Várható értékre és szórásra vonatkozó próbák

1. rész

– egy MATLAB® alapú megközelítés –

Baja Zsolt, Vas Orsolya

Matematika és Informatika Intézet, Babeş–Bolyai Tudományegyetem, Kolozsvár, Románia

(bajazsolt98@gmail.com, vas.orsolya@yahoo.com)

9. labor / 2023. november 27. – december 1.

- Adott statisztikai sokaság összes elemét általában lehetetlen, vagy legalábbis túlságosan költséges akár egyetlen jellemző alapján is felmérni. Ezért célszerű inkább a sokaságnak egy megfelelő méretű független mintavételét tanulmányozni az adott jellemző(k) alapján, majd a kapott eredményeket globálisan kiterjeszteni a sokaság minden elemére, természetesen egy jól meghatározott hibahatár (pontosabban fogalmazva, biztonsági szint) mellett.
- Ezért a mostani és az elkövetkező laborok leírásában eljárásokat adunk valamely statisztikai sokaság viselkedését meghatározó – általában ismeretlen paraméterezésű eloszlást követő – valószínűségi változó két legfontosabb jellemző adatának, a várható érték és a szórásnégyzet, *intervallumbecslésére*, valamint ezekkel kapcsolatos *feltevések* (másképpen *hipotézisek*) adott megbízhatósági szintnek megfelelő *ellenőrzésére*. Ezeket a módszereket *statisztikai próbáknak* nevezzük és hamarosan tapasztalni fogjuk, hogy működésük lényege ugyanaz, csupán a használt eszközökben (eloszlásfüggvények típusában) különböznek egymástól.
- Különbséget teszünk *egy- és kétmintás*, valamint ezeken belül várható értékkel és szórásnégyzettel, illetve ezen paraméterek összehasonlításával kapcsolatos statisztikai próbák között.

- Az ismeretlen elméleti várható értékre és szórásra vonatkozó próbák ismertetése előtt megjegyezzük, hogy függetlenül attól, hogy milyen eloszlást jellemző, ismeretlen paraméterre is alkalmazunk valamilyen feltevést (hipotézist) ellenőrző próbát, mindig *két kockázati tényezővel* számolhatunk, amit az alábbiakban értelmezünk. Érthetőségük végett, viszont néhány fogalmat meg kell előlegeznünk.
- A későbbiekben ismertetendő próbák módszertanát összehasonlítva, egy ismétlődő minta szerint zajló *döntési folyamatot* fedezhetünk fel bennük. Ugyanis egy rögzített megbízhatósági szint mellett, az ismeretlen paraméterre mintavételi tapasztalatok alapján megfogalmazott H_0 *nullhipotézist* egy annak ellentmondó H_1 *alternatív hipotézissel* szembesítjük.
- A H_0 nullhipotézist egy, vagy több független mintavétel előzetes tanulmányozása alapján fogalmazzuk meg, ahol mindig azt feltételezzük, hogy a minták valamilyen ismeretlen paraméterezésű normális eloszlásból származnak. Mindegyik hipotézist ellenőrző próba egy jól ismert (mint például standard normális, Student-, vagy χ^2-) eloszlású valószínűségi változóra épül, amely egyrészt az adott mintavétel(ek)től, másrészt az ismeretlen normális eloszlás paramétereitől is függ.
- Amint hamarosan látni fogjuk, az ellenőrzési folyamatot általában az előbb említett jól ismert eloszlású valószínűségi változó – H_1 alternatív hipotézistől függő – *intervallumbecslésére* vezethetjük vissza. Az így kapott intervallumot *megbízhatósági*, vagy *konfidenciaintervallumnak* nevezzük és annak két végpontját $1-\alpha$ valószínűséggel biztosítjuk, ahol az $\alpha \in (0,1)$ paramétert *szignifikanciaszintnek* nevezzük.

- Behelyettesítve a szóban forgó valószínűségi változó alakjába a H_0 nullhipotézis által igaznak feltételezett ismeretlen normális eloszlást jellemző paramétereket, olyan valós számot kapunk, amit a *próba értékének* nevezünk. Amennyiben a próba értéke beleesik az előzetesen meghatározott konfidenciaintervallumba, akkor a H_0 nullhipotézist elfogadjuk, máskülönben elutasítjuk azt, és a H_1 alternatív hipotézist mellett döntünk az adott α szignifikanciaszint mellett.
- A fentiek alapján már érezhető, hogy milyen két kockázati tényezőt kell figyelembe vennünk. Előfordulhat ugyanis, hogy egy igaz nullhipotézist utasítunk el (ezt az eseményt *elsőfajú hibának* nevezzük), vagy egy hamis nullhipotézist fogadunk el (amihez tartozó eseményt *másodfajú hibának* mondjuk).
- Nyilván szeretnénk minimálisra csökkenti mindkét hiba bekövetkezési valószínűségét. Sajnos, ha valamelyik hiba valószínűségét túlságosan kicsire állítjuk, akkor általában a másik bekövetkezési valószínűsége nő. A legtöbb statisztikai próba esetén, viszont a két hiba bekövetkezési valószínűsége csökken, ha a próbában megjelent független mintavétel(ek) méretét megnöveljük és újra alkalmazzuk az adott módszert.

- Feltételes valószínűségekkel az első és másodfajú hiba valószínűségét az

$$\alpha = P\left(\text{„a } H_0 \text{ nullhipotézist elutasítjuk''} \mid H_0\right),$$

illetve a

$$\beta = P\left(\text{„a } H_0 \text{ nullhipotézist elfogadjuk''} \mid H_1\right)$$

alakra hozhatjuk, vagyis értelmezés szerint az α szignifikanciaszint éppen az elsőfajú hiba valószínűségét rögzíti.

