Egymintás z-próba:

Modell: az Y változó alapsokaságbeli értékei $\mathcal{N}(\mu, \sigma^2)$ ahol μ ismeretlen,

és σ ismert.

Hipotézis: H_0: $\mu = \mu_0$ és az ellenhipotézis H_1: $\mu \neq \mu_0$

Minta: N elemű EVM az Y változóra: $\{y_1, y_2, ..., y_N\}$

Próbastatisztika: $S = \dfrac{y_\bullet - \mu_0}{\sqrt{\frac{\sigma^2}{N}}}$, mely standard normális, ha H_0 igaz.

A fent írt próba **kétoldali.**

Ha az ellenhipotézis H_1: $\mu > \mu_0$ akkor a próba **egyoldali**.

Például 0.05 szigifikanciaszinten a kétoldali esetben H_0 elfogadási tartomány:

$-1.96 \leq S \leq 1.96$ viszont az egyoldali esetben az elfogadási tartomány: $S \leq 1.65$

Kétmintás z-próba:

Modell: az X változó alapsokaságbeli értékei $\mathcal{N}(\mu_X, \sigma_X^2)$

az Y változó alapsokaságbeli értékei $\mathcal{N}(\mu_Y, \sigma_Y^2)$

ahol μ_X és μ_Y ismeretlen, de σ_X és σ_Y ismert.

Hipotézis: H_0: $\mu_X = \mu_Y$ és az ellenhipotézis H_1: $\mu_X \neq \mu_Y$

Minta: N elemű EVM X-re és M elemű EVM Y -ra,

egymástól független két mintavétel.

Próbastatisztika: $S = \dfrac{x_\bullet - y_\bullet}{\sqrt{\frac{\sigma_X^2}{N} + \frac{\sigma_Y^2}{M}}}$, mely standard normális, ha H_0 igaz.

Megjegyezzük, hogy $S =$ a két csoportátlag különbsége osztva a két csoportátlag

különbségének SHjával.

Binomiális próba:

Modell: az Y változó alapsokaságbeli értékei 0 és 1

Hipotézis: H_0: $E\{Y\} = p_0$

Minta: N elemű EVM az Y változóra: $\{y_1, y_2, ..., y_N\}$

Próbastatisztika: $S = y_\bullet$ és ha H_0 igaz, akkor S eloszlása $\dfrac{1}{N} \cdot Binomiális(N, p_0)$

Megjegyezzük, hogy $E\{Y\}=$ az 1 alapsokasági relatív gyakorisága,

$y_\bullet = \dfrac{y_1 + y_2 + ... + y_N}{N} =$ az 1 mintabeli relatív gyakorisága.

Ha a táblázatunk nem tartalmazza a szükséges adatokat, normális eloszlással való közelítést

alkalmazunk, várhatóérték $= p_0$, SH $= \sqrt{\dfrac{p_0 \cdot (1 - p_0)}{N}}$

Egymintás t-próba:

Modell: az Y változó alapsokaságbeli értékei $\mathcal{N}(\mu, \sigma^2)$ ahol μ és σ ismeretlen

Hipotézis: H_0: $\mu = \mu_0$

Minta: N elemű EVM az Y változóra: $\{y_1, y_2, ..., y_N\}$

Próbastatisztika: $S = \dfrac{y_\bullet - \mu_0}{\sqrt{\frac{s_*^2}{N}}}$

ahol $s_*^2 = \dfrac{1}{N-1} \cdot SSQ_Y$, $SSQ_Y = \sum_{n=1}^{N} (y_n - y_\bullet)^2$

S eloszlása $N-1$ szabadsági fokú t , ha H_0 igaz.

Megjegyezzük, hogy s_*^2 eloszlása $\dfrac{1}{N-1} \cdot \sigma^2 \cdot \chi^2_{N-1}$ (akár igaz H_0, akár nem).

Kétmintás t-próba:

Modell: az X változó alapsokaságbeli értékei $\mathcal{N}(\mu_X, \sigma_X^2)$

az Y változó alapsokaságbeli értékei $\mathcal{N}(\mu_Y, \sigma_Y^2)$

ahol μ_X és μ_Y valamint σ_X és σ_Y ismeretlenek, és $\sigma_X = \sigma_Y$

Hipotézis: H_0: $\mu_X = \mu_Y$

Minta: N-EVM X-re és M-EVM Y-ra, egymástól független két mintavétel.

