

**MEM-205 Περιγραφική Στατιστική**  
Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

27-02-2020

Στη συνέχεια θα περιγράψουμε το διάστημα εμπιστοσύνης για την μέση τιμή του πληθυσμού στις ακόλουθες περιπτώσεις:

1. Η μεταβλητή  $X$  ακολουθεί κανονική κατανομή
  2. Η μεταβλητή  $X$  δεν ακολουθεί κανονική κατανομή
    - Σε αυτή τη περίπτωση υποθέτουμε ότι το δείγμα είναι αρκετά μεγάλο ( $n \geq 30$ )
- ▶ Επίσης θα εξετάσουμε χωρίστα αν γνωρίζουμε την τυπική απόκλιση  $\sigma$  ή όχι.
  - ▶ Όταν το  $\sigma$  είναι άγνωστο χρειαζόμαστε τη t-κατανομή.

- ▶ όταν το  $\sigma$  είναι γνωστό θα έχουμε

### Τυπική απόκλιση της $\bar{X}$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

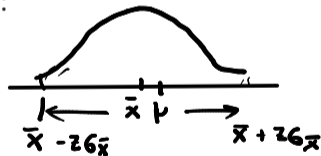
### Διάστημα εμπιστοσύνης της $\mu$

Το  $(1 - \alpha) * 100\%$  διάστημα εμπιστοσύνης για την  $\mu$  είναι:

$$[\bar{X} - z\sigma_{\bar{X}}, \bar{X} + z\sigma_{\bar{X}}]$$

όπου το  $z$  (z-score) λαμβάνεται έτσι ώστε

$$P(Z < z) = 1 - \alpha/2$$



- ▶ Περιθώριο σφάλματος:  $E = z\sigma_{\bar{X}}$

## t-Κατανομή (t-distribution)

- ▶ Είναι γνωστή και ως Student's t distribution και σχετίζεται με την τυπική κανονική κατανομή.
- ▶ Όπως και η τυπική κανονική κατανομή η t-κατανομή είναι συμμετρική γύρω από το μηδέν, έχει καμπανοειδη μορφή και η συνάρτηση πυκνότητας πιθανότητας είναι παντού θετική.
- ▶ Παρουσιάζει μεγαλύτερη διασπορά τιμών σε σχέση τη τυπική κανονική κατανομή.
- ▶ Η μορφή της εξαρτάται από το μέγεθος του δείγματος  $N$ . Μάλιστα η μοναδική παράμετρος της συμβολίζεται με  $df$  και είναι άμεσα συνδεδεμένη με το  $N$ .

$$df = N - 1 \quad (\text{βαθμοί ελευθερίας})$$

- ▶ Όσο το  $df$  αυξάνει η t-κατανομή προσεγγίζει όλο και περισσότερο την τυπική κανονική κατανομή.

## t-Κατανομή (t-distribution)

- ▶ Την t-κατανομή με  $df$  βαθμούς ελευθερίας θα την συμβολίζουμε ως  $t_{df}$
- ▶ Συνάρτηση πυκνότητας πιθανότητας

$$p(t) = \frac{\Gamma(\frac{df+1}{2})}{\sqrt{df\pi}\Gamma(\frac{df}{2})} \left(1 + \frac{t^2}{df}\right)^{-\frac{df+1}{2}}$$

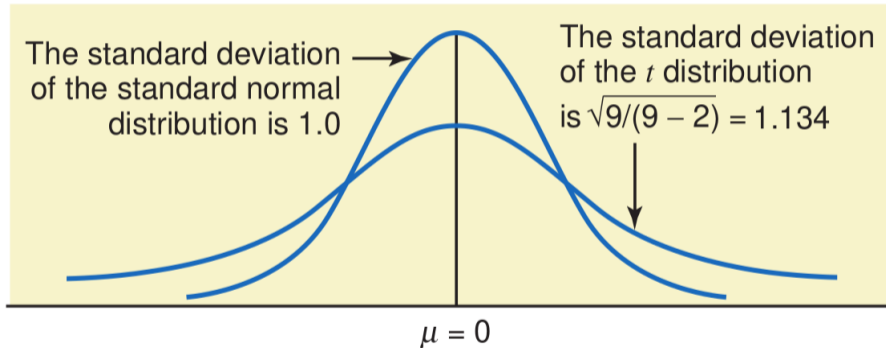
- ▶ Μέση τιμή

$$\mathbb{E}(T) = 0$$
$$\int_{-\infty}^{+\infty} t p(t) dt = 0$$

- ▶ Διασπορά

$$\mathbb{V}(T) = df * (df - 2)^{-1}$$
$$\mathbb{V}(T) = \int_{-\infty}^{+\infty} (t - \mathbb{E}(T))^2 p(t) dt = df(df-2)$$
$$\sigma = \sqrt{\mathbb{V}(T)} = \sqrt{df(df-2)^{-1}}$$

