### Έλεγχος κανονικότητας

- Έλεγχος κανονικότητας : Επειδή πολλές μέθοδοι της Στατιστικής εφαρμόζονται σε δεδομένα που ακολουθούν την κανονική κατανομή, είναι πολύ χρήσιμο να είμαστε σε θέση να γνωρίζουμε εάν τα δεδομένα μας ακολουθούν την κανονική κατανομή ή ανήκουν σε πληθυσμό που ακολουθεί την κανονική κατανομή.
- Μέθοδοι Ελέγχου κανονικότητας :
- Θεωρητικές: Kolmogorov Smirnov, Shapiro Wilk
- Γραφικές: Γραφήματα Normal probability (PP plot), Quantile Quantile plot (QQ plot)

• Τύπος εξέτασης για ένα δείγμα κυρίως όταν είναι μεγάλο.

 Αρχική Θεώρηση, Έλεγχος συγκεκριμένης υπόθεσης: Το δείγμα μας προέρχεται από πληθυσμό με γνωστή συνάρτηση κατανομής F<sub>1</sub>(X), δηλαδή τη συνάρτηση της κανονικής κατανομής.

•  $H_0: F(X) = F_1(X)$ 

 Εναλλακτικά (εναλλακτική υπόθεση) πρέπει να υποθέσουμε ότι το δείγμα μας δεν προέρχεται από την παραπάνω συνάρτηση αλλά από κάποια άλλη συνάρτηση.

•  $H_1: F(X) \neq F_1(X)$ 

 Κατασκευάζεται συνάρτηση η οποία βασίζεται στην κατανομή του δείγματος (εμπειρική συνάρτηση)

 Για να ισχύει η υπόθεση ότι το δείγμα μας ακολουθεί την κανονική κατανομή πρέπει η εμπειρική συνάρτηση σχεδόν να ταυτίζεται με τη συνάρτηση κανονικής κατανομής, αλλιώς τα δεδομένα μας δεν προέρχονται από τον συγκεκριμένο πληθυσμό, αφού πήραμε ως δεδομένο ότι ο πληθυσμός ακολουθεί την κανονική κατανομή.

#### Από τους Kolmogorov – Smirnov:

 Ορίστηκε η απόσταση μεταξύ των δύο υποθέσεων ως το μέγιστο της απόλυτης διαφοράς της συνάρτησης κατανομής από την εμπειρική συνάρτηση.

 Ορίστηκε ένα επίπεδο σημαντικότητας α σύμφωνα με το οποίο η παραπάνω διαφορά είναι αρκετά σημαντική ώστε να γίνει αποδεκτή ή όχι η αρχική υπόθεση (το δείγμα ακολουθεί την κανονική κατανομή)

Σύμφωνα με το επίπεδο σημαντικότητας α γίνεται έλεγχος της τιμής ενός συντελεστή p (p-value ή Sig). Έχει αποδειχθεί και ισχύουν τα εξής:

- Εάν η τιμή p-value είναι μεγαλύτερη του 0.05, τότε η αρχική υπόθεση γίνεται δεκτή, δηλαδή η τυχαία μεταβλητή από την οποία προήλθε το υπό μελέτη δείγμα ακολουθεί την κανονική κατανομή.
- Εάν η τιμή p-value είναι μικρότερη του 0.05, τότε η αρχική υπόθεση απορρίπτεται, δηλαδή η τυχαία μεταβλητή από την οποία προήλθε το υπό μελέτη δείγμα δεν ακολουθεί την κανονική κατανομή.

# Έλεγχος Shapiro – Wilk

# Έλεγχος Shapiro – Wilk

• Τύπος εξέτασης για ένα δείγμα όταν είναι μικρό.

 Αρχική Θεώρηση, Έλεγχος συγκεκριμένης υπόθεσης: Το δείγμα μας προέρχεται από πληθυσμό με γνωστή συνάρτηση κατανομής F<sub>1</sub>(X), δηλαδή τη συνάρτηση της κανονικής κατανομής.

•  $H_0: F(X) = F_1(X)$ 

 Εναλλακτικά (εναλλακτική υπόθεση) πρέπει να υποθέσουμε ότι το δείγμα μας δεν προέρχεται από την παραπάνω συνάρτηση αλλά από κάποια άλλη συνάρτηση.

•  $H_1: F(X) \neq F_1(X)$ 

### Έλεγχος Shapiro – Wilk

• Τύπος εξέτασης για ένα δείγμα όταν είναι μικρό.