- Az egymintás U-próba az alábbi 1. feladatban megfogalmazott kérdésre ad választ. A próba lehetséges változatait és alkalmazását a feladat megoldásából sajátíthatjuk el.

1. Feladat (Egymintás U-próba)

Tekintsük az $X \sim \mathcal{N}(\mu, \sigma)$ valószínűségi változónak az n-elemű független $\{X_j\}_{j=1}^n$ mintavételét ($n \geq 2$), ahol a $\mu \in \mathbb{R}$ elméleti várható értéket ismeretlen, a $\sigma > 0$ szórást pedig ismert paraméternek feltételezzük! Az $\alpha \in (0, 1)$ szignifikanciaszint mellett döntsük el a

$$H_0 : \mu = \mu_0$$

nullhipotézis helyességét a

$$H_k : \mu \neq \mu_0 \text{ (kétoldali próba)},$$
$$H_j : \mu > \mu_0 \text{ (jobb oldali próba)},$$
$$H_b : \mu < \mu_0 \text{ (bal oldali próba)}$$

lehetséges alternatív hipotézisekkel szemben, ahol $\mu_0 \in \mathbb{R}$ egy általunk megfogalmazott tipp az ismeretlen μ várható értékre. Ugyanakkor szerkesszünk mindhárom esetben megbízhatósági (másképpen konfidencia-) intervallumot is az ismeretlen μ paraméterre!

Megoldás

- Az
$$U_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$$
valószínűségi változót használjuk a feladat megoldására, ahol $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ a mintavételi (másképpen becsült) várható értéket jelöli.
- A $\mu = \mu_0$ helyettesítés során kapott
$$U_n^0 = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{1}$$
számot a próba u-értékének nevezzük.
- Mindhárom lehetséges alternatív hipotézis esetén a
$$P(u_{\min} < U_n < u_{\max}) = 1 - \alpha \tag{2}$$
valószínűségben megjelenő intervallum $u_{\min}$ és $u_{\max}$ végpontjaira adunk egy-egy feltételt.

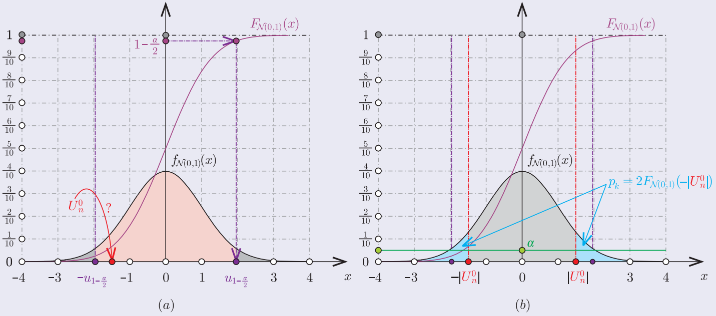
Megoldás – folytatás

- A
$$\begin{cases} H_0 & : \quad \mu = \mu_0, \\ H_k & : \quad \mu \neq \mu_0 \end{cases}$$
kétoldali U-próba esetében az $(u_{\min}, u_{\max})$ intervallumot az origóra nézve szimmetrikusan választjuk meg. Ezért a (2)-es egyenlet megoldását az $u_{\min} = -u$ és $u_{\max} = u$ alakban keressük, ahol $u > 0$.
- Ekkor az
$$\begin{aligned} 1 - \alpha &= P\left(-u < U_n < u\right) \\ &= F_{\mathcal{N}(0,1)}\left(u\right) - F_{\mathcal{N}(0,1)}\left(-u\right) \\ &= F_{\mathcal{N}(0,1)}\left(u\right) - \left[1 - F_{\mathcal{N}(0,1)}\left(u\right)\right] \\ &= 2F_{\mathcal{N}(0,1)}\left(u\right) - 1 \end{aligned}$$
egyenlőségnek kell teljesülnie, ahonnan az
$$u = u_{1-\frac{\alpha}{2}} = F^{-1}_{\mathcal{N}(0,1)}\left(1 - \frac{\alpha}{2}\right)$$
inverz értéket, vagy másképpen az $1 - \frac{\alpha}{2}$ valószínűséghez tartozó kvantilist kapjuk.

Megoldás – folytatás

- Tehát az U_n valószínűségi változóra a
$$-u_{1-\frac{\alpha}{2}} < U_n < u_{1-\frac{\alpha}{2}} \tag{3}$$
feltétel adódik.
- Amennyiben az igaznak feltételezett nullhipotézis által eredményezett (1)-es u-értékre teljesül az
$$U_n^0 \in \left(-u_{1-\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}\right)$$
bennfoglalás, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, ellenkezőleg a $H_k : \mu \neq \mu_0$ alternatív hipotézist fogadjuk el az adott α szignifikanciaszint mellett. Ezt a döntési folyamatot az 1.(a) ábra szemlélteti.

Megoldás – folytatás



1. ábra. (a) A független és azonos $\mathcal{N}(\mu, \sigma)$-eloszlású valószínűségi változókból álló $\{X_j\}_{j=1}^{n}$ mintavétel ismeretlen $\mu \in \mathbb{R}$ elméleti várható értékére szerkesztett megbízhatósági intervallumot láthatjuk a kétoldali U-próba esetén. Ha $U_n^0 \in \left(-u_{1-\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}\right)$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, ellenkezőleg a $H_k : \mu \neq \mu_0$ alternatív hipotézist fogadjuk el az adott $\alpha \in (0, 1)$ szignifikanciaszint mellett. (b) Ugyanezen hipotézisek mellett a két oldali szignifikanciapróbát is alkalmazhatjuk a $p_k = 2P\left(U_n > \left|U_n^0\right| \,\middle|\, H_0\right) = 2F_{\mathcal{N}(0,1)}\left(-\left|U_n^0\right|\right)$ értékkel. Ha $p_k > \alpha$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, máskülönben a $H_k : \mu \neq \mu_0$ alternatív hipotézist fogadjuk el. Mindkét próba esetén az $U_n^0 = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ mennyiség a próbák közös u-értékét jelöli, amelyet az igaznak feltételezett $H_0 : \mu = \mu_0$ nullhipotézis eredményezett.

