ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ
ΜΕΡΟΣ Α

Δημήτρης Κουγιουμτζής
e-mail: ***dkugiu@auth.gr***
Ιστοσελίδα αυτού του τμήματος του μαθήματος:
http://users.auth.gr/~dkugiu/Teach/CivilTransport/index.html

Στατιστική:

- **Δειγματοληψία** **X**
  συλλογή δεδομένων

- **Περιγραφική στατιστική** **V**
  πίνακες, γραφήματα, συνοπτικά μέτρα

- **Στατιστική συμπερασματολογία** **V**
  ανάλυση δεδομένων, εκτίμηση παραμέτρων, μοντέλα

Δεδομένα

Απογραφή → **Πληθυσμός**

Δειγματοληψία → **Δείγμα**
αντιπροσωπευτικό, τυχαίο

Μεταβλητή

χαρακτηριστικό που η τιμή του αλλάζει
από στοιχείο σε στοιχείο του πληθυσμού

αλλάζει σύμφωνα με κάποιους
καθορισμένους νόμους

→ μεταβλητή (κατάστασης, συστήματος)

αλλάζει με αβεβαιότητα
ή τυχαιότητα

→ **τυχαία μεταβλητή (τ.μ.)**

X: μέσο μεταφοράς
Y: χρόνος διαδρομής

τ.μ. X

ποιοτική τ.μ.
κατηγορίες

ποσοτική τ.μ.
αριθμητικές τιμές, μονάδα μέτρησης

διακριτή τ.μ.
διακεκριμένες τιμές

συνεχής τ.μ.
τιμές σε διάστημα

Διάκριση δεδομένων $x_1, x_2, \ldots$

κατηγορικά

ονομαστικά
δεν υπάρχει
έννοια διάταξης

διατακτικά
η διάταξη έχει
σημασία

αριθμητικά (μετρήσεις)
αριθμητικές τιμές, μονάδα μέτρησης

διακριτά

συνεχή

Ομαδοποίηση αριθμητικών δεδομένων ➔ κατηγορικά

Περιγραφική Στατιστική

Πίνακας και διαγράμματα συχνοτήτων

μικρός αριθμός δυνατών διακεκριμένων τιμών (αριθμητικές τιμές ή κατηγορίες)

Δείγμα $x_1, x_2, x_3, x_4, \ldots, x_n$

Διακεκριμένες τιμές $a_1, a_2, \ldots, a_m$

Παράδειγμα

Χρήση μέσου μεταφοράς προς την εργασία

100 εργαζόμενοι που επιλέχθηκαν τυχαία

Δείγμα

$x_1, x_2, x_3, x_4, \ldots, x_{100}$

Μέσο μεταφοράς:

τ.μ. (ποιοτική)

λεωφορείο,

λεωφορείο-μετρό,

αυτοκίνητο,

άλλο (ποδήλατο, πόδια κτλ)

Σύνολο διακεκριμένων τιμών

Ερωτήματα:

Πιο «συχνό» μέσο μεταφοράς ?

Περισσότεροι χρησιμοποιούν αυτοκίνητο ή λεωφορείο?

Σε τι ποσοστό είναι τα μέσα μαζικής μεταφοράς?

f_i : **συχνότητα** εμφάνισης της τιμής a_i

σχετική συχνότητα (ποσοστό): $p_i = \dfrac{f_i}{n}$ (n: μέγεθος δείγματος)

αθροιστική συχνότητα: $F_i = \sum_{j=1}^{i} f_j$ (όπου $a_j < a_i$ για $i<j$)

αθροιστική σχετική συχνότητα: $P_i = \sum_{j=1}^{i} p_j$ (όπου $a_j < a_i$ για $i<j$)

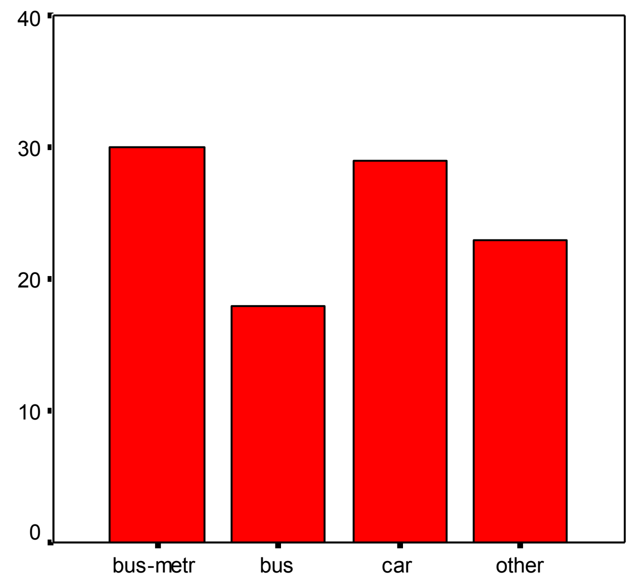
Πίνακας συχνοτήτων

transport mean

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	bus-metr	30	30,0	30,0	30,0
	bus	18	18,0	18,0	48,0
	car	29	29,0	29,0	77,0
	other	23	23,0	23,0	100,0
	Total	100	100,0	100,0	

Ραβδόγραμμα

transport mean



transport mean

Ομαδοποίηση και παρουσίαση αριθμητικών δεδομένων

- Χωρίζουμε τα δεδομένα σε ομάδες
- Παρουσιάζουμε όπως πριν, ομάδα → α_i

Επιλογή ομάδων:

