\frac{\overline{X} - m_0}{\sigma} \le u_p\right) = 1 - p = 2\Phi(u_p) - 1$$

és így a $\Phi(u_p) = 1 - \frac{p}{2}$ összefüggésből adott p-hez u_p értéke a 1. sz. táblázatból visszakereshető. A (*) ún. *kétoldali ellenhipotézis*, mely azt fejezi ki, hogy vagy $m < m_0$ vagy $m > m_0$.

Az (1) képlettel **számított** $u_{sz} = u$ **érték** nagy valószínűséggel (pl. $1 - p = 1 - 0{,}05 = 0{,}95$) az $\left[-u_p; u_p\right]$ elfogadási tartományba esik, és csak kis valószínűséggel (pl. 0,05) esik a kritikus tartományba.

Azt mondjuk, hogy $100(1-p)\%$ biztonsági szinten elfogadjuk a nullhipotézist, ha

$$\left|u_{sz}\right| = \left|\sqrt{n}\,\frac{\overline{X} - m_0}{\sigma}\right| \le u_p = u_t$$

eltérés véletlenszerű, statisztikailag nem igazolható, más szóval az **eltérés nem szignifikáns**.

Azt mondjuk, hogy a H_0 hipotézist $100(1-p)\%$ szinten elutasítjuk, vagyis a H ellentett hipotézist fogadjuk el, ha

$$\left|u_{sz}\right| > u_t,$$

azaz ha u_{sz} a kritikus tartományba esik. Ez utóbbi esetben statisztikailag igazoltnak fogadjuk el, hogy az alapsokaság várható értéke és a feltételezett m_0 érték között **szignifikáns különbség** van.

Mint az előző pontban említettük, mindkét döntésünk véletlen folytán téves lehet. Elsőfajú hibát követünk el, ha H_0 hipotézis igaz, de elvetjük az $M(X) = m_0$ feltevést, mert pl. az u_{sz} érték a kritikus tartományra esett. Másodfajú hibát követünk el, ha H_0 hipotézis nem igaz, de elfogadjuk az $M(X) = m_0$ feltevést pl. mert az u_{sz} érték az elfogadási tartományra esett.

b) **A kétmintás *u*-próba.** Legyen X és Y két normális eloszlású valószínűségi változó, melyekhez $X_1, X_2, \ldots, X_k$ és $Y_1, Y_2, \ldots, Y_m$ független minták tartoznak, valamint szórásuk, σ_x és σ_y, ismert.

Ha

$$H_0 : M(X) = M(Y)$$

nullhipotézis teljesül a

$$H : M(X) \neq M(Y)$$

ellenhipotézissel szemben, akkor az

$$u = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{k} + \frac{\sigma_y^2}{m}}} \qquad (2)$$

valószínűségi változó (próbastatisztika) $N(0;1)$ eloszlású, ahol $\overline{X}$, $\overline{Y}$ a két mintából számított mintaközép. Ekkor a kétmintás *u*-statisztika u konfidencia-intervalluma $1-p$ valószínűséggel megadható, azaz

$$P(-u_p \le u \le u_p) = 1 - p = 2\Phi(u_p) - 1.$$

Az u_p értékét adott p-hez az 1. sz. táblázatból $\Phi(u_p) = 1 - \frac{p}{2}$ összefüggés alapján kiszámíthatjuk. Ha a (2) formulával kiszámított *u*-statisztika nem esik a $\left[-u_p; u_p\right]$ intervallumba, akkor $1-p$ szignifikanciaszinten a nullhipotézist elvetjük, mivel ekkor a kiszámított u értéke a $\left[-u_p; u_p\right]$ intervallumon kívüli kritikus tartományba esik.

23. Hogyan adhatjuk meg megszámlálható valószínűségi mezőn a valószínűséget?

$\Omega=\{\omega_1, \omega_2, \ldots, \omega_n, \ldots\}$, $\mathcal{A}= \mathcal{P}(\Omega)$.
Jelölés: $p_i=P(\omega_i)$, *valószínűségeloszlás*:
$p_i \geq 0$, az összegük 1.
A σ-additivitás miatt tetszőleges A eseményre megy a véges esetre látott számítás:

$$P(A) = P(\cup_{i:\omega_i \in A} \omega_i) = \sum_{i:\omega_i \in A} p_i$$

A megszámlálható számosságú valószínűségi mezők (azaz az olyan kísérletek, melyeknek megszámlálható sok kimenetele van) teljesen leírhatók az ún. diszkrét valószínűségeloszlások segítségével.

1.6. Definíció. *A $p_1, p_2, \ldots$ számsorozatot* ***diszkrét valószínűségeloszlásnak*** *(röviden eloszlásnak) nevezzük, ha*

$$p_i \geq 0, \quad i = 1, 2, \ldots, \quad \sum_{i=1}^{\infty} p_i = 1.$$

Ha $p_1, p_2, \ldots$ egy diszkrét valószínűségeloszlás, akkor legyen $\Omega = \{w_1, w_2, \ldots\}$ egy tetszőleges megszámlálható halmaz, $\mathcal{F} = 2^{\Omega}$. A

$$P(A) = \sum_{w_i \in A} p_i, \quad A \in \mathcal{F}$$

képlet nyilván valószínűséget definiál, melyre $P(w_i) = p_i$, $i = 1, 2, \ldots$.

24. Markov-egyenlőtlenség

2.26. Tétel. *(Markov-egyenlőtlenség) Legyen η nemnegatív valószínűségi változó, $\delta > 0$ rögzített szám. Ekkor*

$$\boxed{P(\eta \geq \delta) \leq \mathbb{E}\eta/\delta}$$

Bizonyítás. A következő egyenlőtlenségek érvényesek:

$$\mathbb{E}\eta = \sum_i y_i P(\eta = y_i) \geq \sum_{\{i: y_i \geq \delta\}} y_i P(\eta = y_i) \geq$$

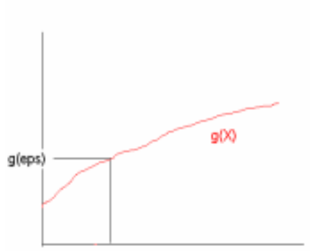
$$\geq \delta \sum_{\{i: y_i \geq \delta\}} P(\eta \geq y_i) = \delta P(\eta \geq \delta).$$

- Legyen $X \geq 0$ valószínűségi változó, g: $\boldsymbol{R}_+ \to \boldsymbol{R}_+$ monoton növő. Ekkor

$$P(X \geq \varepsilon) \leq E(g(X))/g(\varepsilon).$$

- Bizonyítás.