• Ο έλεγχος κάνει χρήση της συνάρτησης W όπου

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

• Διατάσσουμε τις τιμές του δείγματος.

• Επίδραση στη συνάρτηση των διασπορών και μέσων τιμών.

- Ένας ιατρός θέλει να ελέγξει εάν τα δεδομένα του που αφορούν το χρόνο σε μήνες που ο ασθενής χρειάζεται για να επανέλθει σε φυσιολογική κατάσταση μετά τη διάγνωση της ασθένειας ακολουθούν την κανονική κατανομή.
- Για το λόγο αυτό ο ιατρός καταγράφει το χρόνο που απαιτείται από τη στιγμή της διάγνωσης της ασθένειας μέχρι ο ασθενής να αναρρώσει. Οι χρόνοι σε μήνες για 100 ασθενείς είναι στο επισυναπτόμενο αρχείο.

- Κάνουμε χρήση ελέγχου υπόθεσης ότι τα δεδομένα ακολουθούν την κανονική κατανομή.
- A $\pi$ ó το menu Analyze

#### • Aπó το menu Analyze

e	Edit	view	Data	Iransfor	m	Analy
25			ĨD,		0	F
				/ -		D
						E
			time	var		т
	1		9,00			C
	2		11,00			<u>G</u>
- 3	3		45,00			G
- 3	4		82,00			N
	5		25,00			c
- 3	6		65,00			E
	7		53,00			-
- 3	8		30,00			N
6	9		10,00		_	
1	10		29,00			
	11		8,00			<u> </u>
	12		7,00			5
1	13		42,00			7
1	14		68,00			F
1	15		17,00			<u>s</u>
	16		68,00		_	N
1	17		37,00			🐝 M
1	18		50,00			N
1	19		112,00			C
2	20		59,00			冊 s
2	21		19,00			
2	22		12,00			S
2	23		63,00			0
2	24		46,00			-
1	25		19.00			

alyze	<u>G</u> raphs	Utilities	Extensions	Wi	ndow	<u>H</u> elp						
Rep	orts		*			4	G					
Des	criptive Stati	stics	•			14		•				
Baye	esian Statist	ics	•									
Ta <u>b</u> l	es		*		va	r	var		var	var	var	
Com	pare Means	3	•									
Gen	eral Linear N	lodel	•								_	
Gen	eralized Line	ear Models	•									
Mixe	d Models		•	-								
Corr	elate		•					-				
Reg	ression		•					-				_
Logi	inear		•	-				-				_
Neur	ral Net <u>w</u> orks	5	•									_
Clas	si <u>f</u> y		•			-		-				-
Dim	ension Red	uction	•			_		-				-
Scal	е		•									-
Non	parametric 1	Fests	•		ne Sam	ple					-	_
Fore	casting		•	∧ Ir	ndepend	lent Sa	mples					
Survi	ival		•	AR	Related	Sample	s					_
Multi	ple Respon	se	•	1	egacy E	)ialogs		•	Chi c	quoro		٦
Missi	ing Value Ar	nal <u>v</u> sis		-						quare		
Mulți	ple Imputati	on	•							nial		
Com	np <u>l</u> ex Sampl	es	•						Runs	***		
Simu	lation								1-Sar	nple K-S		
Qual	lity Control		•						10 2 Inde	ependent San	nples	
Spat	ial and Tem	poral Mode	ling 🕨	_					K Inde	ependent Sar	nples	
Dire	ct Mar <u>k</u> eting		•						📉 2 Rel	ated Samples	s	-
									🔣 K Rel	ated Samples	S	-

#### • Μεταφορά μεταβλητής και Normal

-	One-Sample Kolmogorov-Smirnov Test
	Test Variable List:       Exact
	Test Distribution         ✓ Normal       Uniform         Poisson       Exponential
1	OK Paste Reset Cancel Help

#### • Options

-	Cne-Sample Kolmogorov-Smirnov Test	×	
	Test Distribution   Mormal Uniform   Pojsson Exponential	Exact Options	One-Sample R-S: Options × Statistics Descriptive Quartiles Missing Values Exclude cases test-by-test Exclude cases listwise Continue Cancel Help

#### Περιγραφικά μέτρα

	Descriptive Statistics											
	N	Mean	Std. Deviation	Minimum	Maximum	25th	50th (Median)	75th				
time	100	50,2900	28,02787	7,00	129,00	27,5000	50,0000	68,0000				

Ενδιάμεσος χρόνος (Median) =50 οι μισοί ασθενείς αναρρώνουν πιο νωρίς από τους 50 μήνες

Το 75% των ασθενών αναρρώνει μετά από 27,5 μήνες (πρώτο τεταρτημόριο) ενώ το 25% αναρρώνει μετά από 68 μήνες (τρίτο τεταρτημόριο).

