ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΣΤΑΤΙΣΤΙΚΗ

Διάλεξη 13η

19/04/2018

Σύγκριση 2 ποσοστών χωρίς ζεύγη

Παράδ 4.10

$$\frac{\Phi}{50} \to 34(E) \quad 16(A) \qquad\qquad \frac{P}{50} \to 9(E) \quad 41(A)$$

H_0: το φάρμακο δεν έχει αποτέλεσμα.

Γενικά: Έστω 2 πληθυσμοί ανεξάρτητοι και τα μέλη των οποίων ανήκουν σε 2 κατηγορίες A και B.

Στην A κατηγορία να έχουν ένα χαρακτηριστικό γνώρισμα $\underline{\underline{\alpha}}$ και B να μην το έχουν.

Έστω p_1 και p_2 οι πιθανότητες τα μέλη των πληθυσμών να ανήκουν στην κατηγορία A αντίστοιχα.

Ένα τυχαίο δείγμα (τ.δ.) μεγέθους n_1 επιλέγεται από τον πληθ. 1 και έστω x_1 από τα μέλη του να είναι της κατηγορίας A.

Επίσης ένα τ.δ. μεγέθους n_2 από τον πληθυσμό 2 και x_2 από τα μέλη του να είναι της κατηγορίας A.

Αυτό που μας ενδιαφέρει είναι η συμπερασματολογία για $p_1 - p_2$ (ή να γίνει ο έλεγχος για $p_1 = p_2$.)

$$\boxed{\hat{p}_1 = \frac{x_1}{n_1}} \to X_1 \sim Bin(n_1, p_1) \to X_1 \overset{\text{προσ.}}{\sim} N(n_1 p_1, n_1 p_1 (1-p_1))$$

$$\boxed{\hat{p}_2 = \frac{x_2}{n_2}} \to X_2 \sim Bin(n_2, p_2) \to X_2 \overset{\text{προσ.}}{\sim} N(n_2 p_2, n_2 p_2 (1-p_2))$$

→ $\hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$ και $\hat{p}_1 - \hat{p}_2$ προσ. (*)

$$\boxed{E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2}$$

$$\boxed{Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$\boxed{\hat{\sigma}^2_{\hat{p}_1 - \hat{p}_2} = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

(*) $\hat{p}_1 - \hat{p}_2 \overset{\text{προσεγ.}}{\sim} N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right) \rightarrow \dfrac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$

▶ Για να ελέγξουμε την υπόθεση

$H_0: p_1 = p_2 = p$ v $H_a: p_1 \neq p_2$

χρησιμοποιώντας το στατιστικό: $Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \left(\overset{\text{προσ.}}{\underset{H_0}{\sim}} N(0,1)\right)$

, $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$

και κρίσιμη περιοχή μεγέθους α , $|Z| > z_{\alpha/2}$

ένα $(1-\alpha)100\%$ Δ.Ε. για $p_1 - p_2$:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$H_0: p_1 = p_2 = p$ v $H_a: p_1 > p_2$

παραδ. 4.10:

$\hat{p}_1 = \frac{34}{50} = 0.68$

$\hat{p}_2 = \frac{9}{50} = 0.18$

$\hat{p}_3 = \frac{34+9}{50+50} = 0.43$

$Z = \dfrac{0.68 - 0.18}{\sqrt{0.43(1-0.43)\left(\frac{1}{50} + \frac{1}{50}\right)}}$ $(= 5.05)$

και περιοχή $\alpha = 0.01$

(συνέχεια παράδ. 4.10)

$Z \geq Z_{0.01} (= 2.326)$, απορρ. η H_0.

p-τιμή: $P(Z \geq 5.05) \approx 0.0 < 0.001$.
απορρ. H_0.

95% Δ.Ε. (διάστημα εμπιστοσύνης): $(0.33, 0.67)$

Συμπερασματολογία για τις διακυμάνσεις (κανονικοί πληθυσμοί)

(A) Μία διακύμανση (χ^2 τεστ)

Παράδ 4.7:

$4.2, 5.3, 10.7, \dots, N(\mu, \sigma^2)$

$H_0: \sigma = 2 \ (\alpha = 0.05)$

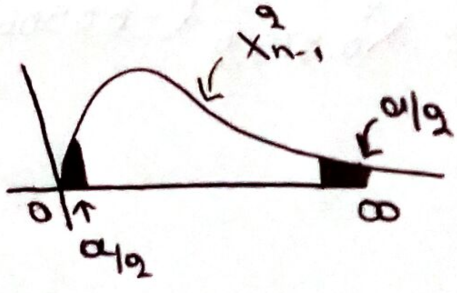
$$S^2 = \frac{1}{n-1} \sum_{1}^{n} (x_i - \bar{x})^2$$

Γενικά: $X_1, \dots, X_n$ από $N(\mu, \sigma^2)$

Αυτό που μας ενδιαφέρει είναι το σ^2

$\hat{\sigma}^2 = S^2$, $E(S^2) = \sigma^2$: αμερόληπτος.

