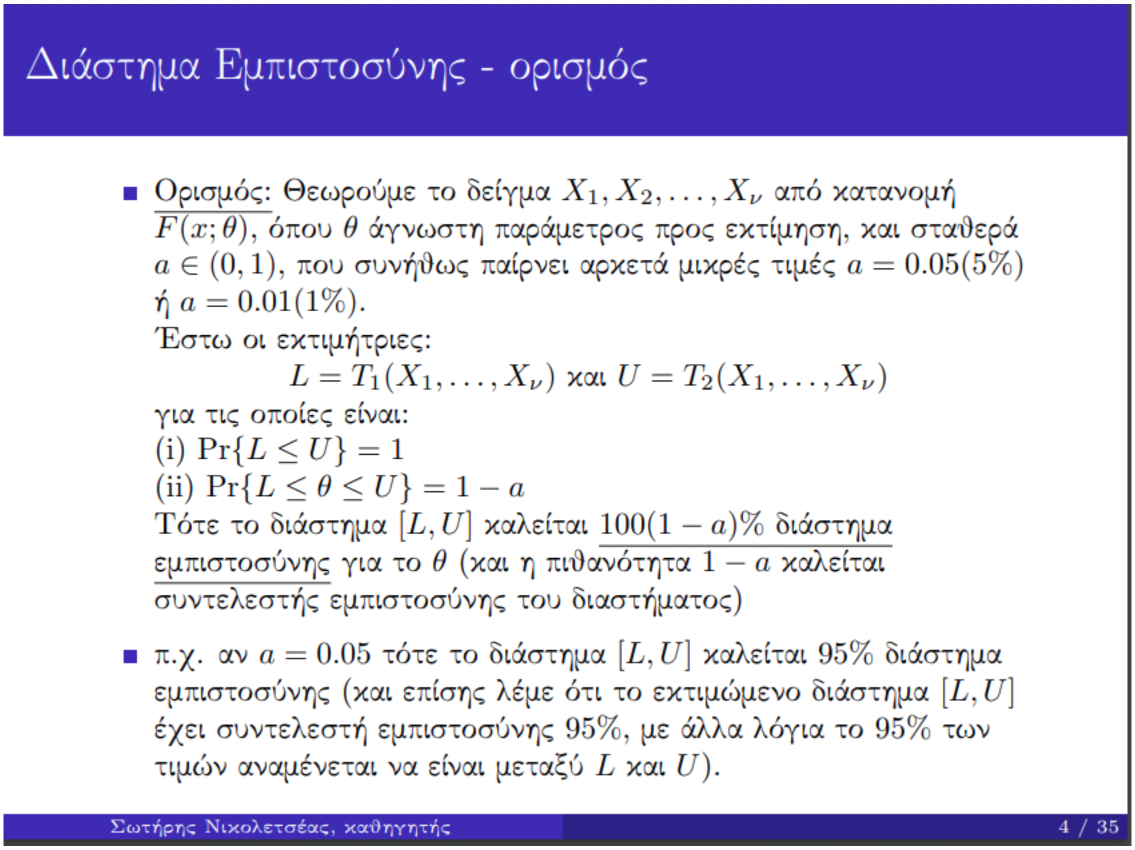
Θεωρία

Tuesday, 17 January 2023 23:54

*Αντί για μία σημειακή εκτίμηση (δηλαδή την εκτίμηση μιας μοναδικής τιμής) θα ήταν συχνά περισσότερο χρήσιμο να προσδιορίσουμε ένα διάστημα τιμών που η παράμετρος παίρνει με "αρκετά μεγάλη" βεβαιότητα (π.χ. Με πιθανότητα 95%).



*Για να βρούμε τα διαστήματα εμπιστοσύνης χρησιμοποιούμε βοηθητικές κατανομές, οι οποίες προέρχονται από την Κανονική Κατανομή:

- Η Κατανομή X^2

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$
$$X \sim \chi_n^2$$

- Η Κατανομή t του Student

$$T_n = \frac{Z}{\sqrt{\frac{\chi_n^2}{n}}}$$

- Η Κατανομή Γ

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{a-1}}{\Gamma(a)} \\ 0, \end{cases}$$

- Η F - Κατανομή του Fisher

$$F_{n,m} = \frac{\chi_n^2/n}{\chi_m^2/m}$$

*Ψάχνουμε διαστήματα εμπιστοσύνης για:

- Μέση τιμή κανονικού δείγματος με γνωστή διασπορά
- Μέση τιμή κανονικού δείγματος με άγνωστη διασπορά
- Μέση τιμή για αρκετά μεγάλα δείγματα από οποιαδήποτε κατανομή (προσεγγιστικά)
- Διασπορά Κανονικού Δείγματος
- Διαφορά των μέσων τιμών δύο κανονικών δειγμάτων

Άσκηση 1: Διάστημα Εμπιστοσύνης Κανονικού Μέσου με Γνωστή Διασπορά

Για τη μελέτη του χρόνου ζωής μιας συσκευής που παράγεται σε ένα εργοστάσιο, λήφθηκε το παρακάτω δείγμα (σε ώρες):

167, 184, 165, 174, 167, 180, 168, 173, 162

Αν ο χρόνος ζωής της συσκευής ακολουθεί κανονική κατανομή με διασπορά 219.04, να βρεθεί το 95% διάστημα εμπιστοσύνης της μέσης τιμής του χρόνου ζωής των συσκευών που παράγει το εργοστάσιο.
Δίνονται $Z_{0.025}$=1.96, $Z_{0.05}$=1.64, $Z_{0.1}$=1.28.

$$\bar{X} = \frac{1}{v} \sum x_i = \frac{1540}{9} = 171.11$$

$$\sigma = \sqrt{\sigma^2} = 14.8$$

Το δ. ε είναι $\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) =$

$$= \left(171.11 - z_{0.025}\cdot\frac{14.8}{3}, 171.11 + z_{0.025}\cdot\frac{14.8}{3}\right)$$

$\alpha = 0,05 \rightarrow \alpha/2 = 0.025$

$$= \left(171.11 - 1.96\cdot\frac{14.8}{3}, 171.11 + 1.96\cdot\frac{14.8}{3}\right)$$

$$= (161.44, 180.77)$$

Άσκηση 2: Διάστημα Εμπιστοσύνης Κανονικού Μέσου με Άγνωστη Διασπορά

Σε μελέτη που έγινε για την περιεκτικότητα του γάλακτος αγελάδος σε μια ουσία, λήφθηκε δείγμα γάλακτος από 20 διαφορετικές αγελάδες. Η ποσότητα (σε ppm) που βρέθηκε για κάθε μια από τις αγελάδες είναι:

16, 0, 0, 2, 3, 6, 8, 2, 5, 0, 12, 10, 5, 7, 2, 3, 8, 17, 9, 1

Να βρεθεί 95% διάστημα εμπιστοσύνης για τη μέση περιεκτικότητα στην ουσία αυτή.

Αφού δεν ξέρουμε τη διασπορά, θα χρησιμοποιήσουμε τη κατανομή t

$$\bar{x} = \frac{1}{v}\sum x_i = \frac{116}{20} = 5.8$$

$$s^2 = \frac{1}{v-1}\left[\left(\sum_{i=1}^{n} x_i^2\right) - n\cdot\bar{x}^2\right] = 25.80$$

$$s = \sqrt{25.8} = 5.085$$

Αφού το δείγμα έχει μέγεθος 20, έχουμε $n-1 = 19$ βαθμούς ελευθερίας

Άρα βρίσκουμε το $t_{19/0.025} = 2.093$

$\downarrow$

$n-1\ ,\ \alpha/2$

Θέλουμε 95% δ.ε, άρα $\alpha = 0.05$

$\alpha/2 = 0.025$

Άρα το δ.ε είναι:

$$\left(\bar{x} - t_{19,0.025} \cdot \frac{S}{\sqrt{n}}, \bar{x} + t_{19,0.025} \cdot \frac{S}{\sqrt{n}}\right) =$$

$$= \left(5.8 - 2.093 \cdot \frac{5.085}{\sqrt{20}}, 5.8 + 2.093 \cdot \frac{5.085}{\sqrt{20}}\right)$$

$$= (2.55, 9.05)$$

Άσκηση 3: Εύρεση μεγέθους δείγματος για Διάστημα Εμπιστοσύνης

Από μελέτες έχει παρατηρηθεί ότι το βάρος του σολομού που μεγαλώνει σε ένα ιχθυοτροφείο ακολουθεί κανονική κατανομή με μέση τιμή που ποικίλλει από χρονιά σε χρονιά με σταθερή όμως τυπική απόκλιση στα 0.3 κιλά. Αν θέλουμε να είμαστε 90% σίγουροι ότι η προσέγγισή του μέσου βάρους του σολομού είναι σωστή με σφάλμα ±0.1 κιλά, πόσο μεγάλο πρέπει να είναι το δείγμα που θα πάρουμε για να κάνουμε την προσέγγιση αυτή μέσω διαστήματος εμπιστοσύνης;

