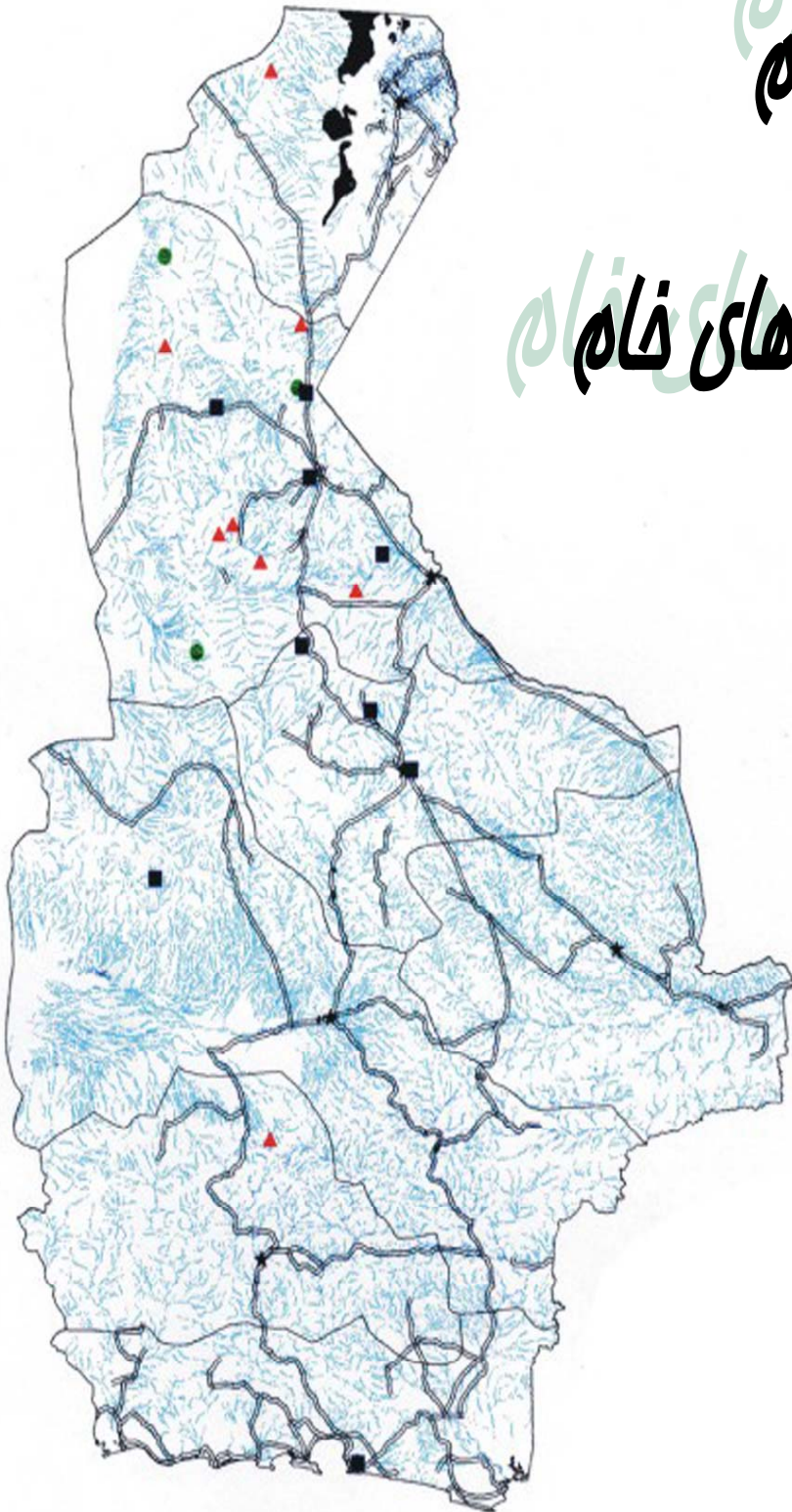


# فصل چهارم

## پژدازش داده‌های خام



### محاسبه پارامترهای آماری داده‌های خام

اولین مرحله پردازش داده‌های ژئوشیمیایی، بررسی پارامترهای آماری مربوط به تک تک عناصر جهت شناخت ماهیت توزیع هریک از آنها می‌باشد که با محاسبه پارامترهای آماری از قبیل میانگین، انحراف معیار، چولگی، کشیدگی، واریانس و ... می‌توان به این موضوع دست یافت. در این قسمت برای هر عنصر به عنوان یک متغیر آماری در یک جدول، تعداد نمونه‌ها، حداقل و حداکثر عیار، میانگین، میانه، انحراف معیار، چولگی و کشیدگی و نمودارهای هیستوگرام توزیع فراوانی محاسبه و ترسیم شده‌اند.

### بررسی مقادیر خارج از رده : ( Outliers )

هنگام بررسی مقادیر داده‌های خام به نمونه‌هایی برخورد می‌شود که در آستانه‌های بالا و پایین جامعه داده‌ها قرار گرفته و از جامعه اصلی جدا افتاده‌اند. اگر نمودار جعبه‌ای ( Boxplot ) آنها ترسیم شود این نمونه‌ها به نحو بارزی خودشان را از بقیه جدا می‌کنند. مقادیر خارج از رده به سه حالت مختلف زیر ممکن است بوجود آیند:

حالت اول) از یک خطای سیستماتیک به هنگام نمونه برداری، آماده‌سازی یا تجزیه شیمیایی نمونه‌ها ناشی شده باشند که باید از مرحله پردازش حذف یا اصلاح شوند.

حالت دوم) مشاهداتی که به صورت یک پدیده فوق‌العاده نمود پیدا می‌کنند که باید پس از بررسی اعتبار آنها در مورد حفظ یا حذف آنها تصمیم گرفت.