$E(g(X)) \geq g(\varepsilon)P(X \geq \varepsilon)$
mert $X \geq \varepsilon$ eseményen
$g(X)) \geq g(\varepsilon)$



25. Visszatevés nélküli mintavétel modellje és a különböző selejtszámok valószínűségei

A **visszatevés nélküli mintavétel** során vagy egyszerre emeljük ki a minta elemeit a sokaságból, vagy egyenként, de ekkor nem tesszük vissza a már kiválasztott elemeket.
Tegyük fel, hogy az N elemű sokaságban s számú elem rendelkezik egy adott T tulajdonsággal. Ezután válasszunk ki véletlenszerűen n darab elemet visszatevés nélkül! ($n \leq s$, és $n \leq N - s$.)
Legyen az X valószínűségi változó értéke az n-elemű mintában kiválasztott T tulajdonságú elemek száma. Bizonyítható, hogy ekkor az X valószínűségi változó hipergeometrikus eloszlású.

- N termék, melyből M selejtes
- n elemű minta visszatevés nélkül
- A: pontosan k selejtes van a mintában

(k=0,…,n)

$$P(A) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

Mintavétel

26. Sűrűségfüggvény becslésére tanult eljárás
27. F-próba és a legfontosabb alkalmazásai

25.2.1. F-próba

Az X és az Y **normális eloszlású**, $D(X) = \sigma_1$ és $D(Y) = \sigma_2$ **ismeretlen szórású** valószínűségi változók **szórásainak egyenlőségére** vonatkozó

$$H_0\colon \sigma_1 = \sigma_2$$

hipotézis helyességét akarjuk ellenőrizni.
A próbastatisztika formula:

$$F_p = \frac{\hat{s}_1^2}{\hat{s}_2^2}.$$

A próbastatisztika $(n_1 - 1, n_2 - 1)$ szabadsági fokú F-eloszlású, ha H_0 fennáll.
n_1 az X-re vonatkozó, n_2 az Y-ra vonatkozó független minták elemszáma.

Megjegyzés: Az, hogy az $\frac{\hat{s}_1^2}{\hat{s}_2^2}$ hányados értéke 1-nél nagyobb vagy kisebb, attól függ, milyen sorrendben dolgozzuk fel a mintákat. Az F-eloszlás táblázatának használatát megkönnyítendő a számlálóba mindig annak a mintának a korrigált tapasztalati szórásnégyzete kerüljön, amelynél ez az érték nagyobb. Ekkor a hányados értéke nem lehet 1-nél kisebb. Technikailag ez azt jelenti, hogy a próbastatisztikát úgy választjuk, hogy értéke legalább 1 legyen:

$$F_p = \max\left(\frac{\hat{s}_1^2}{\hat{s}_2^2}, \frac{\hat{s}_2^2}{\hat{s}_1^2}\right) \geq 1.$$

Ennek az a következménye, hogy F-próba esetén az elfogadási tartomány mindig az egyoldali próbának megfelelő $(0, F_{kr})$ lesz.

28. Véletlen szám generálása Neumann módszerével

Legyen f(x) tetszőleges sűrűségfüggvény,g(x) pedig olyan sűrűségfüggvény, amelyre f(x) < Mg(x), valamely M>1 esetén és g(x)-ből könnyen tudunk mintát venni (tipikus példa az egyenletes eloszlás).

Algoritmus:

1. Vegyünk mintát: *u=>U(0,1)* -ből, *x=>g(x)* -ből
2. Ha *u<f(x)/Mg(x),* akkor x-et elfogadjuk
3. Különben elutasítjuk, és 1-be lépünk.

29. χ-négyzet próba és legfontosabb alkalmazásai

A leggyakrabban használt nem-paraméteres próba a χ 2 -próba. Fontos megjegyezni, hogy míg a korábban ismertetett próbák kis és közepes minták esetében is jól használhatóak, a χ 2 -próba csak nagy elemszámú minták esetében ad megbízható eredményt. Tekintsük át a χ 2 -próba három legfontosabb alkalmazási területét!

a) Illeszkedésvizsgálat

Akkor alkalmazzuk, ha azt akarjuk ellenőrizni, hogy a minta adott eloszlású populációból származik-e. A gyakorlatban ez azt jelenti, hogy azt vizsgáljuk, hogy a mintából származó gyakoriságok mennyire illeszkednek az eloszlásból származó becsült gyakoriságokra. Adott egy n elemű minta, a mintaelemeket osszuk r (páronként diszjunkt) csoportba, azaz tekintsük az $A_1, A_2, \ldots, A_r$ teljes eseményrendszert. Az A_i esemény valószínűségét p_i, gyakoriságát (hány mintaelem esik az i-edik csoportba) pedig μ_i jelölje ($i = 1, 2, \ldots, r$ esetén).
Az eloszlásra vonatkozó $H_0 : P(A_i) = p_i$ ($i = 1, 2, \ldots, r$) nullhipotézist akarjuk ellenőrizni, azaz azt, hogy a megfigyelt valószínűségi változó adott eloszlású.
A próbastatisztika:

$$\chi_p^2 = \sum_{i=1}^{r} \frac{(\mu_i - n \cdot p_i)^2}{n \cdot p_i}$$

Eloszlása: $r - 1$ szabadsági fokú χ^2-eloszlás, ha H_0 fennáll.
(Pontosabban $n \to \infty$ esetén $r - 1$ szabadsági fokú χ^2-eloszlással közelíthető. A gyakorlatban azonban a fenti próbastatisztikát $r - 1$ szabadsági fokú χ^2-eloszlásúnak tekintjük.)
A próba mindig egyoldali, jobb oldali kritikus tartománnyal. (Vegyük észre, ha a feltételezett eloszlás különbözik

b) Homogenitásvizsgálat

Akkor alkalmazzuk, ha azt akarjuk eldönteni, hogy két független minta azonos eloszlásból származik-e. (Anélkül, hogy a konkrét eloszlást meghatároznánk.)
Jelölje X és Y a megfigyelt valószínűségi változókat. Az X-re vonatkozó minta elemszáma legyen m, az Y-ra vonatkozó minta elemszáma n. Soroljuk valamilyen szempont szerint r csoportba a mintaelemeket. Jelölje μ_i az X, ν_i az Y megfigyeléséből származó minta i-edik csoportba eső mintaelemeinek számát (azaz az A_i, illetve B_i események gyakoriságát). $\left(\sum_{i=1}^{r} \mu_i = m, \sum_{i=1}^{r} \nu_i = n\right)$.
A nullhipotézisünk az, hogy a két minta azonos eloszlásból származik.
A próbastatisztika:

$$\chi_p^2 = \frac{1}{m \cdot n} \sum_{i=1}^{r} \frac{(n \cdot \mu_i - m \cdot \nu_i)^2}{\mu_i + \nu_i}$$

