Hipotézisvizsgálat

Feladat: Vásárlói panaszok érkeznek, hogy egy élelmiszerboltban az 1 kg-os feliratú cukros zacskóban valójában kevesebb van. Szeretnénk korrekt módon kivizsgálni az ügyet. Kiszállunk az üzletbe, megnézünk n véletlenszerű zacskót: $X_1, X_2, \ldots, X_n$ minta. Legyen $n = 25$ és azt találjuk, hogy átlaguk 0,98 kg. Tudjuk, hogy a cukros zacskók súlya normális eloszlást követ és a szórásuk: $\sigma_0 = 0,05$. Mit tegyünk?

Megoldási stratégia: Az eltérést okozhatja a véletlen is, hiszen az 1 kg várható értékű, normális eloszlású mintaelemek eltérhetnek a várható értéktől. A következképpen gondolkodhatunk: az ártatlanság vélelme alapján feltehetjük, hogy nem csalnak, vagyis a normális eloszlású háttérváltozó várható értéke valóban 1 kg. Szerkesszünk például 95% -os konfidenciaintervallumot a várható értékre a minta alapján!

Amennyiben az 1 kg hipotetikus várható érték benne van ebben az intervallumban, akkor elfogadjuk, hogy 1 kg a várható érték, az eltérést a véletlen ingadozás természetes volta okozta.

Amennyiben az 1 kg hipotetikus várható érték nincs benne ebben az intervallumban, akkor két eset lehetséges:

1. Mivel az esetek 95% -ában a várható érték benne van ebben az intervallumban, a véletlen folytán lehet, hogy mégis bekövetkezett az az 5% valószínűségű esemény, hogy nincs benne.
2. Nem igaz az eredeti elképzelésünk, hogy 1 kg a várható érték.

Nagyon kis okunk van azt hinni, hogy bekövetkezett egy 5% valószínűségű esemény, emiatt az utóbbi mellett voksolunk, hogy nem 1 kg a várható érték. Azaz 95% biztonsággal úgy döntünk, hogy csaltak!

A konkrét feladatban most a konfidenciaintervallumot $1 - \epsilon = 0,95$, tehát $\epsilon = 0,05$ esetén a következőképpen számoljuk:

$$u_{\epsilon/2} = u_{0,025} = 1,96, \quad n = 25, \quad \sigma_0 = 0,05, \quad \overline{x} = 0,98$$

így a konfidenciaintervallum:

$$(\overline{x} - \frac{u_{\epsilon/2}\sigma_0}{\sqrt{n}}, \overline{x} + \frac{u_{\epsilon/2}\sigma_0}{\sqrt{n}}) =$$

$$(0,98 - \frac{1,96 * 0,05}{\sqrt{25}}; 0,98 + \frac{1,96 * 0,05}{\sqrt{25}}) =$$

$$(0,9604; 0,9996)$$

tehát az 1 kg nincs benne a konfidenciaintervallumban, ezért úgy döntünk, hogy nem 1 kg a várható érték, tényleg csalnak!

Nézzük meg általánosan a feladatot! Egy véletlen jelenség kapcsán egy

$$H_0$$

hipotézisről szeretnénk dönteni egy

$$H_1$$

úgynevezett alternatív hipotézissel szemben.

Ehhez választunk egy $0 < \epsilon < 1$ számot és most $1 - \epsilon$ biztonsággal szeretnénk dönteni. Veszünk egy n elemű mintát: $X_1, X_2, \ldots, X_n$. Ezután a következőképpen döntünk:

1. Ebből elkészítjük egy ún. próbastatisztikát:

   $$T(\underline{X}) = T(X_1, X_2, \ldots, X_n)$$

   értéket.
2. Meghatározzuk az $(1 - \epsilon)$-szintű konfidenciaintervallumot (ez lehet egy félegyenes is).

A döntésünk a következő: Ha a $T(X)$ próbastatisztika benne van a konfidenciaintervallumban, akkor elfogadjuk a H_0 hipotézist, ha nincs akkor benne, akkor elutasítjuk, azaz H_1 alternatív hipotézis mellett döntünk.

	H_0 teljesül	H_0 nem teljesül
H_0-t elfogadjuk	jól döntünk	nem jól döntünk (másodfajú hiba)
H_0-t elutasítjuk	nem jól döntünk (elsőfajú hiba)	jól döntünk

Ekkor ha H_0 tényleg teljesül, akkor $1 - \epsilon$ valószínűséggel jól döntünk, elfogadjuk H_0-t. Ezért annak a valószínűsége, hogy H_0 teljesül, de mi mégsem így döntünk ϵ, ezt az esetet elsőfajú hibának mondjuk, azaz most az elsőfajú hiba valószínűsége éppen az ϵ.