#### • Τεστ κανονικότητας

#### One-Sample Kolmogorov-Smirnov Test

		time
N		100
Normal Parameters <sup>a,b</sup>	Mean	50,2900
	Std. Deviation	28,02787
Most Extreme Differences	Absolute	,065
Most Extreme Differences	Positive	,065
	Negative	-,061
Test Statistic		,065
Asymp, Sig. (2-tailed)		,200 <sup>c,d</sup>

a. Test distribution is Normal.

- b. Calculated from data.
- c. Lilliefors Significance Correction.
- d. This is a lower bound of the true significance.

- Η μεγαλύτερη διαφορά της εμπειρικής από την αναμενόμενη συνάρτηση κατανομής είναι 0,065
- Με επίπεδο σημαντικότητας 5% Το pvalue > 0,2: κρατούμε τη μηδενική υπόθεση - > χρόνος ανάρρωσης ασθενών ακολουθεί κανονική κατανομή

- Έστω ότι το προηγούμενο δείγμα είχε μόνο 25 παρατηρήσεις.
- Data -> Select Cases

ile	<u>E</u> dit	View	<u>D</u> ata	<u>T</u> ransform	<u>A</u> nalyze	<u>G</u> raphs	<u>U</u> tilities	I
P			🍃 Def	fine <u>V</u> ariable F	Properties			l
_			🏄 Set	Measuremen	it <u>L</u> evel for U	Jnknown		
		-	<u>[]</u>	py Data Prope	rties			
		🔗 ti	Net	w Custom Attri	i <u>b</u> ute			var
	1		🗟 Def	fine date and t	time			
	2		Det	fine Multiple R	esponse S	ets		
	3		Val	lidation			•	
	4 r		Ido	ntify Duplicate	Casas		ŕ	
	с С			nuiy D <u>u</u> piicate	04565			
	0 7		ide	ntity Unusual	Cases			
	8		K Co	m <u>p</u> are Datase	ets			
	9		🗟 S <u>o</u> r	t Cases				
1	0		式 Sor	t Varia <u>b</u> les				
1	1		🌉 Tra	<u>n</u> spose				
1	2		🛨 Adj	ust String Wid	ths Across	Files		
1	3		Me	r <u>q</u> e Files			•	
1	4		🐺 <u>R</u> e	structure				
1	15		+ Ra	ke Weights				
1	6		+ Pro	pensity Score	Matching			
1	7		+ Ca	se Control Ma	tching			
1	8			aregate	toring			
1	9	1		hogonal Daai	~~			
2	20			inogonal Desi	gn			
2	21		+ Spi	it into Files				
2	2		🔁 Co	py <u>D</u> ataset				
2	23		📰 Spl	it <u>F</u> ile				
2	24		E Sel	ect Cases				
2	25		ata we	inht Cases				

• Data -> Select Cases

Select Cases	×	:
time time25	Select   All cases   If condition is satisfied   If   Random sample of cases   Sample   Based on time or case range   Range   Use filter variable:   Image:   Image:	
Current Status: Do not filter	cases	ĺ
OK	Paste Reset Cancel Help	

• Επιλογή 1 έως 25 και continue

Select Case	es: Range	×
Observation:	First Case	Last Case
Continue	Cancel	Help

Κάτω του 26 διαγραμμένες

19:			
	🛷 time	🔗 time25	var
1	9,00	9,00	
2	11,00	11,00	
3	45,00	45,00	
4	82,00	82,00	
5	25,00	25,00	
6	65,00	65,00	
7	53,00	53,00	
8	30,00	30,00	
9	10,00	10,00	
10	29,00	29,00	
11	8,00	8,00	
12	7,00	7,00	
13	42,00	42,00	
14	68,00	68,00	
15	17,00	17,00	
16	68,00	68,00	
17	37,00	37,00	
18	50,00	50,00	
19	112,00	112,00	
20	59,00	59,00	
21	19,00	19,00	
22	12,00	12,00	
23	63,00	63,00	
24	46,00	46,00	
25	19,00	19,00	
26	68,00		
27	41,00	•	
20	1 25.00		