حالت سوم) مشاهدات فوق‌العاده‌ای که هیچگونه توضیح مناسبی برای آنها وجود ندارد و کارشناس اگر احساس کند که آنها به عنوان گوشه‌ای از جامعه مورد بررسی هستند می‌تواند آنها را حفظ کند.

وجود مقادیر خارج از رده در جامعه نمونه‌ها موجب افزایش واریانس جامعه و نیز همبستگی بین متغیرها و همچنین افزایش چولگی در نمودار توزیع عناصر می‌شود. برای کاهش این تاثیر راههای مختلفی نظیر محاسبه ضریب همبستگی با استفاده از روشهای ناپارامتری مانند روش اسپیرمن (Spearman)، حذف یا جایگزین نمودن مقادیر استفاده می‌شود در این گزارش از روش جایگزین نمودن مقادیر خارج از رده استفاده شده است. جدول (۴-۱) نمونه‌های دارای مقادیر خارج از رده را نشان می‌دهد.

### نرمال سازی داده‌های خام:

استفاده از برخی روشهای آماری منوط به نرمال بودن تابع توزیع متغیرهای مورد مطالعه است در حالیکه توابع توزیع از نوع لاگ نرمال است، به همین علت قبل از استفاده از این روشها داده‌های خام باید نرمال شوند. در این بخش از نوعی تبدیلات جهت نرمال کردن تابع توزیع داده‌های خام استفاده شده است. این کار شرط لازم کاربرد برخی روشهای آماری مانند تعیین نمونه‌های آنومالی با استفاده از اضافه کردن ضرایبی از انحراف معیار به حد آستانه‌ای و یا محاسبه ضرایب همبستگی پیرسون می‌باشد. روش لاگ نرمال به صورت یک روش توصیفی برای نرمال کردن تابع توزیع جوامعی که دارای چولگی در نمودار خود هستند به کار می‌رود.