Megoldás – folytatás

- Az U_n változó értelmezését használva a (3)-as egyenlőtlenségpárt a

$$\mu_{\min}^{k} = \overline{X} - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} = \mu_{\max}^{k}$$

  alakra hozhatjuk. Ily módon egy megbízhatósági intervallumot szerkesztettünk az ismeretlen μ elméleti várható értékre. Ekkor a következőképpen dönthetünk: amennyiben $\mu_0 \in \left(\mu_{\min}^{k}, \mu_{\max}^{k}\right)$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, máskülönben a $H_k : \mu \neq \mu_0$ alternatív hipotézist fogadjuk el az adott α szignifikanciaszint mellett.

- Vegyük észre, hogy ugyanezen null- és alternatív hipotézisek mellett másképpen is gondolkodhatunk az alábbi sorokban leírt kétoldali szignifikanciapróba segítségével! Tekintsük azt a

$$p_k = 2P\left(U_n > \left|U_n^0\right| \,\middle|\, H_0\right) = 2F_{\mathcal{N}(0,1)}\left(-\left|U_n^0\right|\right)$$

  feltételes valószínűségét, mely szerint legalább akkora elméleti u-értéket kapunk, mint a mintavételhez tartozó U_n^0 érték, amelyet az igaznak feltételezett $H_0 : \mu = \mu_0$ nullhipotézist során tapasztaltunk.

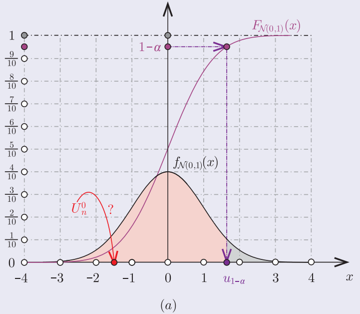
Megoldás – folytatás

- Az így kapott valószínűséget a szignifikanciapróba p-értékének nevezzük és geometriai jelentését az 1.(b) ábrán szemléltetjük. Vegyük észre, hogy a p-érték csak a mintavételtől és az $\mathcal{N}(0,1)$-eloszlástól függ! A p-érték a nullhipotézis elleni bizonyíték erősségének mértékét fejezi ki, azaz minél kisebb a p-érték, annál erősebb bizonyíték szól a nullhipotézis ellen.
- A p-érték alapján a következőképpen dönthetünk: ha $p_k > \alpha$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, ellenben a $H_k : \mu \neq \mu_0$ alternatív hipotézist fogadjuk el.

Megoldás – folytatás

- A
$$\begin{cases} H_0 & : \quad \mu = \mu_0, \\ H_j & : \quad \mu > \mu_0 \end{cases}$$
jobb oldali U-próba esetén az $u_{\min} = -\infty$ és $u_{\max} = u \in \mathbb{R}$ beállításokat végezzük.
- Ekkor az
$$1 - \alpha = P\left(-\infty < U_n < u\right) = F_{\mathcal{N}(0,1)}\left(u\right)$$
egyenletből az
$$u = u_{1-\alpha} = F^{-1}_{\mathcal{N}(0,1)}\left(1 - \alpha\right)$$
inverz értéket, vagy másképpen az $1 - \alpha$ valószínűséghez tartozó kvantilist kapjuk.
- Az így adódó
$$-\infty < U_n < u_{1-\alpha} \tag{4}$$
feltétel alapján a következőképpen dönthetünk: ha $U_n^0 \in \left(-\infty, u_{1-\alpha}\right)$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, máskülönben a $H_j : \mu > \mu_0$ alternatív hipotézist fogadjuk el az adott szignifikanciaszint mellett. A döntéshozatal mikéntjét a 2.(a) ábrán illusztráltuk.

Megoldás – folytatás



2. ábra. (a) A független és azonos $\mathcal{N}(\mu, \sigma)$-eloszlású valószínűségi változókból álló $\{X_j\}_{j=1}^n$ mintavétel ismeretlen $\mu \in \mathbb{R}$ elméleti várható értékére szerkesztett konfidenciaintervallumot figyelhetjük meg a jobb oldali U-próba alkalmazásakor. Ha $U_n^0 \in (-\infty, u_{1-\alpha})$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, ellenkezőleg a $H_j : \mu > \mu_0$ alternatív hipotézist fogadjuk el az adott $\alpha \in (0,1)$ szignifikanciaszint mellett. (b) Ugyanezen hipotézisek mellett a jobb oldali szignifikanciapróbát is alkalmazhatjuk a $p_j = P\left(U_n > \left|U_n^0\right| \middle| H_0\right) = 1 - F_{\mathcal{N}(0,1)}\left(U_n^0\right)$ értékkel. Ha $p_j > \alpha$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, máskülönben a $H_j : \mu > \mu_0$ alternatív hipotézist fogadjuk el. Mindkét próba esetén az $U_n^0 = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ mennyiség a próbák közös u-értékét jelöli, amelyet az igaznak feltételezett $H_0 : \mu = \mu_0$ nullhipotézis eredményezett.