#### Menu Analyze

<b>ta</b> *xi	ronoi_astl	h.sav [Da	ataSet0]	- IBM SPSS Stati	stics Data	Editor					
<u>F</u> ile	Edit	View	<u>D</u> ata	Transform	Analyze	Graphs	<u>U</u> tilities	Extensions	Window	Help	
P			ī.	10 5	Repo	orts		+			
					D <u>e</u> so	criptive Stati	stics	•	123 Erequer	ncies	
19:					<u>B</u> aye	sian Statist	ics	•	Descrip	tives	
		🖉 t	ime	🔗 time25	Ta <u>b</u> le	es		•	A Explore		var
	1		9,00	9,00	Co <u>m</u>	pare Mean	s	•	Crossta	he	
	2		11,00	11,00	Gene	eral Linear I	Model	*			
	3		45,00	45,00	Gene	eralized Lin	ear Models	•	TURF A	nalysis	-
	4		82,00	82,00	Mixed	d Models		•	Ratio		
	5		25,00	25,00	Corre	elate		•	P-P Plot	ts	
1	6		65,00	65,00	Rear	ession			🛃 Q-Q Plo	ts	
	7		53,00	53,00	Logi	near					
	8		30,00	30,00	Nour	neal	_				
1	9		10,00	10,00	Neur	annetwork	5				
	10		29,00	29,00	Clas	sily					
1	11		8,00	8,00	Dime	ension Red	uction				
1	12		7,00	7,00	Sc <u>a</u> le	Ð		,			
1	13		42,00	42,00	Nont	parametric	Tests	*			
	14		68,00	68,00	Fore	casting		•			
	15		17,00	17,00	<u>S</u> urvi	val		•			
1	16		68,00	68,00	M <u>u</u> ltij	ple Respon	ise	•			
	17		37,00	37,00	🔛 Missi	ng Value A	nal <u>v</u> sis				
	18		50,00	50,00	Mulți	ple Imputat	ion	•			1
	19	1	112,00	112,00	Com	plex Sampl	es	•			
1	20		59,00	59,00	🛱 Simu	lation					
1	21		19,00	19,00	Qual	ity Control					
	22		12,00	12,00	Spati	ial and Terr	noral Mode	ling b			
	23		63,00	63,00	Diror	t Markating					
	24		46.00	46.00	Direc	a marketing					

Θέτουμε την μεταβλητή και πιέζουμε το plots

ta Explore		×	
Vime25	Dependent List:	Statistics Plots Options Bootstrap	

ta Explore: Plots	×				
Boxplots Eactor levels together Dependents together None	Descriptive <u>S</u> tem-and-leaf <u>H</u> istogram				
<ul> <li>Normality plots with tests</li> <li>Spread vs Level with Levene Test</li> <li>None</li> <li>Power estimation</li> <li>Transformed Power: Natural log</li> </ul>					
Continue Cancel Help					

	Des	criptives		
			Statistic	Std. Error
time	Mean		39,4400	5,45010
	95% Confidence Interval for Mean	Lower Bound	28,1915	
		Upper Bound	50,6885	
5% Trimmed Mean			37,5333	
	Median	37,0000		
	Variance	742,590		
	Std. Deviation	27,25050		
	Minimum	7,00		
	Maximum	112,00		
	Range	105,00		
	Interquartile Range	46,50		
	Skewness		,786	,464
	Kurtosis		,355	,902

#### Tests of Normality

	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
time	,133	25	,200	,923	25	,062

\*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

P-value οριακά άνω του 0.05. Ίσως ακολουθεί κανονική κατανομή. Είναι μικρό το δείγμα.

Θεωρούμε τυχαίο δείγμα  $Y_1, Y_2, ..., Y_n$  με συνεχή αθροιστική συνάρτηση κατανομής F(y)Θεωρούμε διατεταγμένο δείγμα της τυχαίας μεταβλητής  $X_i = F(Y_i)$  με αναμενόμενη μέση τιμή  $E[F(Y_i)] = \frac{i}{n+1}$  για κάθε i = 1,2,..., n

Παίρνουμε ένα μεγάλο πλήθος παρατηρήσεων  $x_1, x_2, \dots, x_n$  και τοποθετουμε σε γράφημα τα σημεία

$$F(y_i), \frac{l}{n+1}$$

Πρέπει να βρίσκονται κοντά σε ευθεία y = x

Σχηματίζεται Γράφημα με άξονα x την παρατηρούμενη αθροιστική συνάρτηση πιθανότητας και άξονα y την αναμενόμενη αθροιστική συνάρτηση πιθανότητας: Διάγραμμα P-P

