Várható értékre és szórásra vonatkozó próbák

2. 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)

10. labor / 2023. december 4–8.

- A 9. labor anyagában bemutatott egy-, illetve kétmintás U-próbák alkalmazása során azt feltételeztük, hogy a független mintavételezések olyan normális eloszlásokból származnak, amelyek elméleti várható értékét nem ismerjük, az elméleti szórásukat viszont adottnak tekintettük.
- Ebben a pontban azt feltételezzük, hogy a vizsgált statisztikai sokaság szintén normális eloszlású, de annak sem az elméleti várható értékét, sem a szórását nem ismerjük.
- Ezért célunk egy olyan valószínűségi változó bevezetése, amely segítségével ilyen esetekben is el tudjuk dönteni az elméleti várható értékre vonatkozó nullhipotézis helyességét megbízhatósági intervallumok szerkesztésével, vagy szignifikanciapróba használatával.

1. Tétel (Az egymintás T-próba valószínűségi változója)

Tekintsük az $X \sim \mathcal{N}(\mu, \sigma)$ valószínűségi változónak az n-elemű független $\{X_j\}_{j=1}^n$ mintavételét ($n \geq 2$), ahol $\mu \in \mathbb{R}$ és $\sigma > 0$! Ekkor a

$$T_n = \frac{\overline{X} - \mu}{\frac{\overline{\sigma}}{\sqrt{n}}} = \frac{\overline{X} - \mu}{\sqrt{\frac{\overline{\mu}_2^c}{n-1}}} \tag{1}$$

valószínűségi változó Student-eloszlású $n-1$ szabadságfokkal, ahol

$$\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

továbbra is az adott mintavétel becsült várható értékét jelöli, továbbá

$$\overline{\mu}_2^c = \frac{1}{n} \sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2$$

az úgynevezett mintavételi (vagy becsült) másodrendű centrált momentumot,

$$\overline{\sigma} = \sqrt{\frac{n}{n-1} \cdot \overline{\mu}_2^c} = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2}$$

pedig a korrigált mintavételi (becsült) szórást jelöli.

Bizonyítás

A bizonyítás megtalálható a nyomtatott jegyzet 182., illetve a digitális változat 189. oldalán!

- Az 1. tétel fényében már tudunk válaszolni az alábbi feladatban megfogalmazott kérdésekre. A megoldás menetéből pedig új statisztikai módszer születik, amelyet egymintás T-próbának nevezünk.

1. Feladat (Egymintás T-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 sem a $\mu \in \mathbb{R}$ elméleti várható értéket, sem a $\sigma > 0$ elméleti szórást nem ismerjük! 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 elméleti μ várható értékre. Ugyanakkor szerkesszünk mindhárom esetben megbízhatósági intervallumot is az ismeretlen μ paraméterre, határozzuk meg az egyes esetekhez tartozó szignifikanciapróbák p-értékét, továbbá implementáljuk is az így adódó statisztikai módszert!

Megoldás

- A feladat megoldását az (1)-es valószínűségi változó használatára vezetjük vissza.
- A 2. labor anyagában már felhívtuk a figyelmet az $\mathcal{N}(0,1)$- és az $\mathcal{S}(n)$-eloszlások sűrűségfüggvényeinek hasonlóságára, például mindkét sűrűségfüggvény szimmetrikus a függőleges tengelyre nézve, valamint az $\mathbb{R}$ tartományon el nem tűnő értékeket vesznek fel.
- Emiatt minden olyan képlettel és ábrával párhuzamot vonhatunk, amiket az egymintás U-próba esetén kaptunk.
- Néhány fontos dolgot viszont figyelembe kell vennünk!
    - Mivel az elméleti $\sigma > 0$ szórást nem ismerjük, az $U_n \sim \mathcal{N}(0,1)$ valószínűségi változó helyett a $T_n \sim \mathcal{S}(n-1)$ változóval kell dolgoznunk, mert az tartalmazza a mintavételt jellemző $\overline{\sigma}$ becsült korrigált szórást.
    - Ezért az $F_{\mathcal{N}(0,1)}$ eloszlásfüggvény és annak $F_{\mathcal{N}(0,1)}^{-1}$ inverze helyett értelemszerűen az $F_{\mathcal{S}(n-1)}$ eloszlásfüggvényt, illetve ennek $F_{\mathcal{S}(n-1)}^{-1}$ inverzét kell használnunk.