Megoldás – folytatás

- Az U_n változó alakját használva, a (4)-es egyenlőtlenségpárt a
$$\mu_{\min}^{j} = \overline{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha} < \mu < +\infty = \mu_{\max}^{j}$$
alakra hozhatjuk, azaz ugyancsak egy konfidenciaintervallumot szerkeszthetünk az ismeretlen μ elméleti várható értékre.
- Hasonlóan a kétoldali U-próbához, a következőképpen dönthetünk: amennyiben $\mu_0 \in \left(\mu_{\min}^{j}, \mu_{\max}^{j}\right)$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, máskülönben a $H_j : \mu > \mu_0$ alternatív hipotézist fogadjuk el az adott α szignifikanciaszint mellett.
- A $H_j : \mu > \mu_0$ alternatív hipotézis esetén a 2.(b) ábrán bemutatott jobb oldali szignifikanciapróbát is alkalmazhatjuk a
$$p_j = P\left(U_n > U_n^0 \mid H_0\right) = 1 - F_{\mathcal{N}(0,1)}\left(U_n^0\right)$$
értékkel. Ekkor a döntésünk is teljesen hasonló a kétoldali szignifikanciapróbához: ha $p_j > \alpha$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, ellenben a $H_j : \mu > \mu_0$ alternatív hipotézist fogadjuk el.

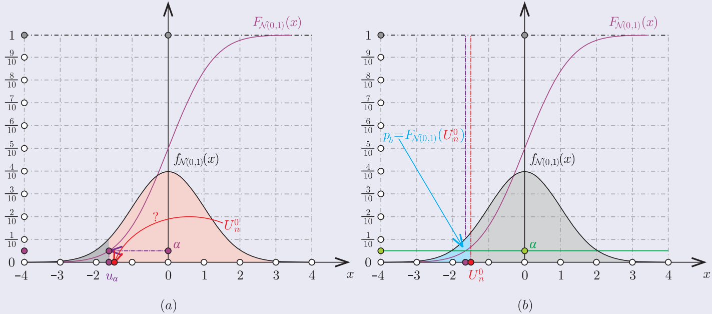
Megoldás – folytatás

- A
$$\begin{cases} H_0 & : \quad \mu = \mu_0, \\ H_b & : \quad \mu < \mu_0 \end{cases}$$
bal oldali U-próbára térve az $u_{\min} = u \in \mathbb{R}$ és $u_{\max} = +\infty$ megválasztásokkal élünk.
- Az
$$1 - \alpha = P\left(u < U_n < +\infty\right) = 1 - F_{N(0,1)}\left(u\right)$$
egyenlet megoldását ekkor az
$$u = u_\alpha = F^{-1}_{N(0,1)}\left(\alpha\right)$$
kvantilis adja, amellyel az
$$u_\alpha < U_n < +\infty, \tag{5}$$
vagy az ezzel ekvivalens
$$\mu^b_{\min} = -\infty < \mu < \overline{X} - \frac{\sigma}{\sqrt{n}} u_\alpha = \mu^b_{\max}$$
egyenlőtlenségpárt kapjuk.
- Ugyanilyen hipotézisek mellett a bal oldali szignifikanciapróba p-értékére a
$$p_b = P\left(U_n < U_n^0 \mid H_0\right) = F_{\mathcal{N}(0,1)}\left(U_n^0\right)$$
valószínűség adódik.

Megoldás – folytatás

- A döntéshozatalunk (lásd a 3.(*a*)–(*b*) ábrákat) is az eddigi mintát követi:
  - ha $U_n^0 \in (u_\alpha, +\infty)$, vagy $\mu_0 \in \left(\mu_{\min}^b, \mu_{\max}^b\right)$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, máskülönben a $H_b : \mu < \mu_0$ alternatív hipotézist fogadjuk el az adott szignifikanciaszint mellett;
  - ha $p_b > \alpha$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, máskülönben a $H_b : \mu < \mu_0$ alternatív hipotézist fogadjuk el.

Megoldás – folytatás



3. ábra. (a) A független és azonos $\mathcal{N}(\mu, \sigma)$-eloszlású valószínűségi változókból álló $\{X_j\}_{j=1}^n$ mintavétel ismeretlen $\mu \in \mathbb{R}$ elméleti várható értékére szerkesztett megbízhatósági intervallumot láthatjuk a bal oldali U-próba alkalmazásakor. Ha $U_n^0 \in (u_\alpha, +\infty)$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, ellenkezőleg a $H_b : \mu < \mu_0$ alternatív hipotézist fogadjuk el az adott $\alpha \in (0,1)$ szignifikanciaszint mellett. (b) Ugyanezen hipotézisek mellett a bal oldali szignifikanciapróbát is alkalmazhatjuk a $p_b = P\left(U_n < \left|U_n^0\right| \,\middle|\, H_0\right) = F_{\mathcal{N}(0,1)}\left(U_n^0\right)$ értékkel. Ha $p_b > \alpha$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, ellenkezőleg a $H_b : \mu < \mu_0$ alternatív hipotézist fogadjuk el. Mindkét próba esetén az $U_n^0 = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ mennyiség a próbák közös u-értékét jelöli, amelyet az igaznak feltételezett $H_0 : \mu = \mu_0$ nullhipotézis eredményezett.

Megoldás – folytatás

- Mindhárom esetet összegezve 1. kódrészletbeli UTest függvényt ajánljuk a különböző oldalú U- és szignifikanciapróbák alkalmazására, valamint az ismeretlen $\mu \in \mathbb{R}$ elméleti várható értékre vonatkozó megbízhatósági intervallumok meghatározására.