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$



$$P\left(\chi^2_{1-\frac{\alpha}{2}, n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2, n-1}\right) = 1 - \alpha \rightsquigarrow$$

$$P\left(\underbrace{\frac{(n-1)S^2}{\chi^2_{\alpha/2, n-1}}}_{L} \leq \sigma^2 \leq \underbrace{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2, n-1}}}_{U}\right) = 1 - \alpha$$

Για να ελέγξουμε υποθέσεις της μορφής:

(i) $H_0: \sigma^2 \le \sigma_0^2 \; (\sigma = \sigma_0)$ v $H_a: \sigma^2 > \sigma_0^2$

(ii) $H_0: \sigma^2 \ge \sigma_0^2$ v $H_a: \sigma^2 < \sigma_0^2$

(iii) $H_0: \sigma^2 = \sigma_0^2$ v $H_a: \sigma^2 \ne \sigma_0^2$

Χρησιμοποιείται το στατιστικό:

$$X^2 = \frac{(n-1)S^2}{\sigma_0^2} \left(\overset{H_0}{\sim} X^2_{n-1} \right)$$ και κρίσιμες περιοχές μεγέθους α

(i) $X^2 \ge X^2_{a,n-1}$

(ii) $X^2 \le X^2_{1-a,n-1}$

(iii) $X^2 \ge X^2_{a/2,n-1}$ και $X^2 \le X^2_{1-\frac{a}{2},n-1}$.

άρα (για το παράδειγμα)

$$S^2 = \frac{1}{n-1} \sum_{1}^{n} (x_i - \bar{x})^2 (= 10.57)$$

$$X^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{2 * 10.57}{4} (= 5.285)$$

$X^2 \le X^2_{0.975,2} (= 0,0506)$, $X^2 \ge X^2_{0.025,2} (= 7.378)$

Δεν απορρ. η H_0.

(B) Σύγκριση 2 διακυμάνσεων - F test.

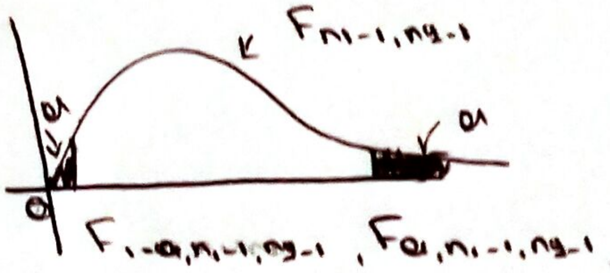
A: $n_1 = 10$, $S_1^2 = 1.95$

B: $n_2 = 8$, $S_2^2 = 0.98$

Υπάρχει διαφορά στην ποιότητα των A και B.
($\alpha = 0.10$).

Γενικά: τ.δ. $X_{11}, \dots, X_{1n}$ από $N(\mu_1, \sigma_1^2)$ και τ.δ. $X_{21}, \dots, X_{2n}$ από άλλο ανεξ. $N(\mu_2, \sigma_2^2)$. Μας ενδιαφέρει η σύγκριση των διακυμάνσεων σ_1^2 και σ_2^2.

$$\left.\begin{array}{l} \dfrac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1} \\[2ex] \dfrac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1} \end{array}\right\} \Rightarrow \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$



Για να ελέγξουμε υποθέσεις τις μορφής:

(i) $H_0: \sigma_1^2 \leq \sigma_2^2$ v $H_a: \sigma_1^2 > \sigma_2^2$

(ii) $H_0: \sigma_1^2 \geq \sigma_2^2$ v $H_a: \sigma_1^2 < \sigma_2^2$.

(iii) $H_0: \sigma_1^2 = \sigma_2^2$ v $H_0: \sigma_1^2 \neq \sigma_2^2$.

Χρησιμοποιώντας το στατιστικό

$$F = \frac{S_1^2}{S_2^2} \quad (\sim F_{n_1-1,n_2-1} \text{, όταν } H_0 \text{ αληθής}) \text{ και}$$

κρίσιμες περιοχές μεγέθους α: (i) $F \geq F_{\alpha,n_1-1,n_2-1}$

(ii) $F \leq F_{1-\alpha,n_1-1,n_2-1}$

(iii) $F \leq F_{1-\frac{\alpha}{2},n_1-1,n_2-1}$ ή $F \geq F_{\frac{\alpha}{2},n_1-1,n_2-1}$

άρα : $H_0 : \sigma_1^2 = \sigma_2^2$ v $H_1 : \sigma_1^2 \neq \sigma_2^2$

$$F = \frac{S_1^2}{S_2^2} \left(\overset{H_0}{\sim} F_{9,7} \right)$$

κρίσ. περιοχή $F \leq F_{0.95,9.7}$ και $F \geq F_{0.05,9.7} \; (= 3.68)$

$$F = \frac{1.95}{0.28} = 4.46$$, επειδή $4.46 > 3.68$ απορ. H_0.

$$F_{0.95,9.7} = \frac{1}{F_{0.05,7.9}} = \frac{1}{3.29} = 0,30$$