Ξέρουμε τη διασπορά ($\sigma = 0,3$), άρα το Δ.Ε 90% για δείγμα n θα είναι:

$$\left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} , \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Θέλουμε 90% Δ.Ε, άρα $\alpha = 0,1 \rightarrow \alpha/2 = 0.05$

$$z_{0.05} = 1.645$$

$$\left(\bar{X} - 1.645 \cdot \frac{\sigma}{\sqrt{n}} , \bar{X} + 1.645 \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Θέλουμε $2 \cdot 1,645 \cdot \frac{\sigma}{\sqrt{n}} \leq 2 \cdot 0,1$ σφάλμα προσέγγισης ±0,1 κιλά

$\Leftrightarrow$ Θέλουμε το μήκος του διαστήματος να είναι το πολύ 0,2

$$\Leftrightarrow \frac{2 \cdot 1,645 \cdot 0,3}{0,2} \leq \sqrt{n} \Leftrightarrow$$

$$\Leftrightarrow n \geq (3 \cdot 1,645)^2 \Leftrightarrow$$

$$\Leftrightarrow n \geq 24,35$$

Πρέπει το μέγεθος του δείγματος να είναι τουλάχιστον $n = 25$

Άσκηση 4: Διάστημα Εμπιστοσύνης διαφοράς μέσων για δύο κανονικούς πληθυσμούς

Τα παρακάτω δεδομένα παριστάνουν τις βαθμολογίες 9 μαθητών Λυκείου σε ένα test μαθηματικών:

3, 8, 4, 11, 8, 6, 9, 10, 5

Επίσης, στο ίδιο test τα αποτελέσματα 6 μαθητριών ήταν

16, 13, 20, 16, 15, 13

Να βρεθεί διάστημα εμπιστοσύνης 95% για τη διαφορά $\mu_1 - \mu_2$ του μέσου βαθμού στους πληθυσμούς μαθητών και μαθητριών, θεωρώντας ότι η διασπορά είναι κοινή και ότι οι βαθμοί και στις δύο περιπτώσεις προέρχονται από κανονική κατανομή (αν και στην πραγματικότητα παίρνουν διακριτές τιμές)

Μαθητές: $n_1 = 9$ Θα βρούμε τα $\bar{X}, \bar{Y}$ και S_1^2, S_2^2 (μαθητές, μαθήτριες)

Μαθήτριες: $n_2 = 6$

$\bar{X} = 7{,}11$ $S_1^2 = 7.67$

$\bar{Y} = 15{,}5$ $S_2^2 = 6{,}7$

$$S_p = \sqrt{\frac{1}{n_1+n_2-2} \cdot \left((n_1-1)\cdot S_1^2 + (n_2-1)\cdot S_2^2\right)} =$$

$$= \sqrt{\frac{1}{13}\cdot\left(8 \cdot 7{,}67 + 5 \cdot 6{,}7\right)} = 2{,}7$$

$$t_{n_1+n_2-2,\,\alpha/2} = t_{13,\,0.025} = 2.16 \text{ (Θα δίνεται στην εκφώνηση)}$$

Το δ.ε είναι:

$$\left[(\bar{X}-\bar{Y}) - S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \cdot t_{n_1+n_2-2,\alpha/2},\ (\bar{X}-\bar{Y}) + S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \cdot t_{n_1+n_2-2,\alpha/2}\right]$$

$$= \left[-8{,}39 - 3{,}07,\ -8{,}39 + 3{,}07\right] = \left[-11{,}46,\ -5{,}32\right]$$

Θεωρία

Wednesday, January 18, 2023 4:44 PM

Ορισμός: Στατιστική υπόθεση είναι μία πρόταση για έναν πληθυσμό, που συνήθως μεταφράζεται σε μια σχέση ως προς μία παράμετρο του πληθυσμού (π.χ. τη μέση τιμή).

Ορισμός: Στατιστικό test TS είναι μία στατιστική συνάρτηση της οποίας η τιμή προσδιορίζεται από ένα δείγμα ενός πληθυσμού. Ανάλογα με την τιμή του test, η μηδενική υπόθεση απορρίπτεται ή όχι.