1. Kódrészlet. Az egymintás U-próba MATLAB®-beli implementálása

```
% ____________
% Description
% ____________
% The function performs a one-sample U-test[1] of the null hypothesis[2] H_0 : mu = mu_0 that the
% data in the vector X = {X_j}_{j=1}^n comes from an N(mu, sigma) distribution, where the
% theoretical mean value mu ∈ R is unknown and the theoretical standard deviation sigma > 0
% is a known parameter.
%
% The test is performed against the alternative hypothesis specified by the input
% parameter tail.
%
% ______
% Input
% ______
% X = {X_j}_{j=1}^n - an independent[3] and identically distributed[4] sample of the normal
%                     distribution N(mu, sigma), where the theoretical mean value mu ∈ R is unknown
```

Implementálás – II

Egymintás U-próba az ismeretlen elméleti várható értékre

```
% mu_0          — denotes the guess[5] $\mu_0$ of the user for the unknown theoretical mean value
% sigma         — stores the known value of theoretical standard deviation $\sigma > 0$
% alpha         — represents the significance level $\alpha \in (0, 1)$
% tail          — a parameter that can be set either by one of the strings 'both',
%                 'right', 'left', or by using one of the numbers 0, 1, -1 (it determines
%                 the type of the alternative hypothesis)
%
% ———————
% Output
% ———————
% ci_u          — confidence interval[6] for the random variable $U_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$
% ci_mu         — confidence interval for theoretical mean value $\mu$
% u_value       — value of the test, assuming that the null hypothesis $H_0 : \mu = \mu_0$ is true
% p_value       — probability of observing the given result, or one more extreme, by
%                 chance if the null hypothesis $H_0 : \mu = \mu_0$ is true (small $p$-values cast
%                 doubt on the validity of the null hypothesis)
% H             — the code of the accepted hypothesis (0 = null hypothesis,
%                 1 = alternative hypothesis defined by the input parameter tail)
%
function [ci_u, ci_mu, u_value, p_value, H] = UTest(X, mu_0, sigma, alpha, tail)

% check the validity of input parameters!
...

% get the size of the sample $\{X_j\}_{j=1}^n$
n = length(X);

```

Implementálás – III

Egymintás U-próba az ismeretlen elméleti várható értékre

```
% observe that $\frac{\sigma}{\sqrt{n}}$ is a constant
s = sigma / sqrt(n);

% calculate the sample mean $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$
X_ = mean(X); % or, equivalently, X_ = sum(X) / n;

% calculate the u-value $U_n^0 = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
u_value = (X_ - mu_0) / s;

% calculate the confidence intervals
switch (tail)
    case {'both', 0}
        % calculate the inverse function value $u_{1-\frac{\alpha}{2}} = F^{-1}_{\mathcal{N}(0,1)}\left(1 - \frac{\alpha}{2}\right)$
        u = norminv(1 - alpha / 2, 0, 1);

        % store the end points of the confidence interval $\left(-u_{1-\frac{\alpha}{2}}, +u_{1-\frac{\alpha}{2}}\right)$
        ci_u(1) = -u;
        ci_u(2) =  u;

        % store the end points of the confidence interval $\left(\overline{X} - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}, \overline{X} + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}\right)$
        ci_mu(1) = X_ - s * u;
        ci_mu(2) = X_ + s * u;
```

```

        % calculate the p-value p = 2F_{N(0,1)}(-|U_n^0|)
        p_value = 2.0 * normcdf(-abs(u_value), 0, 1);

    case {'right', 1}
        % calculate the inverse function value u_{1-alpha} = F^{-1}_{N(0,1)}(1 - alpha)
        u = norminv(1 - alpha, 0, 1);

        % store the end points of the confidence interval (-inf, u_{1-alpha})
        ci_u(1) = -inf;
        ci_u(2) =  u;

        % store the end points of the confidence interval (X̄ - sigma/sqrt(n) u_{1-alpha}, +inf)
        ci_mu(1) = X_ - s * u;
        ci_mu(2) = inf;

        % calculate the p-value p = 1 - F_{N(0,1)}(U_n^0)
        p_value = 1.0 - normcdf(u_value, 0, 1);

    case {'left', -1}
        % calculate the inverse function value u_alpha = F^{-1}_{N(0,1)}(alpha)
        u = norminv(alpha, 0, 1);

        % store the end points of the confidence interval (u_alpha, +inf)
        ci_u(1) = u;
        ci_u(2) = inf;
```

```

        % store the end points of the confidence interval $\left(-\infty, \overline{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha}\right)$
        ci_mu(1) = -inf;
        ci_mu(2) = X_ - s * u;

        % calculate the p-value $p = F_{\mathcal{N}(0,1)}\left(U_n^0\right)$
        p_value = normcdf(u_value, 0, 1);
end

% make your decision based on confidence intervals, and note that, the
% condition (mu_0 > ci_mu(1) && mu_0 < ci_mu(2)) would also be correct!
H = ~(u_value > ci_u(1) && u_value < ci_u(2));

% do you have any doubt? -- if yes, then apply the corresponding significance test!
if (p_value < alpha)
    disp('Warning: small p-value cast doubt on the validity of the null-hypothesis!');
end
```

[1] **test** *próba*; **left-tailed** ∼ *bal oldali próba*; **one-sample** ∼ *egymintás próba*; **right-tailed** ∼ *jobb oldali próba*; **statistical** ∼ *statisztikai próba*; **two-sample** ∼ *kétmintás próba*; **two-tailed** ∼ *kétoldali próba*

[2] **hypothesis** *hipotézis*; **alternative** ∼ *alternatív hipotézis*; **null** ∼ *nullhipotézis*

[3] **independent** *független*; ∼ **random variables** *független valószínűségi változók*

[4] **identically distributed** *azonos eloszlású*; ∼ **random variables** *azonos eloszlású valószínűségi változók*

[5] **guess** *találgatás, feltételezés*

[6] **interval** *intervallum*; **confidence** ∼ *konfidenciaintervallum, vagy megbízhatósági intervallum*

- Ebben a pontban két független, nem feltétlenül azonos normális eloszlású valószínűségi változó ismeretlen elméleti várható értékét hasonlítjuk össze, feltéve, hogy a változók elméleti szórásait, továbbá egy-egy független mintavételüket is ismerjük.
- Az egymintás U-próba esetén már láttuk, hogy a teszt alkalmazásának a kulcsa egy olyan mintavételtől függő valószínűségi változó volt, amely elméleti eloszlását ismertük. Jelen esetben is szükségünk lesz egy jól ismert eloszlású valószínűségi változóra, amely már mindkét változó mintavételezésétől is függ.