Megoldás – folytatás

- Ezért a két-, a jobb és a bal oldali próbák esetén az $1-\frac{\alpha}{2}$, az $1-\alpha$, illetve az α valószínűségekhez tartozó kvantiliseket rendre a

$$\begin{aligned} t_{1-\frac{\alpha}{2}} &= F^{-1}_{\mathcal{S}(n-1)}\left(1-\frac{\alpha}{2}\right), \\ t_{1-\alpha} &= F^{-1}_{\mathcal{S}(n-1)}(1-\alpha), \\ t_{\alpha} &= F^{-1}_{\mathcal{S}(n-1)}(\alpha) \end{aligned} \tag{2}$$

  alakban határozhatjuk meg.

- Mindhárom alternatív hipotézis esetén, az igaznak feltételezett $H_0 : \mu = \mu_0$ nullhipotézis a

$$T_n^0 = \frac{\overline{X}-\mu_0}{\frac{\overline{\sigma}}{\sqrt{n}}}$$

  közös értéket adja, amit a próba t-értékének nevezünk.

Megoldás – folytatás

- A két-, a jobb és a bal oldali próbák esetén rendre az alábbi megbízhatósági intervallumokat kapjuk a T_n valószínűségi változóra és az ismeretlen μ várható értékre:
  - $T_n \in \left(t_{\min}^{k}, t_{\max}^{k}\right) = \left(-t_{1-\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}\right)$,
    $\mu \in \left(\mu_{\min}^{k}, \mu_{\max}^{k}\right) = \left(\overline{X} - \frac{\overline{\sigma}}{\sqrt{n}} t_{1-\frac{\alpha}{2}}, \overline{X} + \frac{\overline{\sigma}}{\sqrt{n}} t_{1-\frac{\alpha}{2}}\right)$;
  - $T_n \in \left(t_{\min}^{j}, t_{\max}^{j}\right) = (-\infty, t_{1-\alpha})$,
    $\mu \in \left(\mu_{\min}^{j}, \mu_{\max}^{j}\right) = \left(\overline{X} - \frac{\overline{\sigma}}{\sqrt{n}} t_{1-\alpha}, +\infty\right)$;
  - $T_n \in \left(t_{\min}^{b}, t_{\max}^{b}\right) = (t_\alpha, +\infty)$, $\mu \in \left(\mu_{\min}^{b}, \mu_{\max}^{b}\right) = \left(-\infty, \overline{X} - \frac{\overline{\sigma}}{\sqrt{n}} t_\alpha\right)$;
- Ha a $H_0 : \mu = \mu_0$ nullhipotézishez rendre a két-, a jobb és a bal oldali alternatív hipotéziseket társítjuk, akkor a megfelelő szignifikanciapróbák p-értékeire rendre az alábbi kifejezéseket kapjuk:
  - $p_k = 2P\left(T_n > \left|T_n^0\right| \mid H_0\right) = 2F_{\mathcal{S}(n-1)}\left(-\left|T_n^0\right|\right)$;
  - $p_j = P\left(T_n > T_n^0 \mid H_0\right) = 1 - F_{\mathcal{S}(n-1)}\left(T_n^0\right)$;
  - $p_b = P\left(T_n < T_n^0 \mid H_0\right) = F_{\mathcal{S}(n-1)}\left(T_n^0\right)$.

Megoldás – folytatás

- Jelölje a továbbiakban $a \in \{k, j, b\}$ az alternatív hipotézis típusát, valamint

$$\mathrm{op}_a = \begin{cases} \neq, & a = k, \\ >, & a = j, \\ <, & a = b \end{cases}$$

az egyes alternatív hipotézisekben megjelenő logikai operátorokat!