Eloszlása: H_0 fennállása esetén $r - 1$ szabadsági fokú χ^2-eloszlás.
A próba most is egyoldali, jobb oldali kritikus tartománnyal. (Vegyük észre, ha a két minta eloszlása különböző, akkor a próbastatisztika értéke a formulából adódóan nagy. Minél nagyobb ez a számított érték, annál nagyobb ez az eltérés.)

c) Függetlenségvizsgálat

Azt akarjuk eldönteni, hogy két valószínűségi változó, X és Y függetlenek-e. Az (X,Y) valószínűségi változó megfigyelésére n elemű mintát veszünk. Az X valószínűségi változó értékeit r, Y értékeit s csoportba soroljuk, azaz létrehozzuk X-re vonatkozóan az $A_1, A_2, \ldots, A_r$, Y-ra vonatkozóan a $B_1, B_2, \ldots, B_s$ teljes eseményrendszereket. Jelölje μ_{i*} az A_i ($i = 1, 2, \ldots, r$), ν_{*j} a B_j ($j = 1, 2, \ldots, s$) bekövetkezésének gyakoriságát, μ_{ij} pedig az (A_i, B_j) együttes bekövetkezésének gyakoriságát. Természetesen ekkor teljesül, $\sum_{i=1}^{r} \sum_{j=1}^{s} \mu_{ij} = 1$,

valamint, hogy $\sum_{i=1}^{r} \mu_{ij} = \nu_{*j}$, és $\sum_{j=1}^{s} \nu_{ij} = \mu_{i*}$. (Ez utóbbiakhoz idézzük fel a valószínűségi vektorváltozók peremeloszlásait diszkrét esetben!)
A feltevésünk, hogy a két valószínűségi változó független, vagyis $(A_1, A_2, \ldots, A_r)$ és $(B_1, B_2, \ldots, B_s)$ teljes eseményrendszerek függetlenek. A nullhipotézis formálisan az, hogy $P(A_i \cdot B_j) = P(A_i) \cdot P(B_j)$ ($i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, s$ esetén).
A próbastatisztika:

$$\chi_p^2 = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n \cdot \mu_{ij} - \mu_{i*} \cdot \nu_{*j})^2}{\mu_{i*} \cdot \nu_{*j}}$$

Eloszlása: $(r - 1) \cdot (s - 1)$ szabadsági fokú χ^2-eloszlás, amennyiben H_0 fennáll.
A próba egyoldali, jobb oldali kritikus tartománnyal, hasonlóan az illeszkedés,- illetve a homogenitásvizsgálathoz.

30. Becsléses illeszkedés vizsgálat és hogyan alkalmazzuk rá a χ2-négyzet próbát
31. Szekvenciális próba(egyszerű hipotézisek esetére)

A valószínűséghányados n elemű mintából

$$V_n = \frac{\prod_{i=1}^{n} f_1(X_i)}{\prod_{i=1}^{n} f_0(X_i)}$$

Addig veszünk mintaelemeket, amíg $V_n \geq B$ vagy $V_n \leq A$ nem teljesül

Tehát az algoritmus:

- $V_n \geq B$: elutasítjuk H_0-t.
- $V_n \leq A$: elfogadjuk H_0-t.
- $B > V_n > A$: új mintaelemet veszünk.

1 valószínűséggel véges az N, ahol

$$N = \min\{n : V_n \leq A \text{ vagy } V_n \geq B\}$$

32. Korrelációs együttható és legfontosabb tulajdonságai

- A kovariancia skálafüggő: cov(*a*X,*b*Y)=*ab*cov(X,Y)
- A változók közötti lineáris kapcsolat erősségét mérő mennyiség a *korrelációs együttható*:

$$R(X,Y) = \frac{\operatorname{cov}(X,Y)}{D(X)D(Y)}$$

- Tulajdonságai:
  - R(X,Y)=0, ha X és Y függetlenek (ez sem fordítható meg)
  - Ez alapján definíció szerint legyen R(X,Y)=0, ha X vagy Y elfajult eloszlású.
  - R(X,aX+b)=1, ha a>0, mert cov(X,aX+b)=aD^2(X).

Két valószínűségi változó közötti kapcsolat szorosságát kifejezhetjük a korrelációs együtthatóval. A korrelációs együttható lényegében két vektor által bezárt szög koszinuszát jelenti.

2.31. Definíció. *Legyen* $0 < \mathbb{D}^2\xi, \mathbb{D}^2\eta < \infty$. *A* ξ *és* η *korrelációs együtthatóján a*

$$\boxed{\operatorname{corr}(\xi,\eta) = \frac{\operatorname{cov}(\xi,\eta)}{\mathbb{D}\xi\mathbb{D}\eta}}$$

mennyiséget értjük. Ha $\operatorname{corr}(\xi,\eta) = 0$, *akkor* ξ*-t és* η*-t* ***korrelálatlanoknak*** *nevezzük. Ha* ξ *és* η *függetlenek, akkor korrelálatlanok, de ez fordítva nem igaz.*

33. Mikor nevezünk egy torzítatlan becslést hatásosabbnak egy másiknál?

A becslés hatásfoka Bárki megkérdezheti tehát, hogy két torzítatlan becslés közül melyik a jobb becslés. A válasz elég nyilvánvaló, ha egy kételemű minta alapján kell a ξ valószínűségi változó $E(\xi)$ várható értékének az alábbi két torzítatlan becslése közül választanunk:

1) $\tilde{\vartheta} = \frac{\xi_1 + \xi_2}{2}$,

2) $\hat{\vartheta} = -7\xi_1 + 8\xi_2$.