A döntésünk során azonban más hibát is véthetünk: lehet, hogy elfogadjuk H_0-t, pedig nem teljesül. Ekkor másodfajú hibáról becsélünk.

Sajnos ennek a valószínűségét nem tudjuk megmondani. A következők teljesülnek erre:

1. Azt mondhatjuk, hogy ez ellentétesen viselkedik az elsőfajú hibával szemben: ha az elsőfajú hibát csökkentjük, akkor cserébe a másodfajú hiba megnő.
2. A mintaelemszám növelésekor a másodfajú hiba valószínűsége csökken.

Statisztikai próbák

Egymintás próba esetén feltesszük, hogy az X háttérváltozóval rendelkezésünkre áll az $X_1, X_2, \ldots, X_n$ minta és az $0 < \epsilon < 1$ szám és az érdekel minket, hogy a várható érték megegyezik-e az általunk gondolt μ_0-val, azaz

$$H_0 : \mu = \mu_0$$

Kétmintás próba esetén az (m_1, σ_1) paraméterekkel rendelkező normális eloszlású X háttérváltozóval kapott $X_1, X_2, \ldots, X_{n_1}$ minta és az (m_2, σ_2) paraméterekkel rendelkező normális eloszlású Y háttérváltozóval kapott $Y_1, Y_2, \ldots, Y_{n_2}$ minta áll rendelkezésre. Ekkor az érdekel minket, hogy a két háttérváltozó várható értéke megegyeik-e, azaz most

$$H_0 : \mu_1 = \mu_2$$

A feladat az, hogy ϵ elsőfajú hibavalószínűséggel döntsünk, ekkor azt mondjuk, hogy $(1 - \epsilon)$-szignifikanciaszinten döntünk.

Ehhez a legfontosabb eseteket külön megnézzük:

u-próba

Az u-próbát akkor alkalmazzuk, ha normális eloszlású v.v. (vagy két v.v.) van a háttérben ez(ek) szórása ismert.

Egymintás u-próba

Tegyük fel, hogy X normális eloszlású v.v. (m, σ_0) paraméterekkel, ahol σ_0 ismert.

Most tehát valamely általunk gondolt μ_0 esetén vizsgáljuk a

$$H_0 : \mu = \mu_0$$

nullhipotézist.

Ekkor a próbastatisztika:

$$u(\underline{X}) = \frac{\overline{x} - \mu_0}{\sigma_0}\sqrt{n}$$

Ekkor az alternatív H_1 hipotézis szerint a következképpen alakulnak a konfidenciaintervallumok:

H_1	konfidenciaintervallum
$\mu \neq \mu_0$	$(-u_{\epsilon/2}, u_{\epsilon/2})$
$\mu > \mu_0$	$(-\infty, u_\epsilon)$
$\mu < \mu_0$	$(-u_\epsilon, \infty)$

ahol

$$\Phi(u_{\epsilon/2}) = 1 - \frac{\epsilon}{2} \quad \Phi(u_\epsilon) = 1 - \epsilon$$

Feladat: Egy 50 kg-os cementeszsákokat gyártó céget meggyanusítanak, hogy ténylegesen kevesebb mint 50 kg az átlagos tömeg. Ehhez egy 25 elemű mintát vesznek a következő adatokkal:

$$49,7 \quad 49,9 \quad 50 \quad 49,8 \quad 50,1 \quad 49,6 \quad 50,1 \quad 49,8 \quad 50$$

$$49,9 \quad 49,8 \quad 50,2 \quad 49,9 \quad 50 \quad 49,7 \quad 49,9 \quad 49,7$$

$$49,8 \quad 49,9 \quad 49,6 \quad 49,9 \quad 50,2 \quad 49,9 \quad 49,9 \quad 50,2$$

Tudjuk, hogy a cég által használt gyártósor szórása $\sigma_0 = 0,3$. Biróság dönt a kérdésben.

1. Az I. fokú bíróság $\epsilon = 0,05$ mellett hozza meg a döntést.
2. Ezután a II. fokú bíróság $\epsilon = 0,01$ mellett hozza meg a döntést.

Milyen döntések születnek?

Ekkor A hipotézisek a következők:

$$H_0 : m = 50 \qquad H_1 : m < 50$$

Az átlagérték:

$$\overline{X} = 49,9,$$

ezért a próbastatisztika:

$$u(X) = \frac{\overline{X} - m}{\sigma_0}\sqrt{n} = \frac{49,9 - 50}{0,3}\sqrt{25} = -1,67$$

Az I. fokú bíróságnál $\epsilon = 0,05$, ezért

$$u_\epsilon = u_{0,05} = 1,65$$

Tehát

$$u(\underline{X}) \notin (-u_\epsilon, +\infty),$$

tehát az I. fokú bíróság elutasítja a H_0 hipotézist.