**Table (4-1) : Outlier Samples For Normal RawData**

Sample Number		
Elements	Outlier (+)	Outlier (-)
Au		
W	MM-241 , MM-239	
Mo		
Se		
Cr		
Co		
Ni		
Cu		
Zn	MK-112	
As	MB-443	
Sr		
Ag		
Be		
Sn		
Sb		
Ba	MB-442 , MD-464	
Pb		
Bi	MK-136	
Hg		
B		
Ti	MK-112	
Mn		

در اینجا از لگاریتم طبیعی مقادیر داده‌های خام به اضافه یا منهای یک مقدار ثابت  $\lambda$  مطابق رابطه تبدیلی زیر استفاده شده است.

$$Z = \text{Ln}(AE \pm \lambda)$$

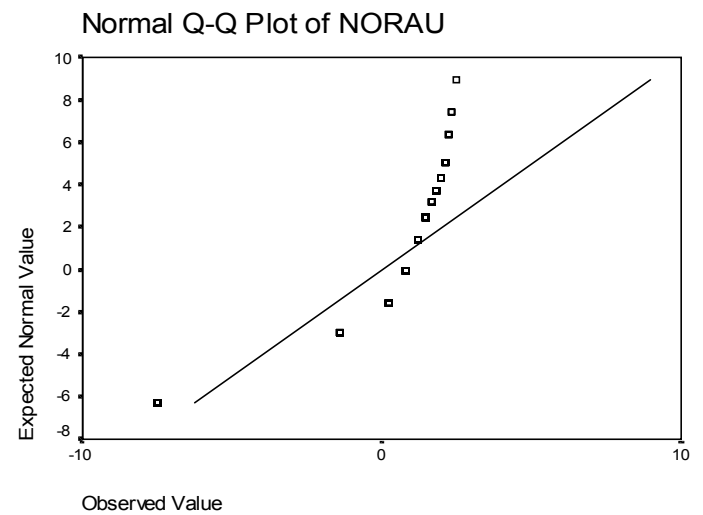
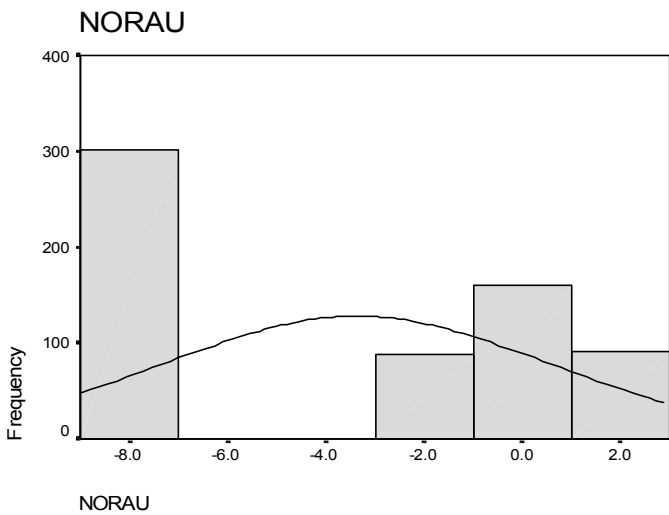
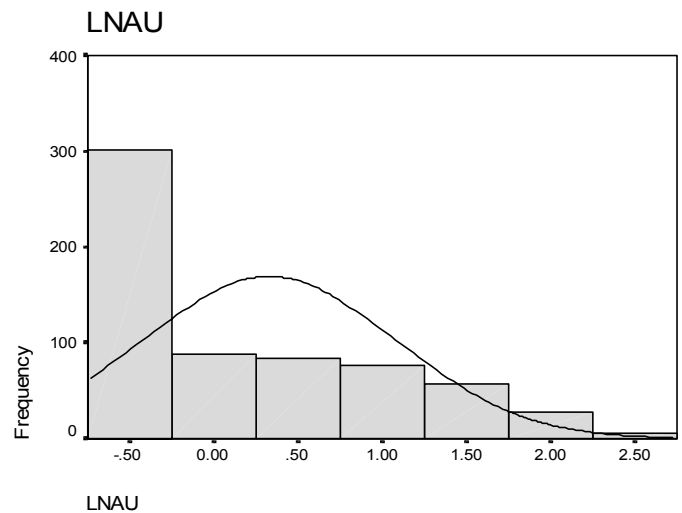
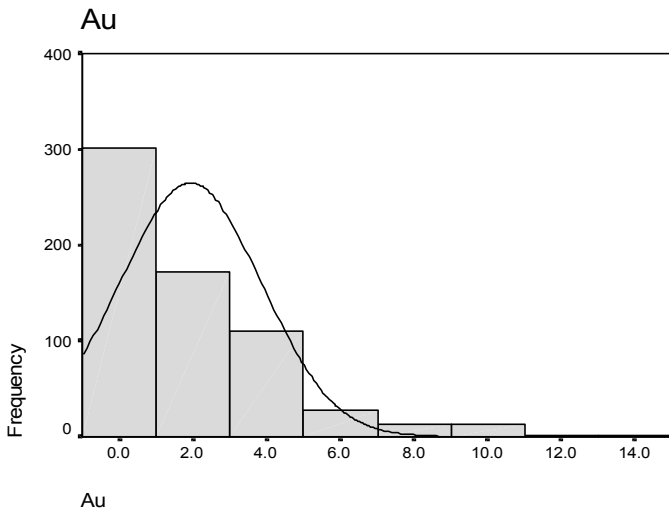
در این رابطه AE آنالیز نمونه برای هر عنصر است.