- Ekkor minden $a \in \{k, j, b\}$ oldalú próba esetén a szokásos módon dönthetünk: amennyiben $T_n^0 \in \left(t_{\min}^a, t_{\max}^a\right)$, vagy $\mu_0 \in \left(\mu_{\min}^a, \mu_{\max}^a\right)$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, máskülönben a $H_a : \mu \,\mathrm{op}_a\, \mu_0$ alternatív hipotézist fogadjuk el.
- Hasonlóképpen, az $a \in \{k, j, b\}$ odalú szignifikanciapróba esetén, ha $p_a > \alpha$, akkor a $H_0 : \mu = \mu_0$ nullhipotézist, ellenben a $H_a : \mu \,\mathrm{op}_a\, \mu_0$ alternatív hipotézist fogadjuk el.
- Az egymintás T-próbát az alábbi MATLAB®-ban írt forráskódban részlegesen implementáltuk az egymintás U-próba kódrészletéhez hasonlóan.
  - Figyeljünk fel arra, hogy az eddig használt normális eloszlás eloszlásfüggvényét és annak inverzét közelítő normcdf, illetve norminv függvények helyett most a Student-féle eloszlásfüggvényt és az ennek inverzét approximáló tcdf, illetve tinv függvényeket hívtuk meg a (2)-es kvantilisek kiértékelése végett!
  - Vegyük észre azt is, hogy a TTest függvény paraméterlistája már nem feltételezi az elméleti σ szórást ismertnek!

Implementálás – I

Egymintás T-próba az ismeretlen elméleti várható értékre

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

```
% ----------------------
% Description
% ----------------------
% The function performs a one-sample $T$-test of the null hypothesis $H_0 : \mu = \mu_0$ that the
% data in the vector $\mathbf{X} = \{X_j\}_{j=1}^{n}$ comes from an $\mathcal{N}(\mu, \sigma)$ distribution, where the
% theoretical mean value $\mu \in \mathbb{R}$ and the theoretical standard deviation $\sigma > 0$
% are unknown parameters.
%
% The test is performed against the alternative hypothesis specified by the input
% parameter tail.
%
% ----------
% Input
% ----------
% $\mathbf{X} = \{X_j\}_{j=1}^{n}$ - an independent and identically distributed sample of the normal distribu-
%                  tion $\mathcal{N}(\mu, \sigma)$, where the theoretical mean value $\mu \in \mathbb{R}$ and the theoretical
%                  standard deviation $\sigma > 0$ are unknown parameters
% mu_0          - denotes the guess $\mu_0$ of the user for the unknown theoretical mean
%                  value $\mu$
% 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
% ------------
```

Implementálás – II

Egymintás T-próba az ismeretlen elméleti várható értékre

```
% ci_t        - confidence interval for the random variable $T_n = \frac{\overline{X} - \mu}{\frac{\overline{\sigma}}{\sqrt{n}}} \sim \mathcal{S}(n-1)$
% ci_mu       - confidence interval for theoretical mean value $\mu$
% t_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_t, ci_mu, t_value, p_value, H] = TTest(X, mu_0, alpha, tail)

% check the validity of input parameters!
...

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

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

% calculate the sample standard deviation $\overline{\sigma} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2}$
sigma_ = sqrt(sum((X - X_).^2) / (n-1)); % or, equivalently, sigma_ = sqrt(var(X));

```

Implementálás – III

Egymintás T-próba az ismeretlen elméleti várható értékre

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

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

% calculate the confidence intervals
switch (tail)
    case {'both', 0}
        % calculate the inverse function value $t_{1-\frac{\alpha}{2}} = F_{\mathcal{S}(n-1)}^{-1}\left(1 - \frac{\alpha}{2}\right)$
        t = tinv(1 - alpha/2, n-1);

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

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

        % calculate the p-value $p = 2F_{\mathcal{S}(n-1)}\left(-\left|T_n^0\right|\right)$
        p_value = 2.0 * tcdf(-abs(t_value), n-1);
```

Implementálás – IV

Egymintás T-próba az ismeretlen elméleti várható értékre

```
    case {'right', 1}
        ...