A $\hat{\vartheta}$ szórása sokkal nagyobb a $\tilde{\vartheta}$ szórásánál, így feltehetően mindenki az 1) számú becslést választaná.

Definíció: Legyen adott a ϑ paraméter két $\hat{\vartheta} = \hat{\vartheta}(\xi_1, \xi_2, \ldots, \xi_n)$ és $\tilde{\vartheta} = \tilde{\vartheta}(\xi_1, \xi_2, \ldots, \xi_n)$ torzítatlan becslése. A $\tilde{\vartheta}$ becslést **hatásosabb**nak nevezzük a $\hat{\vartheta}$ becslésnél, ha

$$D^2\left(\tilde{\vartheta}\right) \leq D^2\left(\hat{\vartheta}\right)$$

a ϑ paraméter minden lehetséges értékére.

34. Trend és a periódus fogalma idősorokra

a)Trend

Az idősorban hosszabb időszakon át tartósan érvényesülő tendencia

b) Periódius

Az i állapot periódusa d, ha d=lnko{k: $p_i^{(k)}$>0}.

35. Nagy számok gyenge törvénye

36. Szita (Poincaré) formula

Képlet az általános esetre:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} (-1)^{i+1} S_i^{(n)}$$

ahol

$$S_i^{(n)} = \sum_{1 \le j_1 < j_2 < \ldots < j_i \le n} P(A_{j_1} \cap A_{j_2} \cap \ldots \cap A_{j_i})$$

az i tényezős metszetek valószínűségeinek összege.

A valószínűségszámításban Poincaré formula néven is ismert állítás a következő:

3. Tétel. (*Nagy számok gyenge törvénye*):
Legyenek az $X_1, X_2, \ldots, X_n, \ldots$ valószínűségi változók függetlenek; eloszlásuk, m-mel jelölt várható értékük, s^2-tel jelölt szórásnégyzetük azonos, akkor az

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

számtani közép sztochasztikusan konvergál az m várható értékhez, ha n minden határon túl nő, azaz

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n - m\right| \ge \varepsilon\right) = 0 \quad \text{vagy} \quad \lim_{n\to\infty} P\left(\left|\overline{X}_n - m\right| < \varepsilon\right) = 1. \quad (1)$$

Ui. $M(\overline{X}_n) = \dfrac{M(X_1) + M(X_2) + \ldots + M(X_n)}{n} = \dfrac{n \cdot m}{n} = m,$

s mivel $X_1, X_2, \ldots, X_n$ független valószínűségi változók, ezért

$$D^2(X_1 + X_2 + \ldots + X_n) =$$
$$= D^2(X_1) + D^2(X_2) + \ldots + D^2(X_n) = n \cdot s^2.$$

Tekintettel arra, hogy

$$D^2(\overline{X}_n) = D^2\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) =$$
$$= \frac{1}{n^2} D^2(X_1 + X_2 + \ldots + X_n) =$$
$$= \frac{1}{n^2}(n \cdot s^2) = \frac{s^2}{n}$$

így a *Csebisev*-egyenlőtlenség (5) alakját használva:

$$P\left(\left|\overline{X}_n - m\right| \ge \varepsilon\right) \le \frac{s^2}{n\varepsilon^2},$$

$$P(A_1 \cup \cdots \cup A_n)$$

ahol $S_i^{(n)}$ az összes i-tényezős metszet valós

$$S_i^{(n)} = \sum_{1 \le k_1 < k_2 < \cdots < k_i \le n} I$$

Sok esetben pedig az az egyszerű átfogalmazás még praktikusabb, ahol unió helyett metszet szerepel:

ahol legyen $S_0^{(n)} := 1$. A két állítás ekvivalenciája abból adódik, hogy a komplementerek metszete éppen azt jelenti, hogy egyik esemény sem következik be – ez pedig éppen az unió komplementere. A jobb oldal pedig éppen 1- a komplementer esemény valószínűsége.

37. Kovariancia és tulajdonságai

- Definíció. Az X és Y kovarianciája: cov(X,Y):=E[(X-E(X))(Y-E(Y))]
- Kiszámítása: cov(X,Y)= E[XY-XE(Y)-YE(X)+E(X)E(Y)]=E(XY)-E(X)E(Y)
- A korábban látott, független val.változókra vonatkozó E(XY)=E(X)E(Y) egyenlőség értelmében cov(X,Y)=0, ha X és Y függetlenek.
- Megj.: Abból, hogy cov(X,Y)=0 nem következik, hogy függetlenek: legyen X szimmetrikus a 0=ra (pl. P(X=1)=P(X=-1)=P(X=0)=1/3) és $Y=X^2$. Ekkor cov(X,Y)=$E(X^3)$-E(X)$E(X^2)$=0-0, hiszen $E(X^3)$=E(X)=0.
- A kovariancia szimmetrikus: cov(X,Y)= cov(Y,X)
- cov(X,X)= $D^2(X)$

Definíció. Legyen X és Y diszkrét, véges valószínűségi változók együttes eloszlása w, a várható értékeket jelölje m_x és m_y, akkor az X és Y változók Cov(X,Y)-nal jelölt **kovarianciáján** a

$$\mathrm{Cov}(X,Y)=\sum_{i,j}(x_i-m_x)(y_j-m_y)w(x_i,y_j)=M\left((X-m_x)(Y-m_y)\right) \quad (4)$$

számot értjük.