Próbastatisztika: $S = \dfrac{x_\bullet - y_\bullet}{\sqrt{SSQ_X + SSQ_Y}} \cdot \sqrt{\dfrac{NM(N+M-2)}{N+M}}$

ahol $SSQ_X = \sum_{n=1}^{N} (x_n - x_\bullet)^2$ és $SSQ_Y = \sum_{m=1}^{M} (y_m - y_\bullet)^2$

Megjegyezzük, hogy S = a két csoportátlag különbsége osztva a két csoportátlag

különbségének a SHbecslésével. S eloszlása $N+M-2$ szab.fokú t , ha H_0 igaz.

Megjegyezzük, hogy ha a modellben nem tesszük fel, hogy $\sigma_X = \sigma_Y$ teljesül, akkor t-próba

helyett Welch-próbát kell végezni, ahol tört szabadsági fokú

t-eloszlással veszik figyelembe a Welch-korrekció alkalmazását.

A szóhasználat gyakran ebben az esetben is változatlan: t-próba.

Két SH (=szórás) egyenlőségére vonatkozó F-próba:

Modell: az X változó alapsokaságbeli értékei $\mathcal{N}(\mu_X, \sigma_X^2)$

az Y változó alapsokaságbeli értékei $\mathcal{N}(\mu_Y, \sigma_Y^2)$

ahol σ_X és σ_Y valamint μ_X és μ_Y ismeretlenek

Hipotézis: H_0: $\sigma_X = \sigma_Y$ és az ellenhipotézis H_1: $\sigma_X \neq \sigma_Y$

Minta: N-EVM X-re és M-EVM Y-ra, egymástól független két mintavétel.

Próbastatisztika: $S = \max\left(\dfrac{SSQ_X}{SSQ_Y} \cdot \dfrac{M-1}{N-1} \ , \ \dfrac{SSQ_Y}{SSQ_X} \cdot \dfrac{N-1}{M-1}\right)$

ahol az elfogadási tartomány határa az F-eloszlás egyoldali küszöbértéke.

Egyszempontos variancia-analízis

Modell: a minta csoportokból áll: $\{\{y_{k,1}, y_{k,2}, \ldots, y_{k,N_k}\} \quad , \quad k = 1 \ldots K\}$

az $Y_k \sim \mathcal{N}\left(\mu_k, \sigma_k^2\right)$ változóra vett N_k-elemű EVM: $\{y_{k,1}, y_{k,2}, \ldots, y_{k,N_k}\}$

Hipotézis: H_0: $\mu_1 = \mu_2 = \ldots \mu_K$, az ellenhipotézis: H_1: van két különböző μ

A modellben feltételezzük, hogy $\sigma_1 = \sigma_2 = \ldots = \sigma_K$

Jelölések: $N = \sum_{k=1}^{K} N_k$ (mintanagyság)

$y_{\bullet\bullet} = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N_k} y_{kn}$ (főátlag) , $y_{k\bullet} = \frac{1}{N_k} \sum_{n=1}^{N_k} y_{kn}$ (k-ik csoportátlag)

$SSQ_Y = \sum_{k=1}^{K} \sum_{n=1}^{N_k} (y_{kn} - y_{\bullet\bullet})^2$ (teljes négyzetösszeg)

$SSQ_{Y.B} = \sum_{k=1}^{K} N_k \cdot (y_{k\bullet} - y_{\bullet\bullet})^2$ (csoportok közötti négyzetösszeg)

$SSQ_{Y.W} = \sum_{k=1}^{K} \sum_{n=1}^{N_k} (y_{kn} - y_{k\bullet})^2$ (csoportokon belüli négyzetösszeg)

ANOVA egyenlet: $SSQ_Y = SSQ_{Y.B} + SSQ_{Y.W}$

Próbastatisztika: $S = \frac{SSQ_{Y.B}}{SSQ_{Y.W}} \cdot \frac{N-K}{K-1}$

$S \sim F_{(K-1),(N-K)}$ (szavakban: S eloszlása $(K-1), (N-K)$ szab.fokú F),

ha H_0 igaz.

Egyoldali próbát kell végezni, azaz a 0.05 szignifikancia szintű elfogadási tartomány felső határa az $F_{(K-1),(N-K)}$ eloszlás 0.95 kvantilise.