A másodfokú bíróságnál $\epsilon = 0,01$, ezért

$$u_\epsilon = u_{0,01} = 2,34,$$

tehát

$$u(\underline{X}) \in (-u_\epsilon, \infty),$$

tehát a II. fokú bíróság már elfogadja a H_0 hipotézist.

Kétmintás u-próba

Kétmintás esetben tehát

$$H_0 : \mu_1 = \mu_2$$

Ekkor a próbastatisztika:

$$u(\underline{X}, \underline{Y}) = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Ekkor az alternatív H_1 hipotézis szerint a következőképpen alakulnak a konfidenciaintervallumok:

H_1	konfidenciaintervallum
$\mu_1 \neq \mu_2$	$(-u_{\epsilon/2}, u_{\epsilon/2})$
$\mu_1 > \mu_2$	$(-\infty, u_\epsilon)$
$\mu_1 < \mu_2$	$(-u_\epsilon, \infty)$

Feladat: Az előző feladatban lévő gépsoron tisztítást végeznek, ami során újra beállítják a gépet. A tisztítás előtt 10 elemű mintát vesznek a következ adatokkal:

$$49,8 \quad 50,1 \quad 49,9 \quad 49,8 \quad 50 \quad 49,8 \quad 49,9 \quad 50,1 \quad 49,8 \quad 50,2$$

A tisztítás után újra beállítják, ekkor a következ 8 elemű mintát kapják:

$$49,7 \quad 49,8 \quad 49,6 \quad 50,1 \quad 49,7 \quad 49,8 \quad 49,7 \quad 50,1$$

Elfogadjuk-e, hogy ugyanúgy sikerült a beállítás, azaz a két várható érték megegyezik $\epsilon = 0,05$ elsőfajú hibával?

Legyen a beállítás előtti normális eloszlás várható értéke m_1, míg a beállítás utáni m_2. Ekkor a hipotézisek:

$$H_0 : m_1 = m_2 \qquad H_1 : m_1 \neq m_2$$

A döntéshez először a próbastatisztikát számoljuk ki:

$$\overline{X} = 49,94, \qquad \overline{Y} = 49,8125,$$

ezért

$$u(\underline{X}, \underline{Y}) = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} =$$

$$\frac{49,94 - 49,8125}{\sqrt{\frac{0,3^2}{10} + \frac{0,3^2}{8}}} = 0,83$$

$\epsilon = 0,05$ esetén

$$u_{\epsilon/2} = u_{0,025} = 1,96,$$

ezért

$$u(\underline{X}) \in (-1,96; 1,96),$$

így elfogadjuk, hogy a beállítás jól történt.

t-próba

Az t-próbát akkor alkalmazzuk, ha normális eloszlású v.v. (vagy két v.v.) van a háttérben és ez(ek) szórása nem ismert.

Egymintás t-próba

Tegyük fel, hogy X normális eloszlású v.v. (m, σ) paraméterekkel, ahol σ nem ismert.

Most tehát valamely általunk gondolt μ_0 esetén vizsgáljuk a

$$H_0 : \mu = \mu_0$$

nullhipotézist.

Ekkor a próbastatisztika:

$$t(\underline{X}) = \frac{\overline{x} - \mu_0}{S_n^*}\sqrt{n}$$

Ekkor az alternatív H_1 hipotézis szerint a következképpen alakulnak a konfidenciaintervallumok:

H_1	konfidenciaintervallum
$\mu \neq \mu_0$	$(-t_{\epsilon/2}(n-1), t_{\epsilon/2}(n-1))$
$\mu > \mu_0$	$(-\infty, t_{\epsilon}(n-1))$
$\mu < \mu_0$	$(-t_{\epsilon}(n-1), \infty)$

ahol $t_{\epsilon/2}(n-1)$ és $t_{\epsilon}(n-1)$ az $(n-1)$-szabadságfokú t-eloszás megfelelő értéke.

Kétmintás t-próba

Kétmintás esetben tehát

$$H_0 : \mu_1 = \mu_2$$

Ekkor a próbastatisztika:

$$t(\underline{X}, \underline{Y}) = \frac{\overline{X} - \overline{Y}}{\sqrt{(n_1 - 1)(S_1^*)^2 + (n_2 - 1)(S_2^*)^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}},$$

ahol S_1^* az $X_1, X_2, \ldots, X_{n_1}$ minta korrigált tapasztalati szórása és S_2^* az $Y_1, Y_2, \ldots, Y_{n_2}$ minta korrigált tapasztalati szórása.