برای هر عنصر مقدار  $\lambda$  به گونه‌ای انتخاب می‌شود که پس از انتخاب داده‌ها به یک مقدار بهینه از چولگی و کشیدگی در منحنی توزیع نرمال دست یافته شود. پارامترهای آماری و هیستوگرام‌های ترسیم شده برای داده‌های نرمال در شکل (۴-۱) تا (۴-۷) آورده شده است. با توجه به این پارامترهای آماری می‌توان دریافت که مقادیر چولگی و کشیدگی متغیرها در مقایسه با مقادیر متناظر مربوط به داده‌های خام نرمال نشده تا چه اندازه کاهش یافته و منحنی توزیع تجمعی آنها به صورت یک خط راست که بیانگر توزیع نرمال می‌باشد، ظاهر شده است. هیستوگرام مقادیر نرمال شده نسبت به هیستوگرام مقادیر نرمال نشده نیز بیانگر مطلب فوق می‌باشد.

**Fig(4-1): Statistical Parameters For Raw Data in Maksan**

**Statistics**

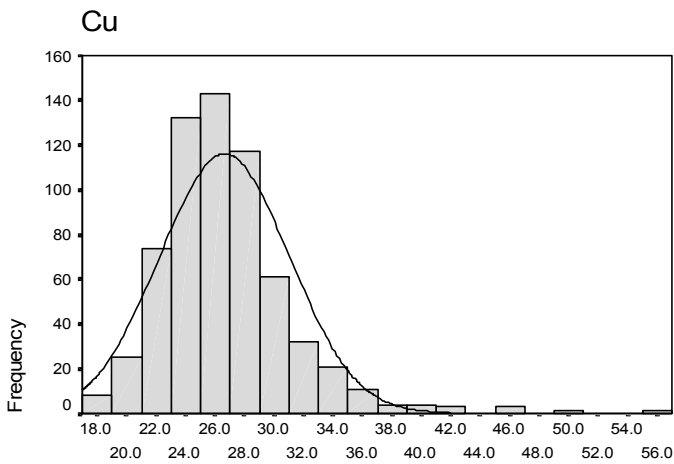
		<b>Au</b>	<b>LNAU</b>	<b>NORAU</b>
N	Valid	640	640	640
	Missing	0	0	0
Mean		1.936	.3282	-3.3748
Median		1.000	.0000	-1.3841
Std. Deviation		1.924	.7522	3.9798
Skewness		2.342	.901	.000
Std. Error of Skewness		.097	.097	.097
Kurtosis		6.291	-.431	-1.870
Std. Error of Kurtosis		.193	.193	.193
Minimum		.8	-.29	-7.51
Maximum		13.0	2.56	2.51