$SSQ_{Y.B}$ interpretációja: a csoportosítás által megmagyarázott négyzetösszeg

$SSQ_{Y.W}$ interpretációja: a csoportosítás által meg nem magyarázott négyzetösszeg

Eta-négyzet $\eta^2 = \frac{SSQ_{Y.B}}{SSQ_Y}$ interpretációja: a csoportosítás által megmagyarázott négyzetösszeg aránya a teljes négyzetösszegen belül.

Regressziós modell (Gauss): $Y(x_n) = \ell(x_n) + E_n$, $n = 1, 2, .., N$, ahol $E_1, .., E_N$ független $\mathcal{N}(0, \sigma^2)$ eloszlású valószínűségi változók, és σ ismeretlen paraméter.

Lineáris regresszió: amikor $\ell(x_n) = a + b \cdot x_n$ ahol a és b ismeretlen paraméterek.
Itt a lineáris regresszió kifejezést mindig a Gauss regressziós modellre használjuk.

Bemenő adatok: $\{(x_1, y_1), (x_2, y_2), .., (x_N, y_N)\}$ ahol y_n minta $Y(x_n)$-re , $n = 1, 2, .., N$

A modellben az X és Y változó szerepe nem egyenrangú: $Y(x_n)$ valószínűségi változó, mely az $X = x_n$ esemény bekövetkezése mellett van értelmezve. X lehet valószínűségi változó (pl ha X=testmagasság és Y=testsúly, akkor $Y(x_n)$ eloszlása írja le a popu- lációban az x_n testmagasságú emberek testsúlyának eloszlását), de az is lehet, hogy X nem-valószínűségi jellegű (pl $x_1, x_2, .., x_n$ adott időpontok, amikor Y megfigyelhető).

Legkisebb négyzetes (LS) paraméterbecslés: az ℓ függvény paramétereinek olyan értékei, melyre a $\sum_{n=1}^{N} (y_n - \ell(x_n))^2$ négyzetösszeg a lehető legkisebb.

Jelölje $\hat{a}$ és $\hat{b}$ és $\hat{\sigma^2}$ a lineáris regressziós modell paramétereinek LS becslését.
Az LS paraméterbecslésekkel vett $\ell(x_n)$ szokott jelölése: $\hat{y}_n$

$$SSQ_{Y.R} = \sum_{n=1}^{N} (\hat{y}_n - \hat{y}_\bullet)^2 \qquad \text{(a regresszió által megmagyarázott SSQ)}$$

$$SSQ_{Y.W} = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 \quad \text{(a regresszió által meg nem magyarázott SSQ)}$$

ANOVA egyenlet: $SSQ_Y = SSQ_{Y.R} + SSQ_{Y.W}$

Jegyezzük meg, hogy a lin.regressziós egyenes LS becslése átmegy a $(x_\bullet, y_\bullet)$ ponton.
Az $\hat{y}_n$-ek átlaga = y_n-ek átlaga, képletben: $\frac{1}{N}\sum_{n=1}^{N} \hat{y}_n = \frac{1}{N}\sum_{n=1}^{N} y_n$

A lineáris regresszió LS paraméterbecslései: $\hat{\sigma}^2 = \frac{1}{N-2} \cdot SSQ_{Y.W}$

$\hat{b} = \frac{SCP_{XY}}{SSQ_X}$ és $\hat{a} = y_\bullet - x_\bullet \cdot \hat{b}$ ahol

$x_\bullet = \frac{1}{N}\sum_{n=1}^{N} x_n$ és $SSQ_X = \sum_{n=1}^{N} (x_n - x_\bullet)^2$ és $SCP_{XY} = \sum_{n=1}^{N} (x_n - x_\bullet) \cdot (y_n - y_\bullet)$

A lin.regresszió LS paraméterbecslései max. likelihood tulajdonságúak és torzítatlanok,

$\hat{b}$ normális eloszlású, SHnégyzete = $\frac{1}{SSQ_X}\sigma^2$

$r_{XY} = \frac{SCP_{XY}}{\sqrt{SSQ_X \cdot SSQ_Y}}$ (tapasztalati korrelációs együttható)

A tapasztalati korrelációs együttható négyzete: $r_{XY}^2 = \frac{SSQ_{Y.R}}{SSQ_Y}$

A $H_0 : b = 0$ hipotézis vizsgálata: a próbastatisztika $S = (N-2) \cdot \frac{SSQ_{Y.R}}{SSQ_{Y.W}}$

$S \sim F_{1,(N-2)}$ (szavakban: S eloszlása $1, (N-2)$ szab.fokú F), ha H_0 igaz.
Egyoldali próbát kell végezni, azaz a 0.05 szignifikancia szintű elfogadási tartomány felső határa az $F_{1,(N-2)}$ eloszlás 0.95 kvantilise.