1. Tétel (A kétmintás U-próba valószínűségi változója)

Tekintsük az $X \sim \mathcal{N}(\mu_1, \sigma_1)$ és $Y \sim \mathcal{N}(\mu_2, \sigma_2)$ független valószínűségi változókat, valamint ezeknek egy-egy független $\{X_i\}_{i=1}^m$ és $\{Y_j\}_{j=1}^n$ mintavételét $(m \geq 2,\ n \geq 2)$, ahol $\mu_1, \mu_2 \in \mathbb{R}$ és $\sigma_1, \sigma_2 > 0$! Ekkor az

$$U_{m,n} = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \tag{6}$$

valószínűségi változó $\mathcal{N}(0,1)$-eloszlású.

Bizonyítás

Lásd a nyomtatott jegyzet 177., vagy az elektronikus változat 184. oldalát!

- Az egymintás változathoz hasonlóan a kétmintás U-próbát is megoldott feladatként ismertetjük.

2. Feladat (Kétmintás U-próba)

Tekintsük az $X \sim \mathcal{N}(\mu_1, \sigma_1)$ és $Y \sim \mathcal{N}(\mu_2, \sigma_2)$ független valószínűségi változókat, valamint ezeknek egy-egy független $\{X_i\}_{i=1}^m$ és $\{Y_j\}_{j=1}^n$ mintavételét ($m \geq 2$, $n \geq 2$), ahol az elméleti $\mu_1, \mu_2 \in \mathbb{R}$ várható értékek ismeretlen, az elméleti $\sigma_1, \sigma_2 > 0$ szórások pedig ismert paraméterek! Az $\alpha \in (0,1)$ szignifikanciaszint mellett döntsük el a

$$H_0 : \mu_1 - \mu_2 = 0$$

nullhipotézis helyességét a

$$H_k : \mu_1 - \mu_2 \neq 0 \text{ (kétoldali próba)},$$
$$H_j : \mu_1 - \mu_2 > 0 \text{ (jobb oldali próba)},$$
$$H_b : \mu_1 - \mu_2 < 0 \text{ (bal oldali próba)}$$

lehetséges alternatív hipotézisekkel szemben! Ugyanakkor szerkesszünk mindhárom esetben megbízhatósági intervallumot az ismeretlen elméleti várható értékek $\mu_1 - \mu_2$ különbségére, továbbá implementáljuk is ezt a statisztikai módszert!

Megoldás

- Mindenekelőtt vegyük észre, hogy a null- és a lehetséges alternatív hipotézisek pontosan azt fejezik ki, hogy az ismeretlen μ_1 és μ_2 elméleti várható értékek megegyeznek (H_0), különböznek (H_k), $\mu_1 > \mu_2$ (H_j), illetve $\mu_1 < \mu_2$ (H_b)! Tehát a kétmintás U-próba esetén valóban összehasonlításról beszélhetünk.
- A feladat megoldásának gondolatmenete teljesen hasonló az egymintás U-próba módszeréhez, amelyet az 1. feladatban írtunk le.
- Az árnyalatnyi különbséget az adja, hogy az U_n valószínűségi változó helyett a (6)-os kifejezésben bevezetett $U_{m,n}$ változó alakját kell használni, továbbá, amíg az egymintás U-próba esetén tetszőlegesen megválaszthattuk a $\mu_0 \in \mathbb{R}$ tippet az ismeretlen elméleti várható értékre, addig a jelen esetben csak rögzített, 0-értékű feltételezésünk van az ismeretlen elméleti várható értékek $\mu_1 - \mu_2$ különbségére nézve.
- Ezért csak felsoroljuk mindhárom alternatív hipotézis esetén kapott megbízhatósági intervallumokat, valamint a megfelelő szignifikanciapróbák p-értékét is ismertetjük. A részletes számításokat az olvasóra bízzuk.

Megoldás – folytatás

- A $H_0 : \mu_1 - \mu_2 = 0$ nullhipotézist igaznak feltételezve, a kétmintás U-próba u-értékére az

$$U_{m,n}^0 = \frac{\left(\overline{X} - \overline{Y}\right) - 0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

  kifejezést kapjuk.
- Akárcsak az 1. feladat megoldásában, az $F_{\mathcal{N}(0,1)}$ eloszlásfüggvénynek az $1 - \frac{\alpha}{2}$, $1 - \alpha$ és α valószínűségekhez tartozó inverz értékeit az $u_{1-\frac{\alpha}{2}}$, $u_{1-\alpha}$, illetve az u_α kvantilisekkel jelöljük.
- Ekkor mindhárom lehetséges alternatív hipotézis esetén az alábbiakat írhatjuk.

Megoldás – folytatás

- A
$$\begin{cases} H_0 & : \quad \mu_1 - \mu_2 = 0, \\ H_k & : \quad \mu_1 - \mu_2 \neq 0 \end{cases}$$
kétoldali, kétmintás U-próba esetén az $U_{m,n}$ változóra a
$$\left(-u_{1-\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}\right), \tag{7}$$
a $\delta = \mu_1 - \mu_2$ különbségre pedig a
$$\left(\delta_{\min}^k, \delta_{\max}^k\right) = \left(\overline{X} - \overline{Y} - \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \cdot u_{1-\frac{\alpha}{2}}, \overline{X} - \overline{Y} + \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \cdot u_{1-\frac{\alpha}{2}}\right) \tag{8}$$
megbízhatósági intervallumot kapjuk.
- Ugyanezen hipotézisek mellett a kétoldali szignifikanciapróba p-értékét a
$$p_k = 2P\left(U_{m,n} > \left|U_{m,n}^0\right| \mid H_0\right) = 2F_{\mathcal{N}(0,1)}\left(-\left|U_{m,n}^0\right|\right)$$
valószínűség adja.