Οι πιο σημαντικοί έλεγχοι:
- Το Z-test για τη μέση τιμή κανονικού πληθυσμού με γνωστή διασπορά (χρησιμοποιούμε τη τυπική κανονική κατανομή Z) ,
- Το t-test για τη μέση τιμή ενός κανονικού πληθυσμού με άγνωστη διασπορά (χρησιμοποιούμε κατανομή t),

Για τον λόγο αυτό διακρίνουμε δύο τύπους σφάλματος του test:

	Αποδοχή της H_0	Απόρριψη της H_0
Αληθής η H_0	Ορθή απόφαση (πιθανότητα $1-\alpha$)	Σφάλμα τύπου I (πιθανότητα α)
Αληθής η H_1	Σφάλμα τύπου II (πιθανότητα β)	Ορθή απόφαση (πιθανότητα $1-\beta$)

Επομένως το σφάλμα τύπου I (απόρριψη αληθούς H_0) είναι σοβαρότερο από το σφάλμα τύπου II (αποδοχή εσφαλμένης H_0).

Η τιμή της πιθανότητας α καλείται <u>επίπεδο σημαντικότητας</u> (significance level) του test.

Αυτό επιτυγχάνεται επιλέγοντας ένα αρκετά χαμηλό επίπεδο σημαντικότητας α ($\alpha = 0.10$, 0.05 ή 0.01), απαιτώντας ουσιαστικά η πιθανότητα απόρριψης της H_0, όταν η H_0 αληθεύει, να είναι το πολύ α, δηλαδή:

$$\alpha = Pr\{\text{απόρριψη της } H_0 \mid H_0 \text{ αληθής}\}$$

Διαισθητικά, όσο πιο μικρό είναι το α, τόσο πιο "χαλαρό" είναι το test (δηλαδή, "είμαστε διατεθειμένοι" να αποδεχθούμε αποκλίσεις στο δείγμα) και απορρίπτουμε την H_0 μόνο όταν είμαστε αρκετά σίγουροι!

Με άλλα λόγια, αν η H_0 δεν "περνάει" τον έλεγχο ενός "χαλαρού" test, τότε είμαστε αρκετά σίγουροι ότι καλώς απορρίπτεται.

Ένα μικρό α ουσιαστικά σημαίνει μια ευρύτερη αποδοχή αποκλίσεων, δηλαδή μια μικρότερη περιοχή απόρριψης C.

Επομένως, είμαστε περισσότερο σίγουροι για την απόφασή μας όταν απορρίπτουμε την H_0 λόγω ασυμφωνίας της με το δείγμα, οπότε ισχύει η εναλλακτική της H_1.

Για τον λόγο αυτό συνήθως ορίζουμε την υπόθεση που θέλουμε να επιβεβαιώσουμε ως H_1. Αντίθετα, όταν θέλουμε να δείξουμε ότι μία υπόθεση δεν ισχύει, την ορίζουμε ως μηδενική υπόθεση H_0.

Η τιμή p (p-value) ενός test

- Καλούμε p-value ενός test το μικρότερο επίπεδο σημαντικότητας α για το οποίο το test απορρίπτει την υπόθεση H_0.
- Επομένως, μια μικρή p-value σημαίνει ισχυρή ένδειξη ότι η H_0 δεν αληθεύει, αφού ακόμα και ένα "χαλαρό" test που "ανέχεται" μεγάλες αποκλίσεις, την απορρίπτει.
- Στην πράξη, πρώτα υπολογίζουμε την p-value και στη συνέχεια, ανάλογα με την περίπτωση, επιλέγουμε κατάλληλο επίπεδο σημαντικότητας α.
- Μια πολύ μεγάλη p-value υπαινίσσεται ότι η H_0 δεν πρέπει να απορριφθεί ενώ μια πολύ μικρή p-value είναι ένδειξη ότι η H_0 πρέπει να απορριφθεί.

Η τιμή p (p-value)

Είναι $\text{TS} = \frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0|$

απόρριψη $H_0 \Leftrightarrow |TS| \geq z_{\alpha/2}$

και p-value είναι η μικρότερη τιμή α για την οποία απορρίπτουμε την H_0. Ας βρούμε αυτήν την μικρότερη τιμή.