    case {'left', -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 = ~(t_value > ci_t(1) && t_value < ci_t(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
```

- A kétmintás U-próbához hasonlóan ebben a pontban is két független, nem feltétlenül azonos normális eloszlású valószínűségi változó ismeretlen elméleti várható értékét hasonlítjuk össze azzal a különbséggel, hogy az ismeretlen elméleti várható értékek mellett az adott valószínűségi változók elméleti szórását sem ismerjük, vagyis az összehasonlítás érdekében a változóknak csak egy-egy független mintavételét tekintjük adottnak.
- A hamarosan ismertetendő kétmintás T-próbát különböző módon kell majd végrehajtani aszerint, hogy az elméleti szórások ismeretlenek, de megegyeznek, vagy az elméleti szórások ismeretlenek, viszont értékeik különbözőek. Éppen ezért két új valószínűségi változót is bevezetünk az alábbi tételekben. (Kiegészítésként lásd a nyomtatott jegyzet 186–187., illetve az elektronikus változat 193–194. oldalait!)

2. Tétel (A kétmintás T-próba valószínűségi változója, ha az elméleti szórások megegyeznek)

Tekintsük az $X \sim \mathcal{N}(\mu_1, \sigma)$ és $Y \sim \mathcal{N}(\mu_2, \sigma)$ 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 $\sigma > 0$! Ekkor a

$$T_{m,n} = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{(m-1)\,\overline{\sigma}_1^2 + (n-1)\,\overline{\sigma}_2^2}} \cdot \sqrt{\frac{m+n-2}{\frac{1}{m} + \frac{1}{n}}} \tag{3}$$

valószínűségi változó $\mathcal{S}(m+n-2)$-eloszlást követ, ahol

$$\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i,\ \overline{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

a mintavételi becsült várható értékeket,

$$\overline{\sigma}_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(X_i - \overline{X}\right)^2,\ \overline{\sigma}_2^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(Y_j - \overline{Y}\right)^2$$

pedig a becsült korrigált szórásnégyzeteket jelölik.

3. Tétel (A kétmintás T-próba valószínűségi változója, ha az elméleti szórások különbözőek)

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 $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 > 0$, továbbá $\sigma_1 \neq \sigma_2$! Ekkor a

$$T_{m,n} = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\overline{\sigma}_1^2}{m} + \frac{\overline{\sigma}_2^2}{n}}} \tag{4}$$

valószínűségi változó $\mathcal{S}(\eta)$-eloszlást követ, ahol az η szabadsáfokot (alakparamétert) súlyozással állíthatjuk elő az

$$\frac{1}{\eta} = \frac{w^2}{m-1} + \frac{(1-w)^2}{n-1},$$

$$w = \frac{\frac{\overline{\sigma}_1^2}{m}}{\frac{\overline{\sigma}_1^2}{m} + \frac{\overline{\sigma}_2^2}{n}}$$

képletek segítségével.

2. Feladat (Kétmintás T-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 sem az elméleti $\mu_1, \mu_2 \in \mathbb{R}$ várható értékeket, sem az elméleti $\sigma_1, \sigma_2 > 0$ szórásokat nem ismerjük! Az $\alpha \in (0,1)$ szignifikanciaszint mellett döntsük el a

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

nullhipotézis helyességét a

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

lehetséges alternatív hipotézisekkel szemben! Ugyanakkor szerkesszünk mindhárom esetben megbízhatósági intervallumot az ismeretlen elméleti várható értékek $\mu_1 - \mu_2$ különbségére, határozzuk meg az egyes esetekhez tartozó szignifikanciapróbák p-értékét, továbbá implementáljuk is ezt a statisztikai módszert!

Megoldás – folytatás

- A feladatot csak vázlatosan oldjuk meg, mert a megoldás mikéntje teljesen hasonló az egymintás T-próba módszeréhez, amelyet az 1. feladatban ismertettünk. A részletes számítások elvégzését és a 2. kódrészlet befejezését az olvasóra bízzuk!
- Ha az $X \sim \mathcal{N}(\mu_1, \sigma_1)$ és $Y \sim \mathcal{N}(\mu_2, \sigma_2)$ teljesen ismeretlen paraméterezésű valószínűségi változók szórásairól a $\sigma_1 = \sigma_2 = \sigma$ megegyezést feltételezzük, akkor, az egymintás T-próbához képest, a T_n valószínűségi változó helyett a (3)-as, máskülönben a (4)-es kifejezésben adott $T_{m,n}$ valószínűségi változó használatára vezethetjük vissza a feladatot.