Minthogy az X és Y diszkrét, véges valószínűségi változók, ezért a

$$\sum_{i,j} y_j w(x_i,y_j)=\sum_j y_j u(y_j)=m_y;$$

$$\sum_{i,j} x_i w(x_i, y_j) = \sum_i x_i v(x_i) = m_x$$

és a

$$\sum_{i,j} w(x_i, y_j) = 1$$

összefüggések felhasználásával a (4) képletet számításra alkalmasabb alakra hozhatjuk:

$$\mathrm{Cov}(X,Y) = \sum_{i,j} (x_i - m_x)(y_j - m_y) w(x_i, y_j) =$$

$$= \sum_{i,j} x_i y_j w(x_i, y_j) - m_x m_y = M(XY) - m_x m_y. \qquad (5)$$

Ha X és Y független valószínűségi változók, akkor

$$w(x_i, y_j) = v(x_i) \cdot u(y_j),$$

és így

$$M(XY) = \sum_{i,j} x_i y_j w(x_i, y_j) = \sum_{i,j} x_i y_j v(x_i) \cdot u(y_j) =$$

$$= \sum_i x_i v(x_i) \cdot \sum_j y_j u(y_j) = M(X) M(Y).$$

Az $M(XY) = M(X)M(Y)$ -ból (5)-re tekintettel következik, hogy $\mathrm{Cov}(X,Y) = 0$, de ez fordítva nem áll fenn, azaz $\mathrm{Cov}(X,Y) = 0$-ból nem következik X és Y változók függetlensége. Mivel a kovariancia nagyságát jelentősen befolyásolja a vizsgálatba bevont (x_i, y_j) adatpárok mértékegysége, ezért ennek kiküszöbölésére az

$$\frac{X - m_x}{D(X)} \quad \text{és az} \quad \frac{Y - m_y}{D(Y)}$$

standardizált változók kovarianciáját képezzük, mely már dimenziómentes:

$$\mathrm{Cov}\left(\frac{X - m_x}{D(X)} \cdot \frac{Y - m_y}{D(Y)} \right) = \frac{\mathrm{Cov}(X,Y)}{D(X) D(Y)}.$$

Az összefüggés szorossága így függetleníthető az (x_i, y_j) adatpárok, a változók mértékegységétől. Ezek figyelembevételével kimondható a következő

38. Abszolút folytonos valószínűségi változók függetlenségének ekvivalens jellemzői

1.32. Tétel. *Legyen $(\xi_1, \ldots, \xi_n)$ abszolút folytonos valószínűségi vektorváltozó. A $\xi_1, \ldots, \xi_n$ valószínűségi változók pontosan akkor függetlenek, ha*

$$f(x_1, \ldots, x_n) = \prod_{k=1}^{n} f_k(x_k)$$

teljesül minden $x_1, \ldots, x_n \in \mathbb{R}$ esetén, ahol f_k a ξ_k sűrűségfüggvénye, továbbá f a $(\xi_1, \ldots, \xi_n)$ sűrűségfüggvénye.

39. Szórásnégyzet és legfontosabb tulajdonságai

Nem mindegy, hogy mekkora a vizsgált véletlen mennyiség ingadozása.

Jobb, ha a buszok pontosan 10 percenként jönnek, mintha időnként 3 jön egymás után, és aztán 30 percet kell várni.

Az ingadozás számszerűsítése: a várható értéktől vett átlagos négyzetes eltérés, elnevezése: szórásnégyzet. Formálisan:

$$D^2(X):=E[(X-E(X))^2].$$

Kiszámítása: $D^2(X)= E[X^2-2XE(X)+E^2(X)]= =E(X^2)-2E(X)E(X)+E^2(X)$

a várható érték linearitása miatt. Azaz

$$D^2(X)=E(X^2)-E^2(X).$$

- $D^2(X)\geq 0$, mert nemnegatív valószínűségi változó várható értéke.
- $D^2(aX+b)=a^2D^2(X)$, mert $D^2(aX+b)= =E[(aX+b-E(aX+b))^2]= E[(aX+b-aE(X)-b)^2]= =E[(aX-aE(X))^2]=a^2 E[(X-E(X))^2]$.
- Abból, hogy E(X) véges, még nem következik $D^2(X)$ végessége, hiszen ha $P(X=k)=c/k^3$ (egyértelműen megadható olyan *c*, amire ez eloszlás lesz) akkor E(X) véges, de $E(X^2)=c(1+1/2+...+1/k+...)$, ami végtelen.

Definíció. Az X valószínűségi változó **szórásnégyzetének (varianciájának)** az $(X-m)^2$ valószínűségi változó várható értékét nevezzük, azaz a

$$D^2(X)=\sum_{i=1}^{n}(x_i-m)^2 v(x_i)=M((X-m)^2) \qquad (1)$$

formulával adott számot, ahol m az X várható értéke ($m=M(X)$).

A szórásnégyzet szokásos jelölése még s^2, σ^2 és Var(X).

A szórásnégyzet jellemzi tehát az X-re vonatkozó egyes mérési adatok eltérését átlaguktól. Más szóval az adatok négyzetes középhibája $D^2(X)$ körül ingadozik.

Definíció. Ha az X folytonos eloszlású valószínűségi változónak $f(x)$ a sűrűségfüggvénye, akkor **szórásnégyzetét** és **szórását** a

$$D^2(X)=M((X-m)^2)=\int_{-\infty}^{\infty}(x-m)^2 f(x)dx \qquad (2)$$

$$D(X)=\sqrt{\int_{-\infty}^{\infty}(x-m)^2 f(x)dx} \qquad (3)$$

képletekkel definiáljuk, ahol $m=M(X)$, feltéve hogy a jobb oldali integrál létezik.

A (2) formula alapján azt is mondhatjuk, hogy a szórásnégyzet az $(X-M(X))^2$ valószínűségi változó várható értéke, ha létezik.

A diszkrét esethez hasonlóan az $m=M(X)$ és $M(X^2)$ létezése szükséges és elegendő feltétel a $D^2(X)$ létezéséhez, és így a *szórásnégyzetet* és a *szórást* a következő képletekkel is számíthatjuk:

$$D^2(X)=M(X^2)-m^2=\int_R x^2 f(x)dx-m^2 \qquad (4)$$

$$D(X)=\sqrt{M(X^2)-m^2}=\sqrt{\int_R x^2 f(x)dx-m^2}. \qquad (5)$$

40. Lineáris regresszió feladata és megoldása

A túlhatározott egyenletrendszerek egy speciális esetének tekinthető a lineáris regresszió. Adott n darab pont, koordinátáik: $P_i=(x_i;y_i)$ $(i=1,...,n)$. Ezek a pontok nem illeszkednek ez egyenesre. Az a feladat. hogyan határozzuk meg a legkisebb négyzetek módszerének a felhasználásával a pontokat legjobban megközelítő egyenes, az ún. regressziós egyenes egyenletét. Az egyenes egyenlete $y=\alpha x+\beta$, ahol az α, β valós paraméterek.

41. Adjon módszert idősorok simítására

Nem stacionárius idősor esetén a megfigyelhető trendet leválasztjuk, majd az egyes periódusokra kiszámoljuk a reziduálisok átlagát. Ezt kivonva az eredeti idősorból, periódusmentes adatokat kapunk.