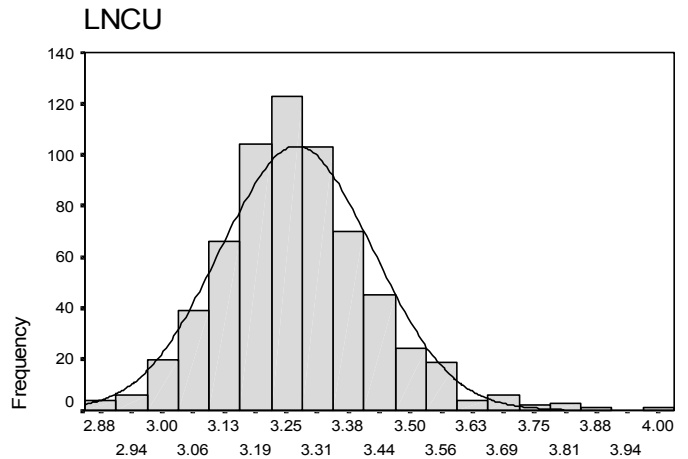
**Fig(4-2):Statistical Parameters For Raw Data in Maksan**

**Statistics**

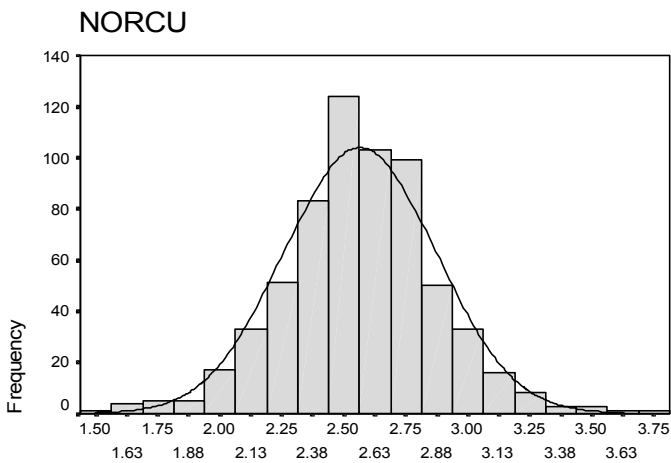
		<b>Cu</b>	<b>LNCU</b>	<b>NORCU</b>
N	Valid	640	640	640
	Missing	0	0	0
Mean		26.647	3.2704	2.5641
Median		25.900	3.2542	2.5549
Std. Deviation		4.392	.1541	.3062
Skewness		1.536	.665	.000
Std. Error of Skewness		.097	.097	.097
Kurtosis		5.394	1.679	1.009
Std. Error of Kurtosis		.193	.193	.193
Minimum		17.7	2.87	1.54
Maximum		55.8	4.02	3.76