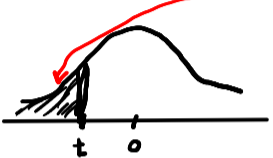
## t-Κατανομή (t-distribution)



$$s = \sqrt{\frac{N-1}{N-3}} \xrightarrow{N \rightarrow \infty} 1$$

# t-Κατανομή (t-distribution)

$P(T \leq t)$



Έστω  $P(T \leq t) = 0.25$

και  $df = 5$

$t = ?$   $t = -0.727$

Έστω  $P(T \leq t) = 0.05$

και  $df = 5000$

$t = -1.645$

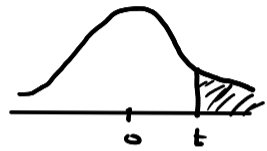
Τιμές του t

πιθανότητες

	cum. prob	$t_{.50}$	$t_{.25}$	$t_{.20}$	$t_{.15}$	$t_{.10}$	$t_{.05}$	$t_{.025}$	$t_{.01}$	$t_{.005}$	$t_{.001}$	$t_{.0005}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001	0.001
df												
2	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62	
3	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599	
4	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924	
5	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610	
6	0.000	-0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869	
7	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959	
8	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408	
9	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041	
10	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781	
11	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587	
12	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437	
13	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318	
14	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221	
15	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140	
16	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073	
17	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015	
18	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965	
19	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922	
20	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883	
21	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850	
22	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819	
23	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792	
24	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768	
25	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745	
26	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725	
27	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707	
28	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690	
29	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674	
30	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659	
40	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646	
60	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551	
80	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460	
100	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416	
1000	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390	
5000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300	
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291	
		0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%

Confidence Level

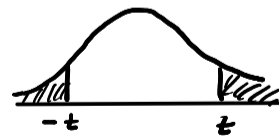
$P(T \geq t)$



Έστω  $P(T \geq t) = 0.01$

και  $df = 11$

$t = 2.718$



# Υπενθύμιση του πίνακα των z-scores

Standard Normal Probabilities

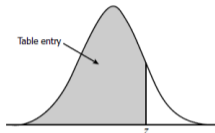


Table entry for  $z$  is the area under the standard normal curve to the left of  $z$ .

Τ. της  $z$

→

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

↑  
Τ. της  $z$

↑  
πιθανότητες



- ▶ όταν το  $\sigma$  δεν είναι γνωστό δεν μπορούμε να υπολογίσουμε το  $\sigma_{\bar{X}}$

### Εκτιμητήρια της τυπικής απόκλισης της $\bar{X}$

$$s_{\bar{X}} = \frac{s}{\sqrt{N}}$$

### Διάστημα εμπιστοσύνης της $\mu$

Το  $(1 - \alpha) * 100\%$  διάστημα εμπιστοσύνης για την  $\mu$  είναι:

$$[\bar{X} - ts_{\bar{X}}, \bar{X} + ts_{\bar{X}}]$$

όπου το  $t$  λαμβάνεται από την  $t_{df}$ ,  $df = N - 1$  έτσι ώστε

$$P(T < t) = 1 - \alpha/2$$

- ▶ Περιθώριο σφάλματος:  $E = ts_{\bar{X}}$

## Παράδειγμα

Έστω ότι η μεταβλητή  $X$  ακολουθεί κανονική κατανομή. Έστω επίσης ότι για ένα δείγμα με 25 στοιχεία λάβαμε:

$$\bar{X} = 186, \quad s = 12$$

1. Κατασκευάστε το 95 % διάστημα εμπιστοσύνης για την μέση τιμή  $\mu$ .
2. Εάν για τη μελέτη μας το περιθώριο του σφάλματος θεωρείται μεγάλο τι θα μπορούσαμε να κάνουμε για να το μειώσουμε;
3. τι θα άλλαζε αν γνωρίζαμε ότι  $\sigma = 12$ .