42. Centrális határeloszlás tétele

7.4 Tétel (Centrális határeloszlástétel független, azonos eloszlású változókra)
Legyenek
$\xi_1, \xi_2, \ldots$ *független, azonos eloszlásúak,* $m := E\xi_1$ *és* $0 < \sigma^2 = D^2\xi_i < \infty$*. Ekkor minden* $x \in \mathbb{R}$*-re*

$$P\left(\frac{\sum_{i=1}^{n} \xi_i - n \cdot m}{\sigma \cdot \sqrt{n}} < x\right) \to \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}\, dt$$

Legyenek X_1, X_2,..., X_n,... független, azonos eloszlású valószínűségi változók. Tegyük fel, hogy $\sigma^2=D^2(X)$ véges ($m:=E(X_i)$). Tekintsük a standardizált összegüket:

$$Z_n := \frac{X_1 + \ldots + X_n - nm}{\sqrt{n}\sigma}$$

Ekkor Z_n gyengén konvergál a standard normális eloszláshoz, azaz

$$P\left(\frac{X_1 + \ldots + X_n - nm}{\sqrt{n}\sigma} < z\right) \to \Phi(z)$$

ahol Φ a standard normális eloszlás eloszlásfüggvénye.

43. Lokális centrális határeloszlás tétele

Legyenek X_1, X_2,..., X_n,... független, azonos eloszlású valószínűségi változók. Tegyük fel, hogy $\sigma^2=D^2(X)$ véges ($m:=E(X_i)$) valamint, hogy abszolút folytonosak, szakaszonként folytonos sűrűségfüggvénnyel. Tekintsük a standardizált összegüket (Z_n) és tegyük fel, hogy legalább egy n-re Z_n sűrűségfüggvénye korlátos. Ekkor a Z_n változó f_n sűrűségfüggvénye (mely a tagok abszolút folytonossága miatt létezik) konvergál a standard normális eloszlás sűrűségfüggvényéhez, azaz

$$f_n(t) \to \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

44. Teljes eseményrendszer fogalma

Definíció. Legyen $A_1, A_2, \ldots, A_n$, $A_k \neq \varnothing$ $(k = 1, 2, \ldots, n)$ az események olyan halmaza, amelyre $A_i \cdot A_j = \varnothing$ $(i, j = 1, 2, \ldots, n;\ i \neq j)$ és $A_1 + A_2 + \ldots + A_n = I$ teljesül, azaz összegük (uniójuk) a biztos esemény és páronkénti szorzatuk (metszetük) a lehetetlen esemény. Ekkor a

$$Q = \{A_1, A_2, \ldots, A_n\}$$

eseményhalmazt **teljes eseményrendszernek** nevezzük.

A teljes eseményrendszer eseményei közül – a definíció értelmében, az eseményekre vonatkozó kísérlet során – mindig egy esemény következik be, és nem több.

Definíció. Események A_1, A_2 , ..., sorozata *teljes eseményrendszer*, ha egymást páronként kizárják és egyesítésük Ω.

Tulajdonság: $P(A_1)+P(A_2)+...=1$

Legtöbbször véges sok elemből álló teljes eseményrendszereket vizsgálunk.

45. Definiálja folytonos eloszlásra a várható érték fogalmát!

Definíció. Ha az X folytonos eloszlású valószínűségi változónak $f(x)$ a sűrűségfüggvénye, akkor **várható értéke**:

$$M(X)=\int_{-\infty}^{\infty} xf(x)dx \qquad (1)$$

feltéve, hogy az improprius integrál abszolút konvergens, azaz

$$\int_{-\infty}^{\infty} |x| f(x)dx < +\infty.$$

Tehát az (1) formula csak akkor áll fenn, ha a jobb oldali integrál létezik és véges.

46. Bernstein tétele

Tétel (Bernstein). Legyen $(X_1, X_2, ..., X_n)$ olyan, hogy $D^2(X_1)+D^2(X_2)...+D^2(X_n)<kn$, valamint tegyük fel, hogy van olyan $h:\boldsymbol{N}\to\boldsymbol{R}_+$ függvény, melyre $|R(X_i,X_j)|\le h(|i-j|)$ és $\frac{1}{n}\sum_{i=1}^{n} h(i)\to 0$ ekkor $(X_1+X_2+...+X_n)/n - (m_1+m_2+...+m_n)/n \to 0$ sztochasztikusan.

Cantor –Schröder-Bernstein-tétel: Tetszőleges A és B halmazok esetén, ha létezik φ:A→B bijektív leképezés is, tehát az A és a B halmaz ekvivalens egymással. Egy halmazon értelmezett összes karakterisztikus függvény halmazának vonatkozó állítás: BA={φ:A→B} az összes A→B leképezés. Ha A={a 1 ,...,a n } és B={b 1 ,...,b m }, akkor az összes A→B leképezés: mn db. {0,1}A={φ:A→{0,1}} az összes A→{0,1} leképezés.

47. Maximum-likelihood módszer

$$L(\theta;\underline{x})=f_\theta(\underline{x})=\prod_{i=1}^{n} f_\theta(x_i)$$

(a likelihood függvény) maximumhelye lesz a θ paraméter maximum likelihood becslése.

Ha a függvény deriválható, a loglikelihood függvény

$$l(\theta;\underline{x})=\ln f_\theta(\underline{x})=\sum_{i=1}^{n}\ln f_\theta(x_i)$$

maximumhelye deriválással

$$\frac{\partial}{\partial\theta}l(\theta;\underline{x})=\frac{\partial}{\partial\theta}\ln f_\theta(\underline{x})=\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f_\theta(x_i)=0$$

megoldásaként megtalálható

48. Abszolút folytonos valószínűségi vektorváltozó fogalma

5.10 Definíció A ξ valószínűségi változó abszolút folytonos eloszlású, ha létezik olyan nemnegatív f függvény, hogy $F_\xi(x) = P(\xi < x) = \int_{-\infty}^{x} f(s)\, ds$. Az f függvényt a valószínűségi változó sűrűségfüggvényének nevezzük.

X valószínűségi vektorváltozó abszolút folytonos, ha létezik olyan $f_x(x_1,x_2,\ldots,x_\alpha)$ függvény, amelyre $F_x(x_1,\ldots x_\alpha) = \int_{-\infty}^{x1} \ldots \int_{-\infty}^{x0} fx(t1, \ldots, td)$ $\alpha t_1,\ldots, \alpha t_d$. Ilyenkor fx (x)-et sűrűségfüggvénynek nevezzük.