Megoldás – folytatás

- A döntés is a már megszokott minta alapján születik:
  - amennyiben $U_{m,n}^0 \in \left(-u_{1-\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}\right)$, vagy $0 \in \left(\delta_{\min}^k, \delta_{\max}^k\right)$, akkor a $H_0 : \mu_1 = \mu_2$ nullhipotézist, máskülönben a $H_k : \mu_1 \neq \mu_2$ alternatív hipotézist fogadjuk el az adott szignifikanciaszint mellett;
  - ha $p_k > \alpha$, akkor a $H_0 : \mu_1 = \mu_2$ nullhipotézist, ellenben a $H_k : \mu_1 \neq \mu_2$ alternatív hipotézist fogadjuk el.
- A

$$\begin{cases} H_0 & : \quad \mu_1 - \mu_2 = 0, \\ H_j & : \quad \mu_1 - \mu_2 > 0 \end{cases}$$

jobb oldali, kétmintás U-próba alkalmazásakor az $U_{m,n}$ változóra a

$$\left(-\infty, u_{1-\alpha}\right), \tag{9}$$

a $\delta = \mu_1 - \mu_2$ különbségre pedig a

$$\left(\delta_{\min}^j, \delta_{\max}^j\right) = \left(\overline{X} - \overline{Y} - \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \cdot u_{1-\alpha}, +\infty\right) \tag{10}$$

konfidenciaintervallum adódik.

Megoldás – folytatás

- Ugyanezen hipotézisek mellett a jobb oldali szignifikanciapróba p-értékét a
$$p_j = P\left(U_{m,n} > U_{m,n}^0 \,\middle|\, H_0\right) = 1 - F_{\mathcal{N}(0,1)}\left(U_{m,n}^0\right)$$
valószínűség adja.
- A nullhipotézis elfogadása, vagy elutasítása az alábbi feltételek alapján történik:
  - amennyiben $U_{m,n}^0 \in (-\infty, u_{1-\alpha})$, vagy $0 \in \left(\delta_{\min}^j, \delta_{\max}^j\right)$, akkor a $H_0 : \mu_1 = \mu_2$ nullhipotézist, ellenkezőleg a $H_j : \mu_1 > \mu_2$ alternatív hipotézist fogadjuk el az adott szignifikanciaszint mellett;
  - ha $p_j > \alpha$, akkor a $H_0 : \mu_1 = \mu_2$ nullhipotézist, máskülönben a $H_j : \mu_1 > \mu_2$ alternatív hipotézist fogadjuk el.
- A
$$\begin{cases} H_0 & : & \mu_1 - \mu_2 = 0, \\ H_b & : & \mu_1 - \mu_2 < 0 \end{cases}$$
bal oldali, kétmintás U-próba esetén az $U_{m,n}$ valószínűségi változóra az
$$\left(u_\alpha, +\infty\right), \tag{11}$$
a $\delta = \mu_1 - \mu_2$ eltérésre pedig a

Megoldás – folytatás

$$\left(\delta_{\min}^{b}, \delta_{\max}^{b}\right) = \left(-\infty, \overline{X} - \overline{Y} - \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \cdot u_{\alpha}, \right) \tag{12}$$

megbízhatósági intervallumot szerkeszthetjük meg.

- Ugyanezen hipotézisek mellett a bal oldali szignifikanciapróba p-értékét a
$$p_b = P\left(U_{m,n} < U_{m,n}^0 \,\middle|\, H_0\right) = F_{\mathcal{N}(0,1)}\left(U_{m,n}^0\right)$$
valószínűség adja.
- Az eddigi esetekhez hasonlóan az alábbi feltételek alapján dönthetünk:
  - amennyiben $U_{m,n}^0 \in (u_\alpha, +\infty)$, vagy $0 \in \left(\delta_{\min}^{b}, \delta_{\max}^{b}\right)$, akkor a $H_0 : \mu_1 = \mu_2$ nullhipotézist, máskülönben a $H_b : \mu_1 < \mu_2$ alternatív hipotézist fogadjuk el az adott szignifikanciaszint mellett;
  - ha $p_b > \alpha$, akkor a $H_0 : \mu_1 = \mu_2$ nullhipotézist, ellenben a $H_b : \mu_1 < \mu_2$ alternatív hipotézist fogadjuk el.
- A kétmintás U-próba részleges implementálását a 2. kódrészletben listáztuk. *A fennmaradt esetek befejezését az olvasóra bízzuk!*

2. Kódrészlet. A kétmintás U-próba részleges MATLAB®-beli implementálása

```
% ___________
% Description
% ___________
% The function performs a two-sample $U$-test of the null hypothesis $H_0 : \mu_1 = \mu_2$ that the
% data in the two independent samples $\mathbf{X} = \{X_i\}_{i=1}^m$ and $\mathbf{Y} = \{Y_j\}_{j=1}^n$, of random variables
% $X \sim \mathcal{N}(\mu_1, \sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2)$, come from distributions with equal theoretical means.
%
% The test is performed against the alternative hypothesis specified by the input
% parameter tail.
%
% _____
% Input
% _____
% $\mathbf{X} = \{X_i\}_{i=1}^m$ - an independent and identically distributed sample of the normal
%               distribution $\mathcal{N}(\mu_1, \sigma_1)$, where only the theoretical mean value $\mu_1 \in \mathbb{R}$ is
%               unknown
% $\mathbf{Y} = \{Y_j\}_{j=1}^n$ - an independent and identically distributed sample of the normal
%               distribution $\mathcal{N}(\mu_2, \sigma_2)$, where only the theoretical mean value $\mu_2 \in \mathbb{R}$ is
%               unknown
% sigma_1     - corresponds the known value of theoretical standard deviation $\sigma_1 > 0$
% sigma_2     - denotes the known value of theoretical standard deviation $\sigma_2 > 0$
% alpha       - represents the significance level $\alpha \in (0, 1)$
% tail        - a parameter that can be set either by one of the strings 'both',
%               'right', 'left', or by using one of the numbers 0, 1, -1 (it determines
%               the type of the alternative hypothesis)
%
```