$$\textcircled{1} \quad s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{12}{5} = 2.4 \quad \bar{X} \pm t s_{\bar{x}} \quad df = n - 1 = 24 \quad \begin{matrix} (1 - \alpha) \cdot 100\% \\ \alpha = 0.05 \end{matrix}$$

Απο τον πίνακα έχουμε  $t = 2.064$

$$\mu \in [186 - 2.064 \cdot 2.4, 186 + 2.064 \cdot 2.4] \quad \text{με πιθανότητα } 0.95$$

$$\mu \in [181.0464, 190.9536]. \quad \left. \begin{array}{l} \textcircled{2} \text{ μεγαλύτερο } N \\ \text{ή} \\ \text{μικρότερο ποσοστό.} \end{array} \right\}$$

$$\bar{x} = 186 \quad \sigma = 12 \quad 95\% \quad (1 - \alpha) \cdot 100\%$$

$$\alpha = 0.05$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{5} = 2.4$$

$$\mu \in [\bar{x} - z \sigma_{\bar{x}}, \bar{x} + z \sigma_{\bar{x}}]$$

$$[186 - 1.960 \cdot 2.4, 186 + 1.960 \cdot 2.4] = [181.296, 190.704]$$

με πιθανότητα 0.95

Δύο μεταβλητές που αναφέρονται στα ίδια στοιχεία λέμε ότι σχετίζονται αν κάποιες τιμές της μια μεταβλητής τείνουν να εμφανίζονται πιο συχνά όταν η δεύτερη μεταβλητή λαμβάνει συγκεκριμένες τιμές.

### Εξαρτημένη μεταβλητή

Ονομάζεται η μεταβλητή για την οποία θέλουμε να περιγράψουμε και να εξηγήσουμε την συμπεριφορά της. Συνήθως συμβολίζεται με  $Y$ .

### Ανεξάρτητη μεταβλητή

Ονομάζεται η μεταβλητή η οποία χρησιμοποιείται για να δικαιολογήσει τις αλλαγές των τιμών της εξαρτημένης μεταβλητής. Συνήθως συμβολίζεται με  $X$ .

### Παράδειγμα

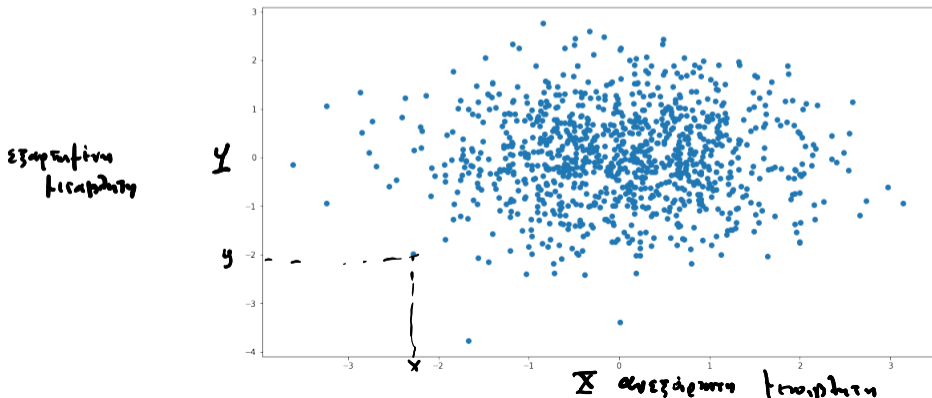
Το αλκοόλ προκαλεί πολλές παρενέργειες στον οργανισμό όπως είναι η πτώση της θερμοκρασίας. Για τη μελέτη του φαινομένου, οι ερευνητές δίνουν διαφορετικές ποσότητες αλκοόλης σε ποντίκια και έπειτα μετρούν την αλλαγή της θερμοκρασίας τους 15 λεπτά μετά τη λήψη. Η **ποσότητα της αλκοόλης** είναι η **ανεξάρτητη μεταβλητή** ενώ η **μεταβολή της θερμοκρασίας** είναι η **εξαρτημένη μεταβλητή**.