49. (n,p) paraméterű binomiális eloszlás várható értékének levezetése

Az (*n*,*p*) paraméterű binomiális eloszlás várható értéke:

$$E(X) = \sum_{k=1}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} np \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} = np$$

50. Hipergeometriai eloszlás várható értékének levezetése

Visszatevés nélküli mintavételnél adott típusú mintaelemek száma.

$$E(X) = \sum_{k=1}^{n} k \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} = \sum_{k=1}^{n} n \frac{M}{N} \frac{\binom{M-1}{k-1}\binom{N-1-(M-1)}{n-1-(k-1)}}{\binom{N-1}{n-1}} = n\frac{M}{N}$$

51. Poisson eloszlás várható érték levezetése

$$E(X) = \sum_{k=1}^{\infty} k\lambda^k \frac{e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} \lambda^k \frac{e^{-\lambda}}{(k-1)!} = \lambda \sum_{k=1}^{\infty} \lambda^{k-1} \frac{e^{-\lambda}}{(k-1)!} = \lambda$$

52. Egyenletes eloszlás várható értékének levezetése

Az x1 , x2 ,…xn számokon egyenletes eloszlás (mindegyik valószínősége 1/n) várható értéke a számok számtani közepe.

Ha X egyenletes eloszlású az [a,b]-ben, akkor

$$E(X) = \int_a^b \frac{y}{b-a} dy = \left[\frac{y^2}{2(b-a)}\right]_a^b = \frac{a+b}{2}$$

53. Exponenciális eloszlás várható értékének levezetése

Ha X exponenciális eloszlású λ paraméterrel, akkor

$$E(X) = \int_0^\infty \lambda y e^{-\lambda y} dy = \left[-ye^{-\lambda y}\right]_0^\infty + \int_0^\infty e^{-\lambda y} dy = \frac{1}{\lambda}$$

54. Standard normális eloszlás várható értékének levezetése

Ha X standard normális eloszlású, akkor

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \left[-\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right]_{-\infty}^{\infty} = 0$$

55. Hogyan generálna normális eloszlású véletlen számot?

Box-Müller módszer

- Legyen U,V független, E[0;1] eloszlású. Ekkor

$$\sqrt{-2\ln U}\sin(2\pi V), \sqrt{-2\ln U}\cos(2\pi V)$$

két független standard normális eloszlású változó lesz.

56. Véletlen szám generálás inverz módszerrel

Tétel: Legyen X val. vált., F eloszlásfüggvénnyel, amely monoton növekedő és folytonos. Ekkor

i. F(X) egyenletes eloszlású [0,1] –en
ii. Ha U ~ U(0,1) akkor $F^{-1}(U)$ eloszlásfüggvénye F .

Pl.: Ha X ~ exp(λ) ⟶ F(x)=1-exp(- λx)
⟶ $F^{-1}(x)$= -ln(1-x)/ λ
⟶ -ln(1-U)/ λ ~ exp(λ)

Kiterjesztése: általánosított inverz: $F^{-1}(x)$= inf{x | F(x)=y}

Az Y_t diszkrét paraméterű sztochasztikus folyamatok k-ad rendű autoregresszív folyamatnak nevezzük, ha

$$Y_t = \alpha_1 * Y_{t-1} + \cdots + \alpha_k * Y_{t-k} + \varepsilon_t$$

Ahol:

- α_i konstansok
- Y_t fehér zaj (várható értéke 0, szórása σ_y)

57. Wilcoxon-próba

A Wilcoxon-féle előjeles rangpróba nem csak az előjeleket, hanem a különbségek közötti nagyságrendeket is figyelembe veszi, így nagyobb erejű, mint az előjelpróba. Az előjelpróbával szemben ennél a próbánál feltétel, hogy a különbség-eloszlás szimmetrikus legyen. Végrehajtása a következő: a mintaelemek közötti különbségeket rangsoroljuk az előjelektől függetlenül, az esetleges nullákat kihagyjuk.

A rangsorolást a következőképpen végezzük: az adatsort nagyság szerint sorba rendezzük, és a legkisebbnek adjuk az 1-es rangszámot, a következőnek a 2-est, stb. Összesen *n* rangszámot osztunk ki. Egyenlő adatok esetén is egyre növekvő rangszámot adunk, majd az egyenlő adatokhoz tartozó rangszámokat utólag korrigáljuk a megfelelő rangszámok átlagával (a korrigált rangszámokat kapcsolt rangszámoknak nevezzük). A rangsorolás helyességét úgy ellenőrizhetjük, hogy összeadjuk a kapott rangszámokat, ennek az összegnek meg kell egyeznie n(n+1)/2-vel (az első „n" egész szám összegével).

A próba során ezután külön összeadjuk a pozitív vagy a negatív különbségekhez tartozó rangszámokat (sőt, elég csak az egyiket). Ha igaz a nullhipotézis és a két sokaság azonos eloszlású, akkor a pozitív és a negatív különbségekhez tartozó rangszámösszegek körülbelül egyformák lesznek. Minél nagyobb az eltérés valamelyik a két sokaság között, annál nagyobb lesz az eltérés a két rangszámösszeg között is. Mekkora eltérést tekinthetünk még véletlenszerűnek? Kis mintaelemszám esetén (nl30 vagy nl50) táblázatok állnak rendelkezésre, amelyek adott -hoz megadják azt az intervallumot, amekkora rangszámösszeg még véletlen eltérésnek tekinthető. Nagy mintaelemszám esetére, vagy ha nagyon sok a kapcsolt rang, egy közelítően normális eloszlású statisztika segítségével a normális eloszlás alapján vizsgálható a szignifikancia:

$$z = \frac{\sum_{i=1}^{n} R_i}{\sqrt{\sum_{i=1}^{n} R_i^2}}$$

Itt a számlálóban az összes előjeles rangszám összege szerepel, a nevezőben pedig ezek négyzetösszegéből vont négyzetgyök. A számítógépes programrendszerek általában csak ezt a normális közelítésből származó *p*-értéket számítják ki még kis mintaelemszám esetén is, amikor pedig a közelítés nem túl jó.