- A továbbiakban – a szórásokkal kapcsolatos esetszétválasztásnak megfelelően – felsoroljuk mindhárom alternatív hipotézishez tartozó konfidenciaintervallumokat, valamint a megfelelő szignifikanciapróbák p-értékét is feltüntetjük.
- Vezessük be az

$$s = \begin{cases} \sqrt{\left((m-1)\overline{\sigma}_1^2 + (n-1)\overline{\sigma}_2^2\right) \cdot \dfrac{m+n}{mn(m+n-2)}}, & \sigma_1 = \sigma_2 = \sigma, \\[2ex] \sqrt{\dfrac{\overline{\sigma}_1^2}{m} + \dfrac{\overline{\sigma}_2^2}{n}}, & \sigma_1 \neq \sigma_2 \end{cases}$$

konstanst.

Megoldás – folytatás

- Ekkor $H_0 : \mu_1 - \mu_2 = 0$ nullhipotézist igaznak tartva, a kétmintás T-próba t-értékére a

$$T_{m,n}^0 = \frac{\left(\overline{X} - \overline{Y}\right) - 0}{s} = \frac{\overline{X} - \overline{Y}}{s}$$

kifejezést kapjuk.

- Jelölje

$$\eta = \begin{cases} m + n - 2, & \sigma_1 = \sigma_2 = \sigma, \\ \dfrac{(m-1)(n-1)}{(m-1)(1-w)^2 + (n-1)w^2}, & \sigma_1 \neq \sigma_2 \end{cases}$$

az esetszétválasztásnak megfelelő szabadságfokot (alakparamétert), ahol

$$w = \frac{n\overline{\sigma}_1^2}{n\overline{\sigma}_1^2 + m\overline{\sigma}_2^2}!$$

Megoldás – folytatás

- Ekkor a

$$T_{m,n} = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{s}$$

valószínűségi változó $\mathcal{S}(\eta)$-eloszlást követ. Ezért, amikor a $H_0 : \mu_1 = \mu_2$ nullhipotézhez a két-, a jobb, és a bal odali alternatív hipotéziseket társítjuk, akkor az $1 - \frac{\alpha}{2}$, $1 - \alpha$, illetve α valószínűségekhez tartozó kvantilisekre rendre a

$$\begin{aligned} t_{1-\frac{\alpha}{2}} &= F^{-1}_{\mathcal{S}(\eta)}\left(1 - \frac{\alpha}{2}\right), \\ t_{1-\alpha} &= F^{-1}_{\mathcal{S}(\eta)}(1 - \alpha), \\ t_{\alpha} &= F^{-1}_{\mathcal{S}(\eta)}(\alpha) \end{aligned} \tag{5}$$

inverz értékeket kapjuk.

Megoldás – folytatás

- A két-, a jobb és a bal oldali próbák esetén rendre az alábbi megbízhatósági intervallumokat kapjuk a $T_{m,n}$ valószínűségi változóra és az ismeretlen várható értékek $\delta = \mu_1 - \mu_2$ különbségére:
  - $T_{m,n} \in (t_{\min}^k, t_{\max}^k) = \left(-t_{1-\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}\right)$,
    $\delta \in (\delta_{\min}^k, \delta_{\max}^k) = \left(\overline{X} - \overline{Y} - s \cdot t_{1-\frac{\alpha}{2}}, \overline{X} - \overline{Y} + s \cdot t_{1-\frac{\alpha}{2}}\right)$;
  - $T_{m,n} \in \left(t_{\min}^j, t_{\max}^j\right) = (-\infty, t_{1-\alpha})$,
    $\delta \in \left(\delta_{\min}^j, \delta_{\max}^j\right) = \left(\overline{X} - \overline{Y} - s \cdot t_{1-\alpha}, +\infty\right)$;
  - $T_{m,n} \in (t_{\min}^b, t_{\max}^b) = (t_\alpha, +\infty)$, $\delta \in (\delta_{\min}^b, \delta_{\max}^b) = \left(-\infty, \overline{X} - \overline{Y} - s \cdot t_\alpha\right)$.
- Ha a $H_0 : \mu_1 = \mu_2$ nullhipotézishez rendre a két-, a jobb és a bal oldali alternatív hipotéziseket társítjuk, akkor a megfelelő oldalú szignifikanciapróbák p-értékeire rendre az alábbi valószínűségeket kapjuk:
  - $p_k = 2P\left(T_{m,n} > \left|T_{m,n}^0\right| \,\middle|\, H_0\right) = 2F_{\mathcal{S}(\eta)}\left(-\left|T_{m,n}^0\right|\right)$;
  - $p_j = P\left(T_{m,n} > T_{m,n}^0 \,\middle|\, H_0\right) = 1 - F_{\mathcal{S}(\eta)}\left(T_{m,n}^0\right)$;
  - $p_b = P\left(T_{m,n} < T_{m,n}^0 \,\middle|\, H_0\right) = F_{\mathcal{S}(\eta)}\left(T_{m,n}^0\right)$.

Megoldás – folytatás

- Minden $a \in \{k, j, b\}$ oldalú kétmintás T-próba esetén a szokásos módon dönthetünk: amennyiben $T_{m,n}^{0} \in \left(t_{\min}^{a}, t_{\max}^{a}\right)$, vagy $0 \in \left(\delta_{\min}^{a}, \delta_{\max}^{a}\right)$, akkor a $H_0 : \mu_1 = \mu_2$ nullhipotézist, máskülönben a $H_a : \mu_1 \operatorname{op}_a \mu_2$ alternatív hipotézist fogadjuk el.
- Hasonlóképpen, az $a \in \{k, j, b\}$ odalú szignifikanciapróba esetén, ha $p_a > \alpha$, akkor a $H_0 : \mu_1 = \mu_2$ nullhipotézist, ellenben a $H_a : \mu_1 \operatorname{op}_a \mu_2$ alternatív hipotézist fogadjuk el.

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