Πολλές ομάδες ➔ μικρές συχνότητες

Λίγες ομάδες ➔ χάνουμε πολλή πληροφορία

Χωρισμός σε ομάδες:

x_{min} : μικρότερη τιμή

x_{max} : μεγαλύτερη τιμή

➔ $R = x_{max} - x_{min}$: εύρος δεδομένων

Επιλογή K ομάδων ➔ R/K : εύρος τιμών κάθε ομάδας

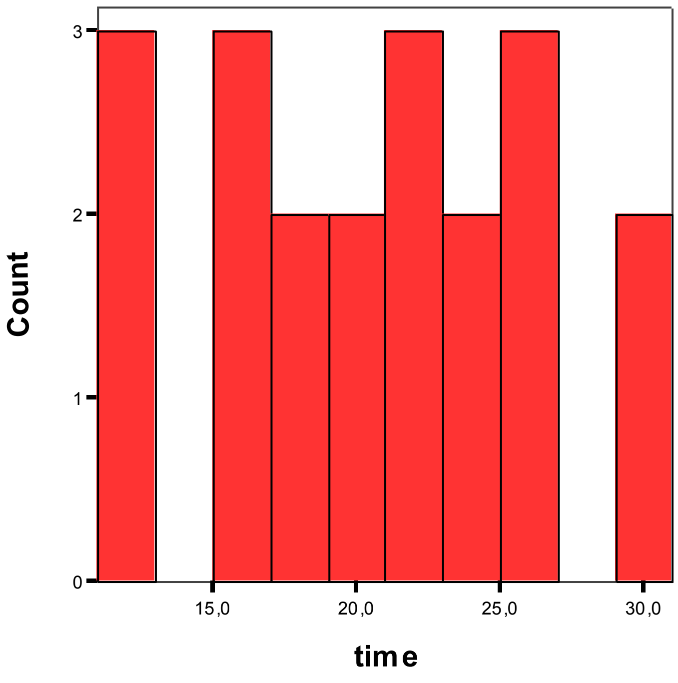
Παράδειγμα

Χρόνος ημι-αστικής διαδρομής συγκεκριμένου μήκους με το ίδιο μέσο μεταφοράς
Δείγμα από 20 διαφορετικές περιοχές σε μια πόλη

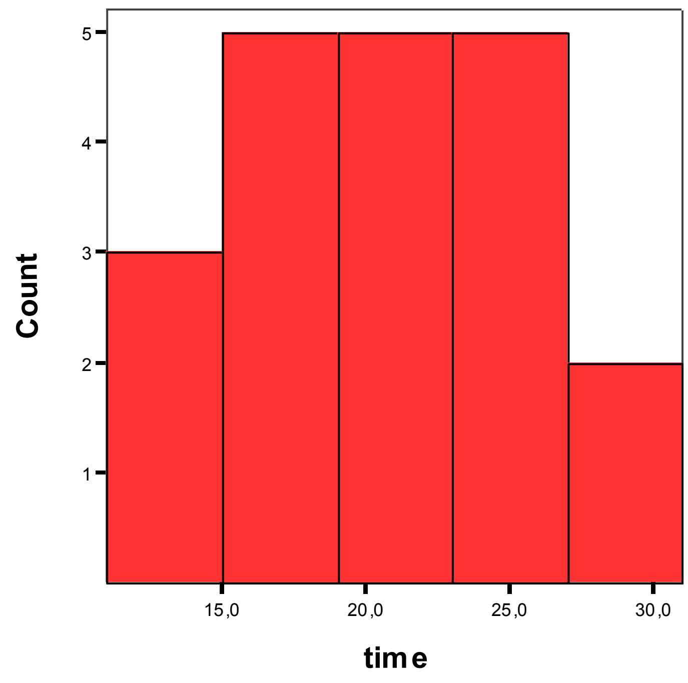
27 24 20 22 16 30 20 31 22 25 11 16 27 16 23 18 19 12 12 26

Ιστογράμματα

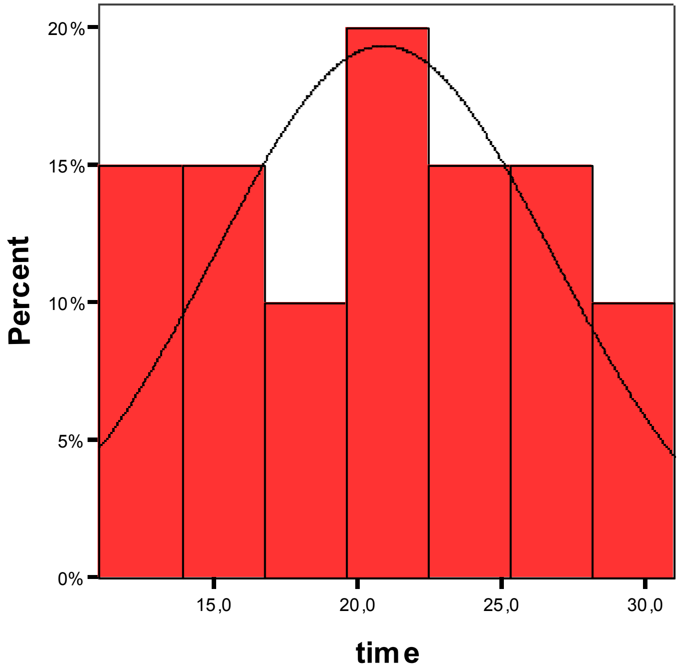
10 ομάδες



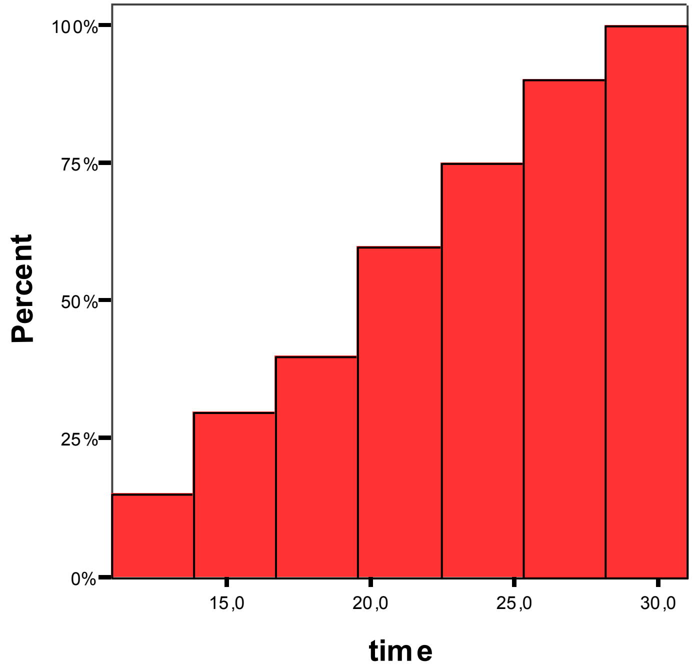
5 ομάδες



Ιστόγραμμα συχνοτήτων



Ιστόγραμμα αθροιστικών συχνοτήτων



Ερωτήματα:

Τι ποσοστό των χρόνων είναι μεγαλύτερο από 25 min;

Μαζεύονται οι χρόνοι γύρω από κάποια κεντρική τιμή;

Είναι η κατανομή των χρόνων **κανονική**;

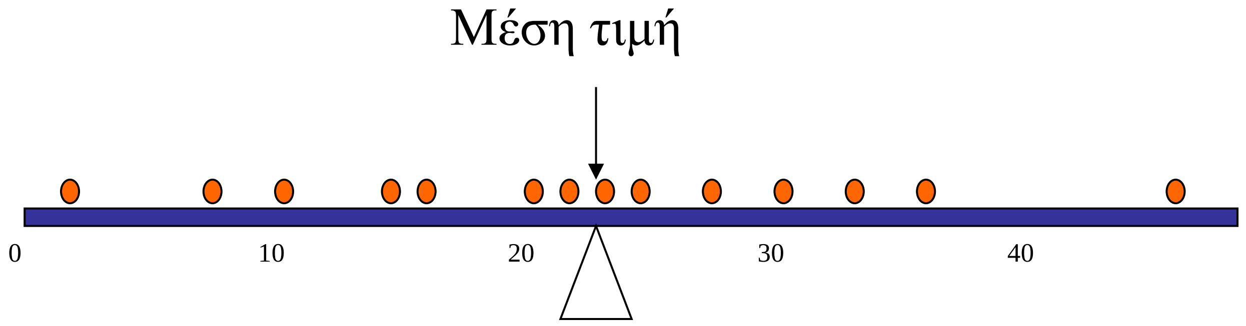
Συνοπτικά μέτρα στατιστικών δεδομένων

Μέτρα θέσης (μέτρα κεντρικής τάσης)

$x_1, x_2, \ldots, x_n$, παρατηρήσεις του δείγματος για τ.μ. X

- **δειγματική μέση τιμή** $\bar{x}$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



- **δειγματική διάμεσος** $\tilde{x}$

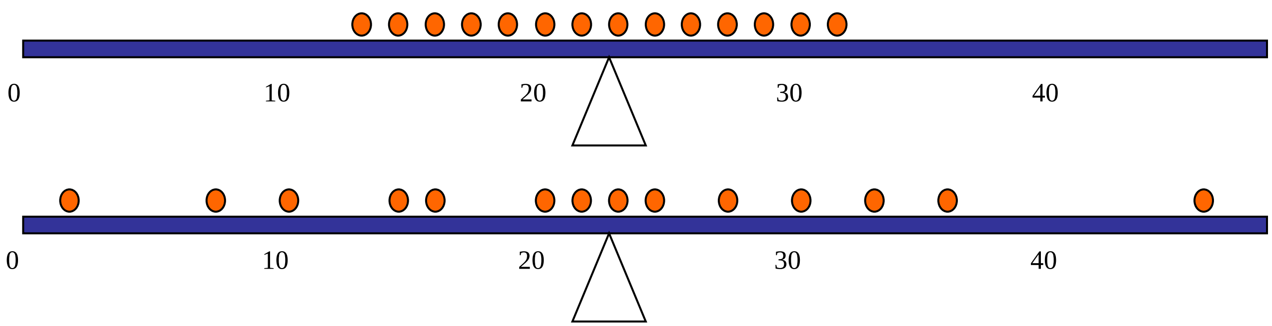
$x_1, x_2, \ldots, x_n$, σε αύξουσα σειρά

n περιττός ➔ $\tilde{x}$ είναι η τιμή στη θέση $(n+1)/2$,

n άρτιος ➔ $\tilde{x}$ είναι το ημιάθροισμα των τιμών στις θέσεις $n/2$ και $n/2+1$

- ☆ $\bar{x}$ χρησιμοποιεί τις τιμές $x_1, x_2, \ldots, x_n$,
- ☆ $\tilde{x}$ χρησιμοποιεί την τάξη των $x_1, x_2, \ldots, x_n$,
- ☆ $\bar{x}$ επηρεάζεται από μακρινές τιμές, όχι η $\tilde{x}$

Μέτρα μεταβλητότητας



- δειγματική διακύμανση (διασπορά) s^2

δειγματική τυπική απόκλιση s

απόκλιση μιας παρατήρησης x_i από $\bar{x}$: $x_i - \bar{x}$

Άθροισμα αποκλίσεων:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Άθροισμα τετραγώνων αποκλίσεων

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

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

$x_1, x_2, \ldots, x_n$, σε αύξουσα σειρά

- **εκατοστιαία σημεία**

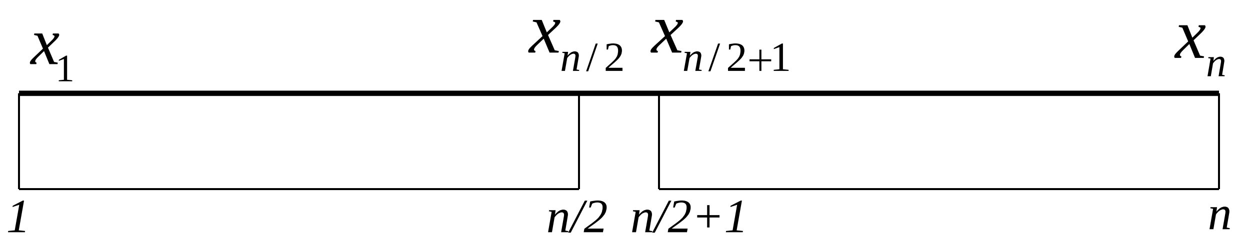
$\tilde{x}$ είναι το 50-εκατοστιαίο σημείο