58. Autoregressziós –és mozgóátlag folyamatok

- Autoregressziós folyamatok (AR)

  $X_t = a_1X_{t-1}+a_2X_{t-2} + a_3X_{t-3} + \ldots + a_pX_{t-p}+\varepsilon_t$

  - Stacionárius, ha az $1-(a_1s+a_2s^2 + \ldots + a_ps^p)=0$ egyenlet gyökei az egységkörön kívül helyezkednek el
- Mozgóátlag folyamatok (MA)

  $X_t = b_1 \varepsilon_{t-1}+b_2 \varepsilon_{t-2} + b_3 \varepsilon_{t-3} + \ldots + b_q \varepsilon_{t-q}+\varepsilon_t$

  - Mindig stacionárius
- Kombináció: ARMA folyamatok

Az Y_t diszkrét paraméterű sztochasztikus folyamatot k-ad rendű mozgóátlag folyamatnak nevezzük, ha

$$Y_t = \beta_0 * U_t + \beta_1 * U_{t-1} + \cdots + \beta_k * U_{t-k}$$

Ahol

- β_k konstansok
- U_t diszkrét fehér zaj (várható érték 0, szórás σ_u)

ARMA(p,q)

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t.2} + \cdots + \alpha_p Y_{t-}$$
$$\varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q},$$

Ahol

- ε_t fehér zaj
- p és q az autoregresszív és mozgóátlag folyamat rendje

Autoregresszív és Mozgóátlag modellek (autoregressive and moving-average)

Sztochasztikus idősorelemzés legegyszerűbb és leginkább elterjedt módszere

AR és MA folyamatokat egyesít

Paraméterek megállapítása általában empirikus idősor alapján

59. Mit jelent a becsléses illeszkedésvizsgálat és hogyan alkalmazzuk rá a χ-négyzet próbát?
Ha a vizsgálat csak arra vonatkozik, hogy valamilyen típusú eloszlás feltételezhetö-e, az eloszlás paramétereit a mintából becsülve, akkor becsléses illeszkedésvizsgálatokat alkalmazunk.

A gyakorlatban az eloszlásfüggvény alakjára van feltételezésünk, azonban az eloszlásfüggvény bizonyos paraméterei nem ismertek. Legyen

$$H_0 : P(X < t) = F(t; \vartheta_1, \ldots, \vartheta_s),$$

ahol az $F(t; \vartheta_1, \ldots, \vartheta_s)$ eloszlásfüggvényében a $(\vartheta_1, \ldots, \vartheta_s)$ (egydimenziós) paraméterek ismeretlenek. A minta alapján becsüljük meg az ismeretlen paramétereket maximum-likelihood módszerrel. Jelölje $\widehat{\vartheta}_i$ a ϑ_i maximum-likelihood becslését. Vizsgáljuk a

$$H_0' : P(X < t) = F(t; \widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_s)$$

hipotézist a korábban ismertetett eljárással. A módszerben a változás csupán annyi, hogy a $\mathcal{X}^2$-eloszlás szabadsági fokát a becsült paraméterek számával kell csökkenteni, azaz a kritikus értéket $\mathcal{X}^2_{r-1-s}$ táblázatából kell kikeresni.

H_0 hipotézis: az $A_1, A_2, \ldots, A_r$ teljes eseményrendszerre teljesül $P(A_1)=p_1$, $P(A_2)=p_2$, …, $P(A_r)=p_r$

A tesztstatisztika: $\sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}$

ami aszimptotikusan r-1 szabadságfokú χ-négyzet eloszlású, ha igaz a nullhipotézis.

Kritikus tartomány: ha a statisztika értéke nagyobb, mint az r-1 szabadságfokú χ-négyzet eloszlás 1- α kvantilise, elutasítjuk a nullhipotézist.

60. Bayes-tétele

Legyen B_1, B_2 , ..., pozitív valószínűségű eseményekből álló teljes eseményrendszer, $A \in \mathcal{A}$ pozitív valószínűségű. Ekkor

$$P(B_k \mid A) = \frac{P(A \mid B_k)P(B_k)}{\sum P(A \mid B_i)P(B_i)}$$

(Visszakövetkeztetés az első lépés eredményére.)

Bizonyítás. A nevező éppen $P(A)$ a teljes valószínűség tétele miatt.

A számláló pedig $P(A \cap B)$, definíció szerint.

61. Adja meg a legkisebb négyzetes becslést a lineáris modellben

A lineáris modell általános alakja:

$$Y = X\beta + \varepsilon,$$

ahol az Y vektor i-edik eleme a függő változó i-edik megfigyelése, az X mátrix i-edik sora a független (magyarázó) változókra vonatkozó i-edik megfigyelés, β becsülendő paraméter, ε nem megfigyelhető véletlen hiba. Legyen Y és X sorainak száma n, X oszlopai száma p. A β legkisebb négyzetes becslése $\widehat{\beta} = (X'X)^{-1}X'Y$. A homoszkedasztikus esetben (azaz amikor ε 0 várható értékű, azonos szórású és korrelálatlan koordinátákból áll) a fenti $\widehat{\beta}$ legjobb lineáris torzítatlan becslés.

`[b,bint,r,rint,stats]=regress(Y,X,alfa)` az alábbiakat számolja: `b` a fenti $\widehat{\beta}$ becslés (numerikusan a QR dekompozícióval számolva), `bint` β-ra konfidencia intervallum (t-eloszlás alapján számolva), `r` a maradék vektor, `rint` erre konfidencia intervallum. A maradék definíciója: $r = Y - X\widehat{\beta}$, ennek alkalmas normalizáltja t-eloszlású, ami alapján a konfidencia intervallum megszerkeszthető. A konfidencia intervallumok szintje: $100(1-\alpha)\%$. A `stats` négy értéket tartalmaz: az R^2 statisztikát (amely a maradék négyzetösszeg és a teljes négyzetösszeg hányadosa, $R^2 \leq 1$, és minél közelebb van 1-hez, annál jobb az illeszkedés), az F statisztikát, annak szignifikancia szintjét és a hiba szórásnégyzetének becslését. A program kódja alapján a fenti F-próba a [konstans + zaj] modellt teszteli az általános modellel szemben. Ha a modell ténylegesen [konstans + zaj] alakú (és X oszlopai között szerepel a konstans), akkor F eloszlása centrált $F(p-1, n-p)$.