Ekkor az alternatív H_1 hipotézis szerint a következképpen alakulnak a konfidenciaintervallumok:

H_1	konfidenciaintervallum
$\mu_1 \neq \mu_2$	$(-t_{\epsilon/2}(n_1 + n_2 - 2), t_{\epsilon/2}(n_1 + n_2 - 2))$
$\mu_1 < \mu_2$	$(-t_{\epsilon}(n_1 + n_2 - 2), \infty)$
$\mu_1 > \mu_2$	$(-\infty, t_{\epsilon}(n_1 + n_2 - 2))$

Feladat: Egy erdei fenyves kapcsán rendelkezésre állnak 1960-ból származó 40 éves fákra vonatkozó adatok, 10 véletlenszerűen kiválasztott fa magassága méterben:

19 27 23 24 22 26 22 29 24 22

Most is megnéznek 10 véletlenszerűen kiválasztott 40 éves fenyt és a következő adatokat kapják:

23 28 22 26 25 18 27 30 26 24

A fenyők magassága normális eloszlást követ. Lehet-e azt vélelmezni, hogy változott a fenyők átlagmagassága az elmúlt 50 évben? Válaszoljuk meg a kérdést $\epsilon = 0,1$ elsőfajú hibával!

Megoldás: Legyen X az 1960-as fenyők magassága, ennek várható értéke legyen m_1, míg jelölje Y a mostani fenyők magasságát, az ő várható értéke m_2

A hipotézisünk a következő:

$$H_0 : m_1 = m_2 \qquad H_1 : m_1 \neq m_2$$

Ennek eldöntéséhez kiszámoljuk a próbastatisztikát:

$$\overline{X} = 23,8, \quad \overline{Y} = 24,9$$

$$(S_1^*)^2 = \frac{(19-23,8)^2 + (27-23,8)^2 + \cdots + (22-23,8)^2}{9} = 8,4$$

$$(S_2^*)^2 = \frac{(23-24,9)^2 + (28-24,9)^2 + \cdots + (24-24,9)^2}{9} = 10,9$$

Így

$$t(\underline{X}, \underline{Y}) = \frac{23,8-24,9}{\sqrt{9*8,4+9*10,9}}\sqrt{\frac{1800}{20}} = -0,79$$

Másrészt a konfidenciaintervallumhoz $\epsilon = 0,05$ mellett

$$t_{\epsilon/2}(n_1 + n_2 - 2) = t_{0,025}(18) = 2,1,$$

ezért

$$t(\underline{X}, \underline{Y}) \in (-t_{0,025}(18), t_{0,025}(18)),$$

tehát elfogadjuk H_0-t, nincs lényeges változás a fenyők átlagos magasságában.

Relatív gyakoriságok összehasonlítása két csoportban

Azt szeretnénk tesztelni, hogy két (nem feltétlenül azonos méretű) csoportban valamely tulajdonság megléte azonos valószínűségű-e, azaz ha p_1 az 1. csoportban a tulajdonság valószínűsége és p_2 a másik csoportban, akkor

$$H_0 : p_1 = p_2 \qquad H_1 : p_1 \neq p_2$$

Megoldás: Veszünk az első csoportból egy n_1 elemű mintát, itt a relatív gyakoriság legyen r_1 és veszünk a második csoportból egy n_2 elemű mintát, itt a relatív gyakoriság legyen r_2.

Ekkor a próbastatisztika:

$$u = \frac{r_1 - r_2}{\sqrt{\frac{r_1(1-r_1)}{n_1} + \frac{r_2(1-r_2)}{n_2}}}$$

és a konfidenciaintervallum:

$$(-u_{\epsilon/2}, u_{\epsilon/2})$$

Feladat: Egy közvéleménykutató cég a pártok támogatásának felmérésekor 1000 embert kérdez meg. A szeptemberi felmérésen az egyik pártra 120 ember szavazott, míg a novemberi felmérésen már 140 ember. Következtethetünk-e ezekből az adatokból arra, hogy nőtt a párt népszerűsége. A válaszát $\epsilon = 0,1$ elsőfajú hiba mellett adja meg!

Megoldás: Jelölje p_1 a szeptemberi és p_2 a novemberi tényleges támogatottságot. A hipotéziseink a következők:

$$H_0 : p_1 = p_2 \qquad H_1 : p_1 \neq p_2$$

Ekkor

$$r_1 = \frac{120}{1000} = 0,12 \quad r_2 = 1401000, \quad n_1 = n_2 = 1000,$$

így a próbastatisztika

$$u = \frac{0,12 - 0,14}{\sqrt{\frac{0,12*0,88}{1000} + \frac{0,14*0,86}{1000}}} = -1,33$$

Másrészt $\epsilon = 0,1$ mellett

$$u_{\epsilon/2} = u_{0,05} = 1,65$$

és így

$$u \in (-u_{\epsilon/2}, u_{\epsilon/2})$$

tehát elfogadjuk H_0-t, nincs ers bizonyítékunk arra, hogy lényeges változás történt.