Q_1: 25-εκατοστιαίο σημείο, **πρώτο τεταρτομόριο**

Q_3: 75-εκατοστιαίο σημείο, **τρίτο τεταρτομόριο**

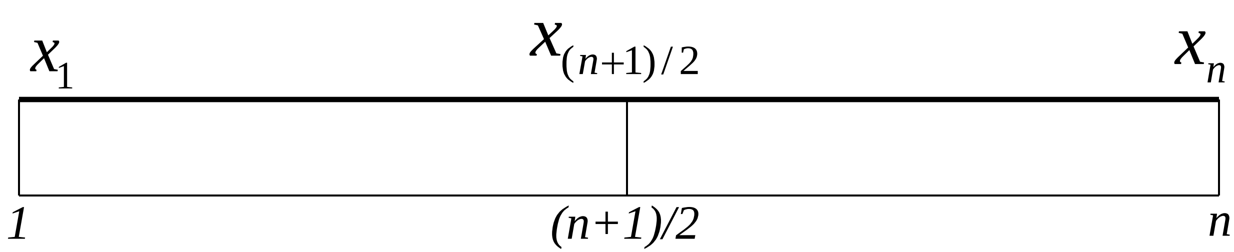
$I = Q_1 - Q_3$: **ενδοτεταρτομοριακό εύρος**

$n = 2k$



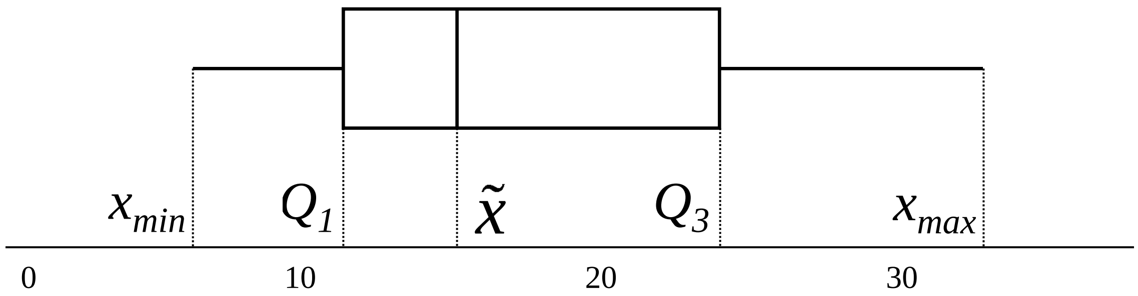
Q_1: διάμεσος των $x_1, \ldots, x_{n/2}$ Q_3: διάμεσος των $x_{n/2+1}, \ldots, x_n$

$n = 2k+1$



Q_1: διάμεσος των $x_1, \ldots, x_{(n+1)/2}$ Q_3: διάμεσος των $x_{(n+1)/2}, \ldots, x_n$

Σύνοψη 5 αριθμών, θηκόγραμμα



μακρινή τιμή x_i	
ύποπτη απόμακρη τιμή (outlier)	SPSS → o
απόμακρη τιμή (extreme)	SPSS → *

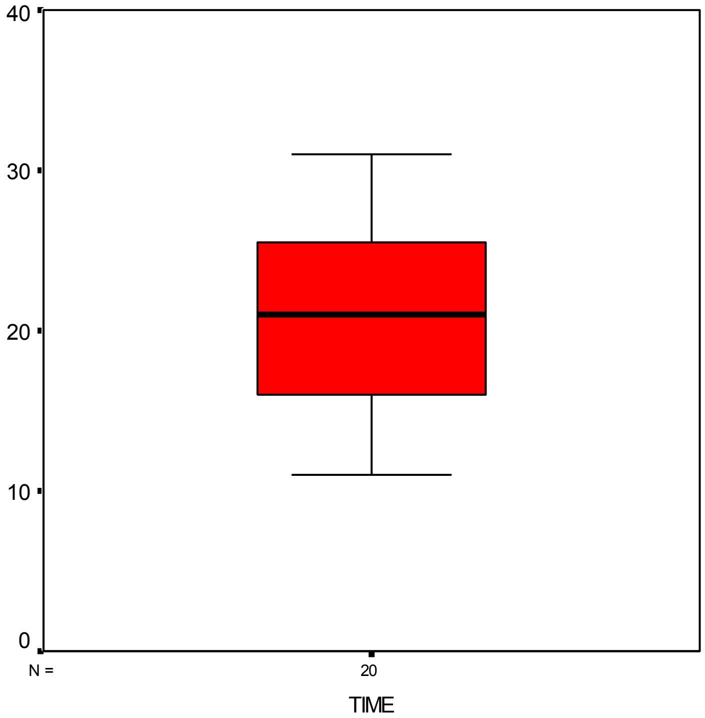
Παράδειγμα

Χρόνος ημι-αστικής διαδρομής συγκεκριμένου μήκους με το ίδιο μέσο μεταφοράς
Δείγμα από 20 διαφορετικές περιοχές σε μια πόλη

11 12 12 16 16 16 18 19 20 20 22 22 23 24 25 26 27 27 30 31

11 12 12 16 16 16 18 19 20 20 22 22 23 24 25 26 27 27 30 31

11 16 21 25.5 31



Κανονική κατανομή:

1. Συμμετρικότητα
2. Όχι μακριές ουρές
3. Όχι απόμακρα σημεία

Εκτιμητική

Γενικά

πληθυσμός

τ.μ. X

παράμετρος θ

μέση τιμή, $\boldsymbol{\mu}$

δείγμα $\{x_1, x_2, \ldots, x_n\}$

στατιστική $\hat{\theta}$ (σημειακή εκτίμηση)