```
% ----------------
% Description
% ----------------
% The function performs a two-sample $T$-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
% with unknown (and not necessarily equal) theoretical standard deviations.
%
% 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 both parameters $\mu_1 \in \mathbb{R}$ and $\sigma_1 > 0$
%                 are 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 both parameters $\mu_2 \in \mathbb{R}$ and $\sigma_2 > 0$
%                 are unknown
% equal_std   - a boolean variable which specifies whether the unknown standard
%                 deviations $\sigma_1$ and $\sigma_2$ are equal
% 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_t        — confidence interval for the random variable
%
%              $T_{m,n} = \begin{cases} \frac{\left(\overline{X}-\overline{Y}\right)-(\mu_1-\mu_2)}{\sqrt{(m-1)\overline{\sigma}_1^2+(n-1)\overline{\sigma}_2^2}} \cdot \sqrt{\frac{m+n-2}{\frac{1}{m}+\frac{1}{n}}}, & \sigma_1 = \sigma_2, \\ \frac{\left(\overline{X}-\overline{Y}\right)-(\mu_1-\mu_2)}{\sqrt{\frac{\overline{\sigma}_1^2}{m}+\frac{\overline{\sigma}_2^2}{n}}}, & \sigma_1 \neq \sigma_2 \end{cases}$
%
% ci_delta    — confidence interval for the difference $\delta = \mu_1 - \mu_2$ of the theoretical
%               mean values
% t_value     — value of the test, assuming that the null hypothesis $H_0 : \mu_1 = \mu_2$ is true
% p_value     — probability of observing the given result, or one more extreme, by chance
%               if the null hypothesis $H_0 : \mu_1 = \mu_2$ 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_t, ci_delta, t_value, p_value, H] = TTest2D(X, Y, equal_std, alpha, tail)