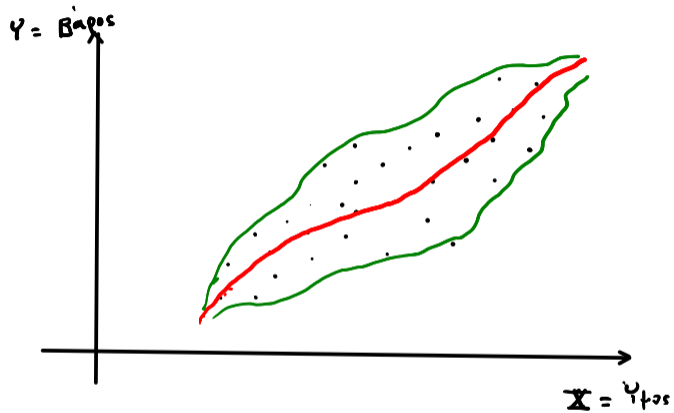
Για τη μελέτη του κατά πόσο δύο μεταβλητές συσχετίζονται, ακολουθούμε τα ακόλουθα βήματα:

- ▶ Γραφική αναπαράσταση και υπολογισμός των περιγραφικών μέτρων
- ▶ Αναγνώριση προτύπων και μελέτη των αποκλίσεων των τιμών.
- ▶ Όταν τα πρότυπα είναι αρκετά ευδιάκριτα, επιλογή κατάλληλου μαθηματικού μοντέλου για τη περιγραφή τους.

## Διάγραμμα Διασποράς (Scatter Plot)

Το **διάγραμμα διασποράς** παρουσιάζει τη σχέση μεταξύ των τιμών δύο ποσοτικών μεταβλητών που αναφέρονται στα ίδια στοιχεία. Ο οριζόντιος άξονας εκφράζει τις τιμές της μιας μεταβλητής (συνήθως της ανεξάρτητης μεταβλητής) ενώ ο κάθετος τις τιμές της άλλης μεταβλητής (συνήθως της εξαρτημένης μεταβλητής). Κάθε ζεύγος τιμών  $(x, y)$  για τα στοιχεία του πληθυσμού ή του δείγματος απεικονίζοντε με ένα συμβολο.



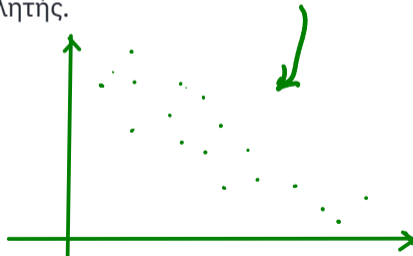
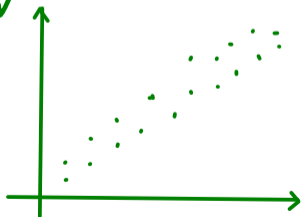


### Θετικά συσχετισμένες μεταβλητές

Όσο μεγαλύτερες τιμές μιας μεταβλητής τείνουν να συνοδεύονται με όλο και μεγαλύτερες τιμές της άλλης μεταβλητής.

### Αρνητικά συσχετισμένες μεταβλητές

Όσο μεγαλύτερες τιμές μιας μεταβλητής τείνουν να συνοδεύονται με όλο και μικρότερες τιμές της άλλης μεταβλητής.





- ▶ Αν  $X$  είναι η ανεξάρτητη μεταβλητή και  $Y$  είναι η εξαρτημένη μεταβλητή η συναρτησιακή σχέση των δύο μεταβλητών περιγράφεται μέσω μιας συνάρτησης  $f$  στη μορφή  $Y = f(X)$ .
- ▶ Για δεδομένη τιμή  $x$  της ανεξάρτητης μεταβλητής, η συνάρτηση  $f$  δίνει την αντιστοιχη τιμή  $y$  της εξαρτημένης μεταβλητής  $Y$ .
- ▶ Η  $f$  δύναται να είναι στοχαστική συνάρτηση. Σε αυτή την περίπτωση ακόμη και για ίδιες τιμές της μεταβλητής  $X$  μπορούν να προκύψουν διαφορετικές τιμές για την  $Y$ .

$$y = f(x) = x + 2$$

## Συναρτησιακή Σχέση μεταξύ Μεταβλητών

$f(x) = x + \eta$  , όπου  $\eta$  είναι το αποτέλεσμα ενός γαιριού

$$x_1 = 3 \quad \eta_1 = 3 \quad \rightarrow \quad y_1 = 3 + 3 = 6$$

$$x_2 = 5 \quad \eta_2 = 4 \quad \rightarrow \quad y_2 = 5 + 4 = 9$$

$$x_3 = 8 \quad \eta_3 = 5 \quad \rightarrow \quad y_3 = 8 + 5 = 13$$