διάστημα εμπιστοσύνης $[\theta_1, \theta_2]$

μέση τιμή

$\mu \equiv \mathrm{E}[X]$

δειγματική μέση τιμή

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Εκτίμηση μέσης τιμής

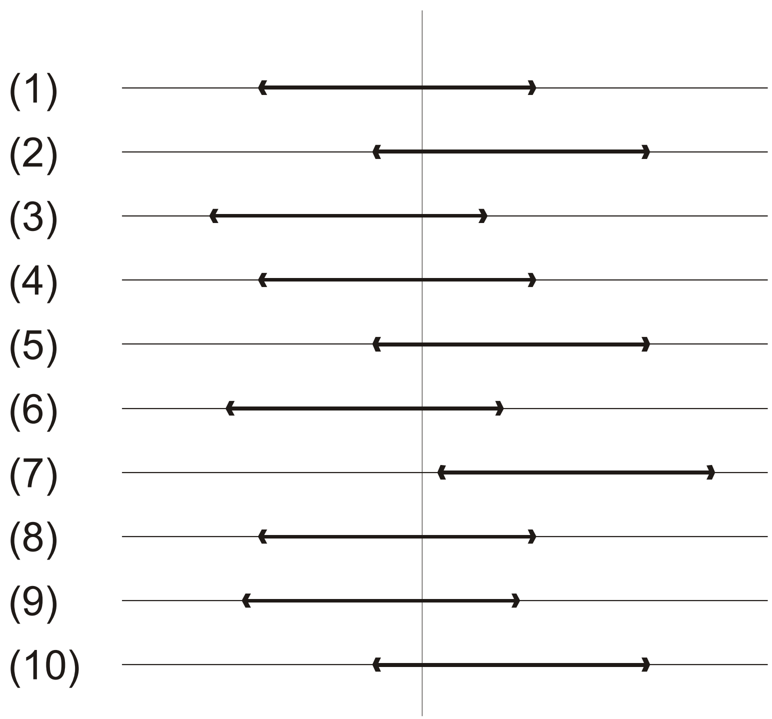
διάστημα εμπιστοσύνης (δ.ε.) $[\mu_1, \mu_2]$

σε ***επίπεδο εμπιστοσύνης*** $1\text{-}\alpha$ (α: ***επίπεδο σημαντικότητας***)

$$\mathrm{P}\left(\mu \in [\mu_1, \mu_2]\right) = 1 - \alpha \qquad \mathrm{P}\left(\mu \notin [\mu_1, \mu_2]\right) = \alpha$$

$\overline{x}$ αλλάζει με το δείγμα ➔ $\overline{x}$ είναι τ.μ.

Πραγματική τιμή της μ



(1-α)% διάστημα εμπιστοσύνης για μ

$$\boxed{\overline{x} \pm (\text{κρίσιμη τιμή}) \cdot s_{\overline{x}}}$$

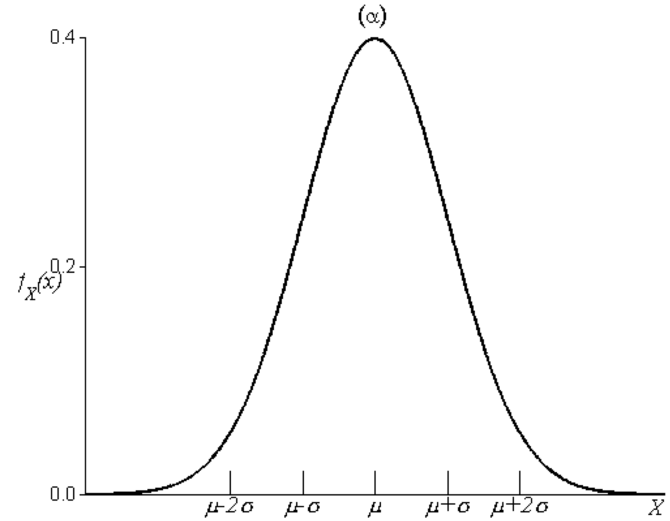
εκτίμηση της τυπικής απόκλισης του $\overline{X}$ ή τυπικό σφάλμα του $\overline{X}$

Κρίσιμες τιμές ορίζονται από τις ουρές κάποιας βασικής κατανομής

σ^2 γνωστό

$$n < 30 \text{ και } X \sim N\left(\mu, \sigma^2\right) \Rightarrow \quad z \equiv \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

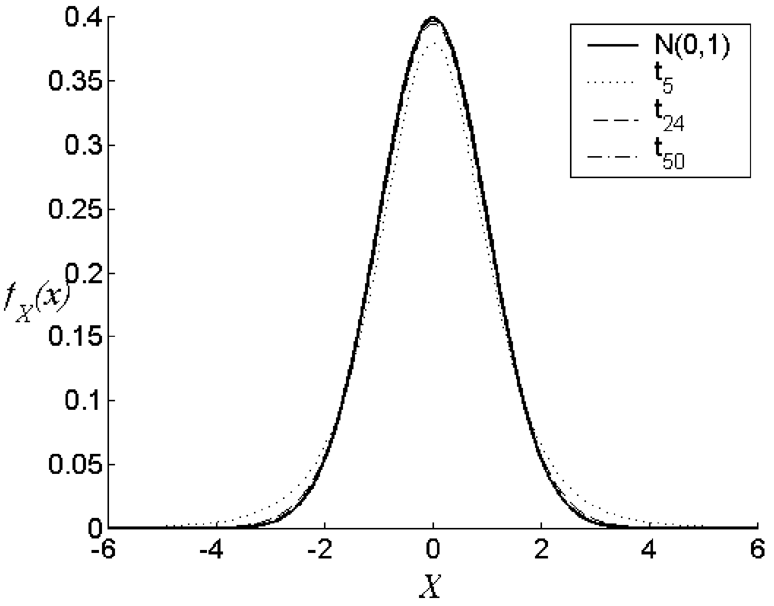
Τυπική κανονική κατανομή



σ^2 άγνωστο

$$n < 30 \text{ και } X \sim N\left(\mu, \sigma^2\right) \Rightarrow \quad t \equiv \frac{\bar{x} - \mu}{s / \sqrt{n}} \sim t_{n-1}$$

Κατανομή student

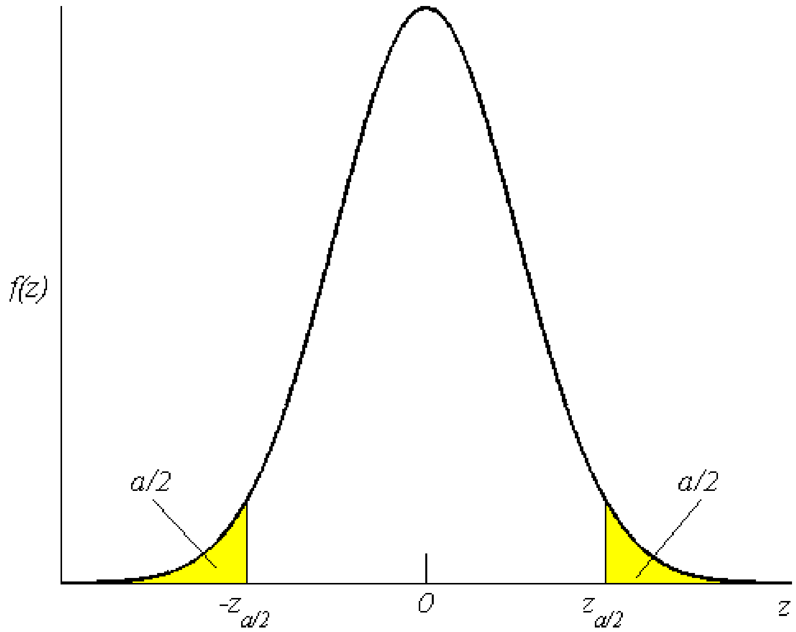


Τυπική κανονική κατανομή

επίπεδο εμπιστοσύνης $1\text{-}\alpha$

$z_{1-\alpha/2}$: η $1\text{-}a/2$ *κρίσιμη τιμή της τυπικής κανονικής κατανομής*