Kategória értékű változó eloszlásának vizsgálata (illeszkedésvizsgálat)

Modell: az X változó alapsokaságbeli értékei $X = 1, 2, .., I$

Minta: $\{x_1, x_2, .., x_N\}$ EVM X-re, illetve az ebből számolt $\{f_1, f_2, .., f_I\}$ gyakoriságok,

ahol f_i = az $\{X = i\}$ mintabeli előfordulásainak száma.

Hipotézis: H_0: $\{p_1, p_2, .., p_I\}$ az alapsokaságban X eloszlása

Megjegyzés: f_i várhatóértéke $= e_i = N \cdot p_i$, $i = 1, ..I$, ha H_0 igaz

Próbastatisztika: $S = \sum_{i=1}^{I} \frac{(f_i - e_i)^2}{e_i}$ mely (közelítőleg) χ^2_{I-1}

(szavakban: $I - 1$ szab.fokú khi-négyzet) eloszlású, ha H_0 igaz.

Két kategória értékű változó függetlenségének vizsgálata

Modell: az X változó alapsokaságbeli értékei $X = 1, 2, .., I$

az Y változó alapsokaságbeli értékei $Y = 1, 2, .., J$

Hipotézis: H_0: X és Y függetlenek

Minta: $\{(x_1, y_1), (x_2, y_2), .., (x_N, y_N)\}$ EVM az (X, Y) változópárra

illetve az ebből számolt $\{f_{1,1}, f_{1,2}, ..., f_{2,1}, f_{2,2}, .., f_{I,J}\}$ gyakoriságok,

$f_{i,j}$ = az $\{X = i\}$ és $\{Y = j\}$ mintabeli együttes előfordulásainak száma.

Legyen $e_{i,j} = \frac{f_{i+} \cdot f_{+j}}{f_{++}}$ minden (i, j) párra

Megjegyzés: $f_{i,j}$ várhatóértéke $= e_{i,j}$ várhatóértéke, ha H_0 igaz.

Próbastatisztika: $S = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{i,j} - e_{i,j})^2}{e_{i,j}}$ melynek eloszlása (közelítőleg)

$\chi^2_{(I-1)\cdot(J-1)}$ (szavakban: $(I - 1) \cdot (J - 1)$ szab.fokú khi-négyzet), ha H_0 igaz.

Mind az illeszkedés, mind a függetlenség vizsgálatánál egyoldali próbát kell végezni, azaz a 0.05 szignifikancia szintű elfogadási tartomány felső határa a (megfelelő szab.fokú) khi-négyzet eloszlás 0.95 kvantilise.

Két kategória értékű változó együttes eloszlása

Az (X, Y) változópár együttes eloszlását gyakran táblázatos formában írjuk:

$X \backslash Y$	Y=1	Y=2	...	Y=J	*row total*
X=1	$p_{1,1}$	$p_{1,2}$	...	$p_{1,J}$	$p_{1,+}$
X=2	$p_{2,1}$	$p_{2,2}$	...	$p_{2,J}$	$p_{2,+}$
...	...	...	...	...	...
X=I	$p_{I,1}$	$p_{I,2}$	...	$p_{I,J}$	$p_{I,+}$
col total	$p_{+,1}$	$p_{+,2}$	...	$p_{+,J}$	$p_{+,+}$

ahol $p_{i,j} = P\{X = i \text{ és } Y = j\}$ ami az $\{X = i\}$ és $\{Y = j\}$ alapsokasági együttes előfordulások relatív gyakorisága.

X és Y függetlensége azt jelenti, hogy $p_{i,j} = p_{i,+} \cdot p_{+,j}$ teljesül minden (i, j) párra.

Ezzel teljesen egyenértékű az a feltétel az, hogy $P\{X = i \mid Y = j\} = P\{X = i\}$

azaz $\frac{p_{i,j}}{p_{+,j}} = p_{i,+}$ teljesül minden (i, j) párra (feltéve, hogy $p_{+,j} \neq 0$).