Implementálás – II


```
% ____________
% Output
% ____________
% ci_u        — confidence interval for the random variable
%
```

$$U_{m,n} = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim \mathcal{N}(0,1)$$

```
%
% ci_delta    — confidence interval for the difference δ = μ_1 − μ_2 of the theoretical
%               mean values
% u_value     — value of the test, assuming that the null hypothesis H_0 : μ_1 = μ_2 is true
% p_value     — probability of observing the given result, or one more extreme, by
%               chance if the null hypothesis H_0 : μ_1 = μ_2 is true (small p—values
%               cast doubt on the validity of the null hypothesis)
% H           — the code of the accepted hypothesis (0 = null hypothesis,
%                1 = alternative hypothesis defined by the input parameter tail)
%
function [ci_u, ci_delta, u_value, p_value, H] = UTest2D(...
    X, Y, sigma_1, sigma_2, alpha, tail)

% check the validity of input parameters!
...

% get the size of the given samples {X_i}_{i=1}^m and {Y_j}_{j=1}^n
m = length(X);
n = length(Y);

```

Implementálás – III

Kétmintás U-próba az ismeretlen elméleti várható értékek összehasonlítására

```
% observe that $\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$ is a constant
s = sqrt(sigma_1^2 / m + sigma_2^2 / n);

% calculate the difference $\overline{X} - \overline{Y} = \frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{n}\sum_{j=1}^{n} Y_j$ of sample means
X_Y_ = mean(X) - mean(Y);

% calculate the u-value $U_{m,n}^0 = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$
u_value = X_Y_ / s;

% calculate the confidence intervals
switch (tail)
    case {'both', 0}
        % calculate the inverse function value $u_{1-\frac{\alpha}{2}} = F_{\mathcal{N}(0,1)}^{-1}\left(1 - \frac{\alpha}{2}\right)$
        u = norminv(1 - alpha / 2, 0, 1);

        % store the end points of the confidence interval $\left(-u_{1-\frac{\alpha}{2}}, +u_{1-\frac{\alpha}{2}}\right)$
        ci_u(1) = -u;
        ci_u(2) =  u;

        % store the end points of the confidence interval
        % $\left(\overline{X} - \overline{Y} - \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \cdot u_{1-\frac{\alpha}{2}}, \overline{X} - \overline{Y} + \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \cdot u_{1-\frac{\alpha}{2}}\right)$
```

Implementálás – IV

Kétmintás U-próba az ismeretlen elméleti várható értékek összehasonlítására

```
        ci_delta(1) = X_Y_ - s * u;
        ci_delta(2) = X_Y_ + s * u;

        % calculate the p-value p = 2F_{N(0,1)}(-|U^0_{m,n}|)
        p_value = 2.0 * normcdf(-abs(u_value), 0, 1);

    case {'right', 1}
        ...
    case {'left', -1}
        ...
end

% make your decision based on confidence intervals, and note that, the
% condition (0 > ci_delta(1) && 0 < ci_delta(2)) would also be correct!
H = ~(u_value > ci_u(1) && u_value < ci_u(2));

% do you have any doubt? -- if yes, then apply the corresponding significance test!
if (p_value < alpha)
    disp('Warning: small p-value cast doubt on the validity of the null-hypothesis!');
end
```

1. feladat

- Egy gyógynövényekből készült étrend-kiegészítő italport gyártó és forgalmazó cégnél ellenőrzést tartanak, amely során lemérik néhány a gyártósoron levő automata adagológép által megtöltött italporos tasak súlyát. A mérések grammban kifejezett eredményét az

$$X = [197.21, 200.92, 200.53, 200.11, 197.65, 202.0, 200.60, 199.53,\\ 200.10, 199.55, 196.81, 199.58, 198.56, 198.25, 197.90, 199.02,\\ 196.45, 201.70]$$

  tömb tartalmazza.
- Tudva, hogy a kapott adatok normális eloszlást követnek, ha mintavétel hibája $\sigma = 1.65\ g$ döntsétek el 94%-os valószínűség mellett, hogy az adagológép átlagosan több italport adagol-e a tasakokba, mint az előírt 200 grammos érték!

2. feladat

- Egy kozmetikai termékeket gyártó üzemben az éjszakai arcregeneráló szérumot két különböző automata géppel töltik 30 milliliteres kiszerelésű üvegcsékbe. A havi rutinellenőrzés során az automata gépek működéséért felelős munkás leméri mindkét gép esetén néhány véletlenszerűen kiválasztott üvegcse szérumtartalmát. A mérések milliliterben kifejezett eredményét az

$$X = [28.4, 27.0, 30.8, 27.8, 30.1, 31.2, 30.6, 31.1, 31.8, 27.4, 30.5, 28.3, 27.6, 30.7, 30.0],$$

$$Y = [29.9, 31.9, 31.2, 30.7, 32.6, 32.3, 30.5, 28.6, 28.6, 32.5, 27.6, 29.9, 28.9, 31.1, 31.8, 30.6, 29.9]$$

  tömbök tartalmazzák.
- Tudva, hogy a üvegcsékben levő éjszakai arcregeneráló szérum mennyisége normális eloszlást követ, valamint a gépek hibája rendre $\sigma_1 = 1.6$ és $\sigma_2 = 1.4$ milliliter, döntsétek el 91%-os valószínűség mellett, hogy a második töltőgép átlagosan több arcregeneráló szérumot tölt-e az üvegcsékbe, mint az első gép!