$$a = 0.05 \Longrightarrow z_{1-\alpha/2} = z_{0.975} = 1.96$$



$$\mathrm{P}\left(-z_{1-\alpha/2} < z < z_{1-\alpha/2}\right) = 1 - \alpha$$

$(1\text{-}\alpha)\%$ διάστημα εμπιστοσύνης για μ

Συνθήκες

1. μέγεθος του δείγματος n: μεγάλο / μικρό
2. κανονική κατανομή της τ.μ. X : ναι / όχι
3. διασπορά σ^2: γνωστή / άγνωστη

Γενική μορφή

$$\bar{x} \pm (\text{κρίσιμη τιμή}) \cdot s_{\bar{x}}$$

(1-α)% διάστημα εμπιστοσύνης της μέσης τιμής

Διασπορά	*κατανομή της X*	*N*	*διάστημα εμπιστοσύνης*
γνωστή	κανονική		$\bar{x} \pm z_{1-\alpha/2} \dfrac{\sigma}{\sqrt{n}}$
γνωστή	μη κανονική	$n > 30$	$\bar{x} \pm z_{1-\alpha/2} \dfrac{\sigma}{\sqrt{n}}$
γνωστή	μη κανονική	$n < 30$	μη-παραμετρικό
άγνωστη		$n > 30$	$\bar{x} \pm z_{1-\alpha/2} \dfrac{s}{\sqrt{n}}$
άγνωστη	κανονική	$n < 30$	$\bar{x} \pm t_{n-1,1-\alpha/2} \dfrac{s}{\sqrt{n}}$
άγνωστη	μη κανονική	$n < 30$	μη-παραμετρικό

Παράδειγμα

Χρόνος ημι-αστικής διαδρομής συγκεκριμένου μήκους με το ίδιο μέσο μεταφοράς
Δείγμα από 20 διαφορετικές περιοχές σε μια πόλη

11 12 12 16 16 16 18 19 20 20 22 22 23 24 25 26 27 27 30 31

Κανονική κατανομή, άγνωστη διασπορά, μικρό μέγεθος δείγματος

$$\left[\bar{x} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

Πίνακας κατανομής student

$$n-1=19$$

$$\alpha = 0.05 \Rightarrow 1-\alpha/2 = 0.975$$

$$t_{19,0.975} = 2.093$$

$$\bar{x} = 20.85$$

$$s = 5.896$$

$$s^2 = 34.766$$

$$\left[20.85 - 2.093 \cdot \frac{5.896}{\sqrt{20}}, 20.85 + 2.093 \cdot \frac{5.896}{\sqrt{20}}\right] = [18.090, 23.610]$$

Με εμπιστοσύνη σε επίπεδο 95% μπορούμε να πούμε πως ο μέσος χρόνος διαδρομής είναι μεταξύ 18.09min και 23.61min

Εκτίμηση διαφοράς μέσων τιμών

δύο ανεξάρτητων τ.μ. X_1 και X_2

Μέσες τιμές: μ_1 και μ_2

Διασπορές: σ_1^2 και σ_2^2

Δείγμα μεγέθους n_1, n_2 από τον πληθυσμό της X_1, X_2

Δειγματικές μέσες τιμές: $\bar{x}_1$ και $\bar{x}_2$

Δειγματικές διασπορές: s_1^2 και s_2^2

Θέλουμε να εκτιμήσουμε τη διαφορά των μέσων τιμών μ_1 και μ_2

Σημειακή εκτίμηση της διαφοράς μ_1 - μ_2: $\bar{x}_1 - \bar{x}_2$

Διάστημα εμπιστοσύνης για τη διαφορά μέσων τιμών

διασπορές των X_1, X_2	*Κατανομές των X_1, X_2*	*n_1, n_2*	*διάστημα εμπιστοσύνης για $\mu_1 - \mu_2$*
γνωστές	κανονική		$\left(\bar{x}_1 - \bar{x}_2\right) \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
γνωστές	μη κανονική	μεγάλα	$\left(\bar{x}_1 - \bar{x}_2\right) \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
γνωστές	μη κανονική	μικρά	μη-παραμετρικό
άγνωστες άνισες ή ίσες		μεγάλα	$\left(\bar{x}_1 - \bar{x}_2\right) \pm z_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
άγνωστες και ίσες	κανονική	μικρά	$\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{n_1+n_2-2,1-\alpha/2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
άγνωστες και ίσες	μη κανονική	μικρά	μη-παραμετρικό
άγνωστες και άνισες		μικρά	----

Εκτίμηση κοινής διασποράς $$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