% check the validity of input parameters!
...

% get the size of the given samples $\{X_i\}_{i=1}^m$ and $\{Y_j\}_{j=1}^n$
```

Implementálás – III

Egymintás T-próba az ismeretlen elméleti várható értékre

```
m = length(X);
n = length(Y);

% calculate the constant s =
```

$$s = \begin{cases} \sqrt{\left((m-1)\,\overline{\sigma}_1^2 + (n-1)\,\overline{\sigma}_2^2\right) \cdot \dfrac{m+n}{mn\,(m+n-2)}}, & \sigma_1 = \sigma_2, \\[2ex] \sqrt{\dfrac{\overline{\sigma}_1^2}{m} + \dfrac{\overline{\sigma}_2^2}{n}}, & \sigma_1 \neq \sigma_2 \end{cases}$$

```
sigma_1_2_ = var(X);
sigma_2_2_ = var(Y);

if (equal_std)
    s = sqrt(((m-1) * sigma_1_2_ + (n-1) * sigma_2_2_) * (m + n) / m / n / (m + n - 2));
else
    s = sqrt(sigma_1_2_ / m + sigma_2_2_ / n);
end

% determine the degree of freedom η =
```

$$\eta = \begin{cases} m+n-2, & \sigma_1 = \sigma_2 = \sigma, \\[2ex] \dfrac{(m-1)\,(n-1)}{(m-1)\,(1-w)^2 + (n-1)\,w^2}, & \sigma_1 \neq \sigma_2, \end{cases}$$

```
% where w =
```

$$w = \frac{n\overline{\sigma}_1^2}{n\overline{\sigma}_1^2 + m\overline{\sigma}_2^2}$$

```
if (equal_std)
    eta = m + n - 2;
else
    w = n * sigma_1_2_ / (n * sigma_1_2_ + m * sigma_2_2_);
    eta = (m - 1) * (n - 1) / ((m - 1) * (1 - w)^2 + (n - 1) * w^2);
```

Implementálás – IV

Egymintás T-próba az ismeretlen elméleti várható értékre

```
end

% 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 t-value $T^0_{m,n} = \frac{\overline{X} - \overline{Y}}{s}$
t_value = X_Y_ / s;

% calculate the confidence intervals
switch (tail)
    case {'both', 0}
        % calculate the inverse function value $t_{1-\frac{\alpha}{2}} = F^{-1}_{\mathcal{S}(\eta)}\left(1 - \frac{\alpha}{2}\right)$
        t = tinv(1 - alpha / 2, eta);

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

        % store the end points of the confidence interval
        % $\left(\overline{X} - \overline{Y} - s \cdot t_{1-\frac{\alpha}{2}}, \overline{X} - \overline{Y} + s \cdot t_{1-\frac{\alpha}{2}}\right)$
        ci_delta(1) = X_Y_ - s * t;
        ci_delta(2) = X_Y_ + s * t;

```

```
        % calculate the p-value p = 2F_{S(eta)}(-|T^0_{m,n}|)
        p_value = 2.0 * tcdf(-abs(t_value), eta);

    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 = ~(t_value > ci_t(1) && t_value < ci_t(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 kézműves aszalt gyümölcsöket forgalmazó kisvállalatnál ellenőrzést tartanak, amely során lemérik néhány találomra kiválasztott fél kilogrammos tasak aszalt gyümölcs-tartalmát. A mérések grammban kifejezett értékét az

$$X = [501.24, 497.36, 497.54, 498.14, 493.23, 499.76, 498.44, 500.73, 501.32, 498.01, 499.31, 500.75, 498.26, 500.15, 490.18]$$

  tömb tartalmazza.
- Tudva, hogy a tasakokban levő aszalt gyümölcsök súlya normális eloszlást követ, döntsétek el 95%-os valószínűség mellett, hogy a tasakokban átlagosan kevesebb aszalt gyümölcs van-e, mint fél kilogramm!

2. feladat

Oldjátok meg a 9. labor feladatait úgy, hogy a szórás(oka)t egyik feladat esetén sem tekintitek ismert paraméter(ek)nek!

3. feladat

- Egy acélmegmunkálással foglalkozó üzem vezetője két új automata darabológépet vásárol, amelyeket üzembeállítás előtt egy tesztkör keretén belül próbálnak ki. A próbakör során a gépeknek egy méter hosszúságú acélszalagokat kell vágniuk. A gépek által darabolt acélszalagok centiméterben mért hosszát az

$$X = [104, 100, 102, 98, 103, 99, 100, 97, 107, 102, 103, 100, 98, 102, 100],$$
$$Y = [98, 97, 101, 100, 99, 100, 97, 102, 101, 98, 99, 100, 102, 99, 101, 96, 99]$$

  tömbök tartalmazzák.
- Tudva, hogy a gépek által vágott acélszalagok hossza normális eloszlást követ, döntsétek el 92%-os valószínűség mellett, hogy a második automata gép átlagosan rövidebb acélszalagokat vág-e, mint az első gép!