Παρόμοιο γράφημα είναι αυτό που στον άξονα x έχουμε την παρατηρούμενη τιμή και στον άξονα y την αναμενόμενη κανονικοποιημένη τιμή: Διάγραμμα Q-Q

#### Έλεγχος κανονικότητας Διαγράμματα πιθανότητας περιληπτικά

- Αρχικά βρίσκουμε τη συνάρτηση πιθανότητας. Αποτελεί τη συνάρτηση που περιγράφει μαθηματικά τον πληθυσμό από τον οποίο προέρχεται το προς εξέταση δείγμα.
- Για τον σχεδιασμό των γραφημάτων πιθανότητας
   χρησιμοποιούμε τους άξονες του επιπέδου βαθμονομημένους με
   τη δεδομένη κατανομή, δηλαδή την κανονική.

- Διατάσσουμε τις παρατηρήσεις από την μικρότερη στη μεγαλύτερη.
- <u>Σχεδιάζουμε</u> τα γραφήματα.

• Data -> Select cases -> All cases

Select Cases	×
	Select
🔗 time	All cases
	◎ If <u>c</u> ondition is satisfied
4	If
	Random sample of cases
	Sample
	O Based on time or case range
	Range 1 thru 25
	O Use filter variable:
	· · · · · · · · · · · · · · · · · · ·
	Output
	<u>Filter out unselected cases</u>
	O Copy selected cases to a new dataset
	Dataset name:
	Delete unselected cases
Current Statue: Do not filter	
Guirent Status, Do not litter	cases
ОК	Paste Reset Cancel Help

#### • Analyze

atis	tics Data Ec	altor									
n	Analyze	<u>G</u> raphs	Utilities	Extension	ns	Win	dow	Help			
-	Repor	rts		•					A (	2	6
4	Descr	iptive Stati:	stics	•		123 <u>E</u> re	equen	cies		0	
	Bayes	ian Statisti	cs	•		De	escript	ives			_
	Ta <u>b</u> le	s		*		A Ex	plore.			var	
	Comp	are Means		•		Cr	nestal	hs			
	Gener	ral Linear M	lodel	*				valveie			
_	Gener	ralized Line	ear Models	•				lalysis			_
	Mixed	Models		*		<u>III</u> <u>R</u> a	10				
_	Corre	late		•		<u>P-</u>	P Plots	5			_
-	Regre	ssion		*	Ľ	<u>Q</u> -	Q Plot	S		-	_
-	Loglin	near		*					-		
	Neura	al Net <u>w</u> orks		*					-		-
-	Class	i <u>f</u> y		۴					-		
	Dime	nsion Redu	uction	*					-		-
	Sc <u>a</u> le			*							
	Nonpa	arametric T	ests	•					1		-
	Forec	asting		*					+		_
	Surviv	al		*					1		_
	Multip	le Respon	se	*							
	🔛 Missin	ig Value An	al <u>v</u> sis								
	Multip	le Imputati	on	•							
	Comp	lex Sample	es	•							
	Bimula	ation									_
	Qualit	y Control		*							
	<u>S</u> patia	al and Tem	poral Model	ing 🕨							
	Direct	Mar <u>k</u> eting		•							
						9					

P-P Plots	×
Variables: Time Transform Natural log transform Standardize values Difference: 1 Seasonally difference: 1 Current Periodicity: None	Test Distribution   Normal   df.   - Distribution Parameters   Image: Distri



Όλες οι τιμές γύρω από την ευθεία άρα ακολουθεί  $\tau\eta\nu$ κανονική κατανομή



0.2

0.4

**Observed Cum Prob** 

0.6

0.8

1.0

#### • Eπιλογή: Analyze -> Descriptive -> QQ plot



• Επιλογή της μεταβλητής και της κατανομής από το menu κατανομών

P-P Plots		×
	Variables:	Test Distribution   Normal   dt   dt   - Distribution Parameters   Image: Estimate from data   Location:   0   Scale:   1   Proportion Estimation Formula Image: Solution S
	OK Paste Reset Cancel	Help

• Πίνακας εκτιμημένων παραμέτρων της κανονικής κατανομής

	Estimated Distribution Parameters					
			time			
♦	Normal Distribution	Location	50,2900			
		Scale	28,02787			
	The cases are unweighted.					

- Η μέση τιμή (location) εκτιμήθηκε 50,29
- Η τυπική απόκλιση (Scale) εκτιμήθηκε 28,027



 Όλες οι τιμές βρίσκονται κοντά στη διαγώνιο -> τα δεδομένα επαληθεύεται ότι ακολουθούν την κανονική κατανομή