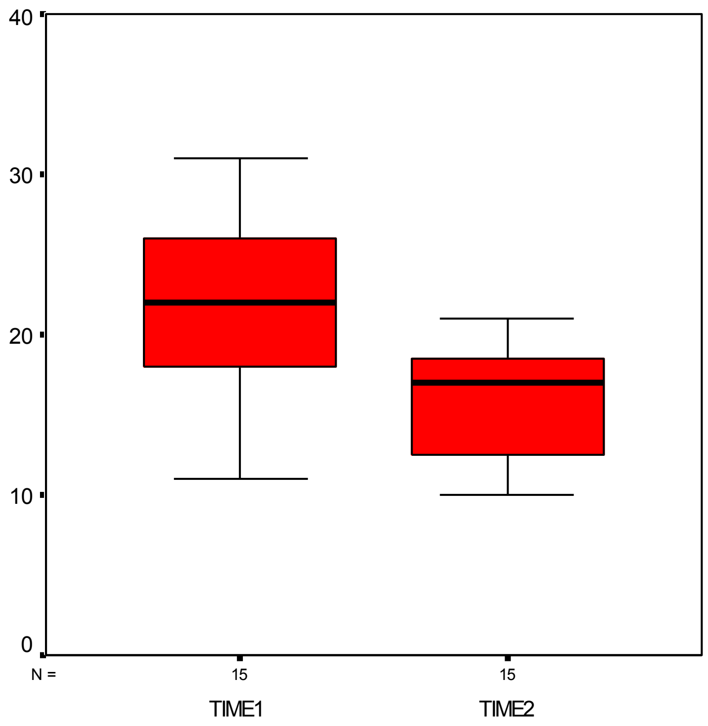
Παράδειγμα

Χρόνος ημι-αστικής διαδρομής συγκεκριμένου μήκους με το ίδιο μέσο μεταφοράς
Δείγμα από 20 διαφορετικές περιοχές σε μια πόλη Α

11 12 12 16 16 16 18 19 20 20 22 22 23 24 25 26 27 27 30 31

Δείγμα από 15 διαφορετικές περιοχές σε μια πόλη Β

10 10 11 11 14 15 15 17 17 17 18 19 19 19 21



X_1: χρόνος διαδρομής για πόλη Α

X_2: χρόνος διαδρομής για πόλη Β

- Κανονική κατανομή για X_1 και X_2 ?
- Διασπορές των X_1 και X_2 περίπου ίδιες ?
- Μέσος χρόνος διαδρομής μεγαλύτερος στην πόλη Α ?

Κανονικές κατανομές, άγνωστες αλλά ίσες διασπορές, μικρά δείγματα

$$\left(\bar{x}_1-\bar{x}_2\right)\pm t_{n_1+n_2-2,1-\alpha/2}\cdot s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$\left.\begin{aligned}\bar{x}_1&=20.85\\ \bar{x}_2&=15.53\end{aligned}\right\}\quad \bar{x}_1-\bar{x}_2=5.32$$

Πίνακας κατανομής student

$$\left.\begin{aligned}n_1+n_2-2&=33\\ \alpha=0.05\Rightarrow 1-\alpha/2&=0.975\end{aligned}\right\}\quad t_{33,0.975}=2.04$$

$$s_p=5.058$$

$$\left.\begin{aligned}s_1^2&=34.766\\ s_2^2&=13.124\end{aligned}\right\}\quad s_p^2=\frac{(20-1)34.766+(15-1)13.124}{20+15-2}=25.584$$

$$5.32\pm 2.04\cdot 5.058\cdot\sqrt{\frac{1}{20}+\frac{1}{15}}\quad\Longrightarrow\quad [1.80,\ 8.83]$$

Σε επίπεδο εμπιστοσύνης 95%, ο χρόνος διαδρομής στην πόλη Α είναι μεγαλύτερος από το χρόνο διαδρομής στην πόλη Β κατά 1.8min με 8.8min

Έλεγχος Υπόθεσης

Διαφέρουν οι μέσες τιμές μ_1 και μ_2;

Έλεγχος στατιστικής υπόθεσης

Μηδενική υπόθεση H_0: $\mu_1 = \mu_2$ Εναλλακτική υπόθεση H_1: $\mu_1 \neq \mu_2$

Στατιστική ελέγχου q με γνωστή κατανομή $q \equiv z \sim \mathrm{N}(0,1)$ $q \equiv t \sim \mathrm{t}_{n_1+n_2-2}$

Απορριπτική περιοχή R
(απίθανες τιμές της q όταν ισχύει η H_0)

$$R = \left\{ |z| > z_{1-\alpha/2} \right\}$$

α: επίπεδο σημαντικότητας

Δειγματική στατιστική ελέγχου $\tilde{q}$

$\tilde{q} \in R$ ⇨ Απόρριψη της H_0

$\tilde{q} \notin R$ ⇨ Μη απόρριψη της H_0

Πως ορίζεται η q?

Όπως για τα διαστήματα εμπιστοσύνης

p-τιμή: χαμηλότερη στάθμη σημαντικότητας α που μπορούμε να απορρίψουμε την H_0

$$p = P\left(|z| > z_{1-p/2} \right)$$

Προηγούμενο παράδειγμα

$$\mathrm{H}_0:\ \mu_1 = \mu_2 \qquad \Rightarrow \qquad \mathrm{H}_0:\ \mu_1 - \mu_2 = 0$$

$$\mathrm{H}_1:\ \mu_1 \neq \mu_2 \qquad\qquad\quad \mathrm{H}_1:\ \mu_1 - \mu_2 \neq 0$$

$$q \equiv t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{n_1+n_2-2,1-\alpha/2}$$

$$R = \left\{ |t| > t_{n_1+n_2-2,1-\alpha/2} \right\}$$

$$R = \left\{ |t| > 2.04 \right\}$$

$$\tilde{q} = \frac{5.32}{5.058 \cdot \sqrt{\dfrac{1}{20} + \dfrac{1}{15}}} = 3.077$$

$$|\tilde{q}| = 3.077 > 2.04 \ \Rightarrow \ \tilde{q} \in R$$

⇒ Απόρριψη της H_0

Σε επίπεδο εμπιστοσύνης 95%, ο χρόνος διαδρομής στην πόλη Α είναι μεγαλύτερος από το χρόνο διαδρομής στην πόλη Β

Ανάλυση Διασποράς

Σχεδιασμός πειράματος

Μελέτη χρόνου διαδρομής

μεταβλητή ενδιαφέροντος

- μέσο μεταφοράς
- ημιαστικές περιοχές
- πόλεις
- …

παράγοντας

Μονόδρομη ανάλυση διασποράς

Πλήρως τυχαιοποιημένος σχεδιασμός

Μεταβλητή ενδιαφέροντος Y

Ένας παράγοντας με k ομάδες

Δείγμα 1: $y_{11}, y_{12}, \ldots, y_{1n_1}$

Δείγμα 2: $y_{21}, y_{22}, \ldots, y_{2n_2}$

Δείγμα k: $y_{k1}, y_{k2}, \ldots, y_{kn_k}$

Υποθέτουμε πως η Y στις k ομάδες ακολουθεί κανονική κατανομή με κοινή διασπορά άρα οι k ομάδες μπορεί να διαφέρουν μόνο ως προς τη μέση τιμή τους