Άρα $z_{\alpha/2} \leq |TS| \Leftrightarrow$
$\Pr\{Z \geq z_{\alpha/2}\} \geq Pr\{Z \geq |TS|\} \Leftrightarrow$
$\frac{\alpha}{2} \geq Pr\{Z \geq |TS|\} \Leftrightarrow$
$\alpha \geq 2Pr\{Z \geq |TS|\}$

Άρα η μικρότερη τιμή (δηλαδή η p-value) είναι :

$$\alpha = p\text{-value} = 2Pr\{Z \geq |TS|\}$$

p-value (Σύνοψη)

Επομένως, η p-value του Z-test υπολογίζεται ως εξής:

- Στην αρχή υπολογίζουμε την τιμή του στατιστικού test:

$$TS = \sqrt{n}\frac{\bar{X} - \mu_0}{\sigma} = \text{v}$$

- Στη συνέχεια η p-value είναι:

$$p\text{-value} = 2Pr\{Z \geq |\text{v}|\}$$

- Αν p-value < a -> Απορρίπτουμε την ***Ho***
- Αν p-value >= a -> Δεν απορρίπτουμε την ***Ho***

A. Z-test

Table 9.1 Hypothesis Tests Concerning the Mean μ of a Normal Population with Known Variance σ^2

$X_1, \ldots, X_n$ are sample data, and $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

H_0	H_1	Test statistic TS	Significance-level-α test	p value if $TS = v$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\sqrt{n}\,\frac{\overline{X}-\mu_0}{\sigma}$	Reject H_0 if $\lvert TS \rvert \geq z_{\alpha/2}$ Do not reject H_0 otherwise	$2P\{Z \geq \lvert v \rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\sqrt{n}\,\frac{\overline{X}-\mu_0}{\sigma}$	Reject H_0 if $TS \geq z_\alpha$ Do not reject H_0 otherwise	$P\{Z \geq v\}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\sqrt{n}\,\frac{\overline{X}-\mu_0}{\sigma}$	Reject H_0 if $TS \leq -z_\alpha$ Do not reject H_0 otherwise	$P\{Z \leq v\}$

B. t-test

Table 9.2 Hypothesis Tests Concerning the Mean μ of a Normal Population with Unknown Variance σ^2

$X_1, \ldots, X_n$ are sample data; $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, $S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$

H_0	H_1	Test statistic TS	Significance-level-α test	p value if $TS = v$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\sqrt{n}\,\frac{\overline{X}-\mu_0}{S}$	Reject H_0 if $\lvert TS \rvert \geq t_{n-1,\alpha/2}$ Do not reject otherwise	$2P\{T_{n-1} \geq \lvert v \rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\sqrt{n}\,\frac{\overline{X}-\mu_0}{S}$	Reject H_0 if $TS \geq t_{n-1,\alpha}$ Do not reject otherwise	$P\{T_{n-1} \geq v\}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\sqrt{n}\,\frac{\overline{X}-\mu_0}{S}$	Reject H_0 if $TS \leq -t_{n-1,\alpha}$ Do not reject H_0 otherwise	$P\{T_{n-1} \leq v\}$

T_{n-1} is a t random variable with $n-1$ degrees of freedom, and $t_{n-1,\alpha}$ and $t_{n-1,\alpha/2}$ are such that $P\{T_{n-1} \geq t_{n-1,\alpha}\} = \alpha$ and $P\{T_{n-1} \geq t_{n-1,\alpha/2}\} = \alpha/2$.

Άσκηση 1

Wednesday, January 18, 2023 4:44 PM

Ένα παλαιότερο δείγμα από ψάρια στη λίμνη Michigan υποδείκνυε ότι η μέση τιμή της συγκέντρωσης μιας τοξικής ουσίας ήταν 11.2 ppm ανά ψάρι, με τυπική απόκλιση 2 ppm.
Έστω τώρα ένα νέο δείγμα 10 ψαριών με τις εξής συγκεντρώσεις στην ουσία αυτή:

11.5, 12, 11.6, 11.8, 10.4, 10.8, 12.2, 11.9, 12.4, 12.6

Αν η τυπική απόκλιση του δείγματος παραμένει στα 2 ppm, να ελεγχθεί η υπόθεση ότι η μέση συγκέντρωση στην τοξική αυτή ουσία συνεχίζει να έχει την τιμή 11.2 ppm. Να χρησιμοποιηθεί επίπεδο σημαντικότητας 5%.