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

Μονόδρομη Ανάλυση Διασποράς (ANOVA)

$\bar{y}$: μέσος όρος από όλα τα δείγματα

$\bar{y}_i$: μέσος όρος στο δείγμα i

$$y_{ij} \quad = \quad \bar{y} \quad + \quad (\bar{y}_i - \bar{y}) \quad + \quad (y_{ij} - \bar{y}_i)$$

παρατήρηση ολικός μέσος απόκλιση λόγω παράγοντα τυχαίο σφάλμα

SSA: $\sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}\right)^2$ SSE: $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}\right)^2$

σφάλμα μεταξύ δειγμάτων σφάλμα μέσα σε κάθε δείγμα

Στατιστική ελέγχου q : $\boxed{F = \dfrac{\mathrm{SSA}/(k-1)}{\mathrm{SSE}/(n-k)} \sim F_{k-1,n-k}}$ ← Fisher κατανομή

Απορριπτική περιοχή $\boxed{R = \{F > F_{k-1,n-k;1-a}\}}$

Δειγματική στατιστική ελέγχου $\tilde{F}$ $\tilde{F} > F_{k-1,n-k;1-a}$ ⇨ Απόρριψη της H_0

Συνέχεια της ανάλυσης όταν απορρίπτεται η H_0

Ποιες ομάδες διαφέρουν; → Πολλαπλές συγκρίσεις ζευγαριών ομάδων

Έλεγχος ορθότητας των υποθέσεων

Υποθέσεις:

1. Η κατανομή της Y στις k ομάδες είναι κανονική
2. Η διασπορά της Y στις k ομάδες είναι ίδια

Τα θηκογράμματα μπορούν να δείξουν αν οι υποθέσεις είναι σωστές

Παράδειγμα

Χρόνος ημι-αστικής διαδρομής συγκεκριμένου μήκους με το ίδιο μέσο μεταφοράς

Δείγμα από 20 διαφορετικές περιοχές σε μια πόλη Α

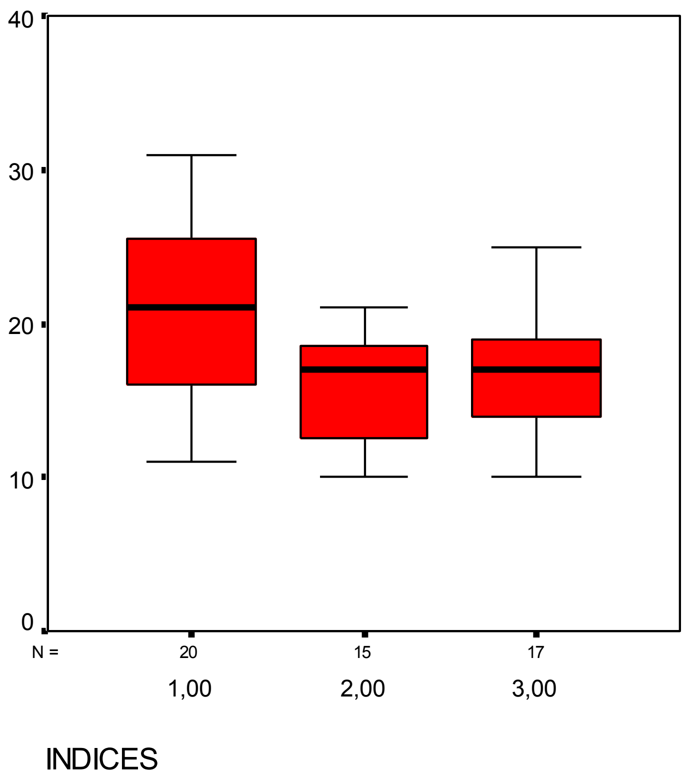
11 12 12 16 16 16 18 19 20 20 22 22 23 24 25 26 27 27 30 31

Δείγμα από 15 διαφορετικές περιοχές σε μια πόλη Β

10 10 11 11 14 15 15 17 17 17 18 19 19 19 21

Δείγμα από 17 διαφορετικές περιοχές σε μια πόλη Γ

10 12 12 14 14 15 15 17 17 18 18 19 19 21 23 25 25



- Κανονική κατανομή ?
- Ίσες διασπορές ?
- Μέσος χρόνος διαδρομής μεγαλύτερος στην πόλη Α από ότι στις άλλες δύο πόλεις ?

$$n = 20 + 15 + 17 = 52$$

Κρίσιμη τιμή

Απορριπτική περιοχή

$$\left.\begin{array}{l} k-1=2 \\ n-k=49 \\ 1-\alpha=0.95 \end{array}\right\} \quad F_{2,49;0.95}=3.187 \quad \Rightarrow \quad R=\{F>3.187\}$$

Δειγματική στατιστική ελέγχου

$$\left.\begin{array}{l} \text{SSA} = 260.956 \\ \text{SSE} = 1157.813 \end{array}\right\} \quad \tilde{F} = \frac{260.956/2}{1157.813/49} = 5.522$$

$\tilde{F} = 5.522 > 3.187 \Rightarrow \tilde{F} \in R$ ⇒ Απόρριψη της H_0

> Σε επίπεδο εμπιστοσύνης 95%, ο χρόνος διαδρομής δεν είναι ίδιος στις τρεις πόλεις

Σε ποιες πόλεις διαφέρει ο χρόνος διαδρομής; ⇒ Πολλαπλές συγκρίσεις

Ανάλυση Διασποράς με Δύο Παράγοντες

Δύο παράγοντες

πόλη (Α, Β, Γ)

μέσο μεταφοράς (λεωφορείο, αυτοκίνητο, άλλο)

Ερωτήματα:

- Επηρεάζει ο πρώτος παράγοντας από μόνος του τη μεταβλητή ενδιαφέροντος;
- Επηρεάζει ο δεύτερος παράγοντας από μόνος του τη μεταβλητή ενδιαφέροντος;
- Έχουν οι δύο παράγοντες μαζί συνδυασμένη επίδραση στη μεταβλητή ενδιαφέροντος;

Μια μηδενική υπόθεση για κάθε ερώτημα

Ανάλυση διασποράς δύο δρόμων