· Έχουμε $\mu_0 = 11{,}2$ ppm , $\sigma = 2$ ppm , $n = 10$

την μηδενική υπόθεση

$H_0 : \mu_0 = 11{,}2$ ppm ένατι της εναλλακτικής υπόθεσης $H_1 : \mu \neq 11{,}2$

Αφού ξέρουμε τη διασπορά του πληθυσμού, θα χρησιμοποιήσουμε z-test. Άρα πρέπει να βρούμε το:

$$TS = \frac{\sqrt{n}}{\sigma} \cdot |\bar{X} - \mu_0|$$

Και θα ελέγξουμε αν είναι μεγαλύτερο από το $z_{\alpha/2}$ όπου $\alpha = 0{,}05 \rightarrow \alpha/2 = 0{,}025$

$$\bar{X} = \frac{1}{v} \cdot \sum_{i=1}^{n=10} x_i = 11{,}72$$

· Άρα $TS = \frac{\sqrt{10}}{2} \cdot |11{,}72 - 11{,}2| = \frac{3{,}16}{2} \cdot 0{,}52 = 0{,}8222$

·

Από τον πίνακα τυπικής κατανομής $z_{0,025} = 1.96$

Άρα $TS < z_{0,025}$. Επομένως, δεν απορρίπτουμε την H_0. Τα δεδομένα δεν είναι ικανά να απορρίψουν την αρχική υπόθεση.

Το ίδιο υποδεικνύει και η p-value, η οποία είναι:

$$p\text{-value} = 2 \cdot P(Z > 0{,}8222) = 2 \cdot 0{,}2061 = 0{,}4122$$

η οποία είναι $> \alpha$, άρα δεν απορρίπτουμε την H_0

Άσκηση 2

Tuesday, 17 January 2023 16:47

Έστω ένας στατιστικός έλεγχος με H_0: $\mu \leq 100$ και H_1: $\mu > 100$.
Έστω ένα δείγμα μεγέθους 20 με δειγματικό μέσο $\bar{X}$=105.
Να βρεθεί το p value του ελέγχου, αν η τυπική απόκλιση είναι γνωστή και ίση με
(α) 5
(β) 10

Αφού ξέρουμε τη διασπορά, ο έλεγχος είναι z-test

α) $n=20$, $\bar{X}=105$, $\sigma=5$

$$TS = \frac{\sqrt{n}}{\sigma}|\bar{X}-\mu_0| = \frac{\sqrt{20}}{5}(105-100) = 4,472$$

και p-value $= P(Z > 4,472) \simeq 0$ (από πίνακα Z)

β) $n=20$, $\bar{X}=105$, $\sigma=10$

$$TS = \frac{\sqrt{n}}{\sigma}|\bar{X}-\mu_0| = \frac{\sqrt{20}}{10}(105-100) = \frac{\sqrt{20}\cdot 5}{10} =$$

$$= 2,2361$$

και p-value $= P(Z > 2,2361) = 0,0125$

$$1 - P(Z \leq 2,2361) = \Phi(2,2361)$$

Άσκηση 3

Tuesday, 17 January 2023 16:56

Μια επιχείρηση έχει έσοδα 2000 ευρώ ανά μέρα. Για να ελέγξουν οι διοικούντες αν τα έσοδα της επιχείρησης έχουν επηρεαστεί από την αλλαγή στην οικονομία (είτε θετικά, είτε αρνητικά), αποφάσισαν να μελετήσουν τα έσοδα των επόμενων 8 ημερών. Αν αυτά ήταν τα παρακάτω:

2050, 2212, 1880, 2121, 2205, 2018, 1980, 2188

(α) ποια είναι η μηδενική και η εναλλακτική υπόθεση;

(β) είναι τα δεδομένα αρκετά, με επίπεδο σημαντικότητας 5%, για να αποδείξουν ότι έχει επηρεαστεί η επιχείρηση;

α) $H_0: \mu = 2000$, $H_1: \mu \neq 2000$

β) $n=8$ → μέγεθος δείγματος

Αφού δεν γνωρίζουμε τη διασπορά, θα χρησ/με t-test

$\bar{X} = 2031{,}7$, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) =$

$= \frac{1}{n-1}\left[\left(\sum_{i=1}^{n=8} x_i^2\right) - n\bar{x}^2\right] =$

$= 14448.04$

Άρα $S = 120{,}2$

$TS = \sqrt{8} \cdot \frac{2031{,}7 - 2000}{120{,}2} = 1{,}92$

Θα συγκρίναμε το TS με τη τιμή της κατανομής

$t: \; t_{7,0{,}025} = 2{,}365$

$\underset{n-1,}{\downarrow}\;\underset{\alpha/2}{\downarrow}$

Επομένως, $TS < t_{7,0{,}025}$

Άρα τα δεδομένα δεν είναι ικανά να απορρίψουν τη μηδενική υπόθεση H_0