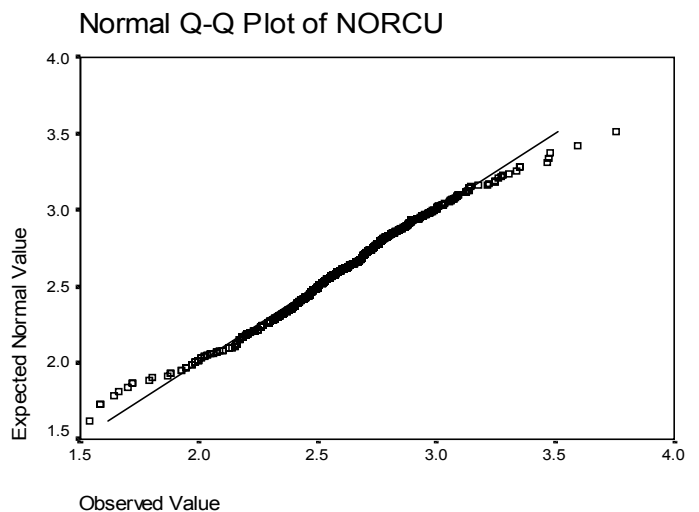
Cu



LNCU



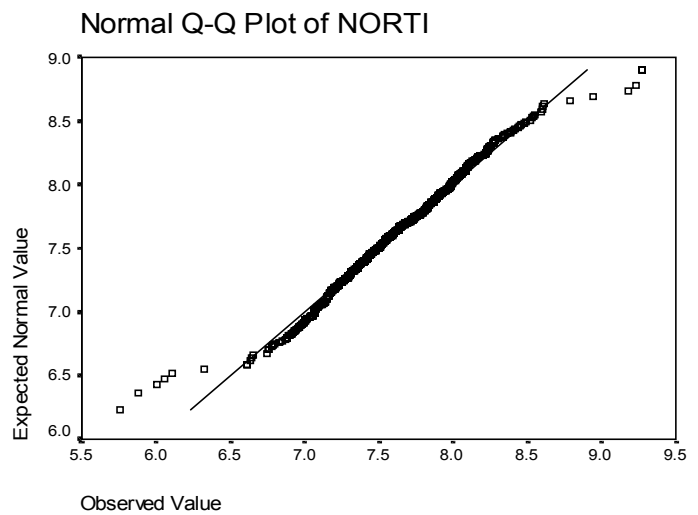
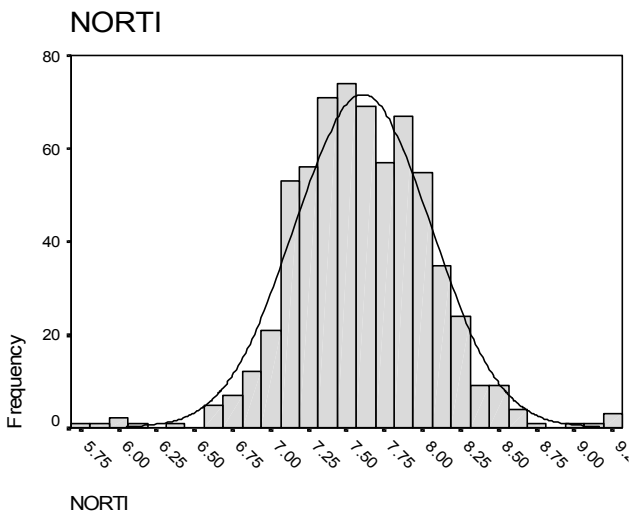
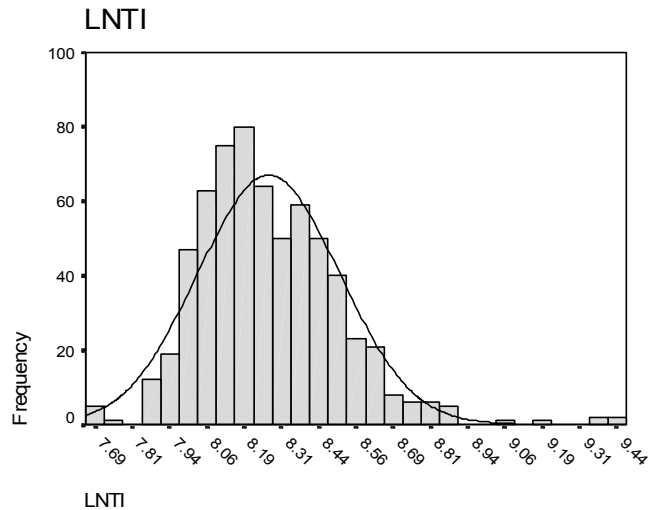
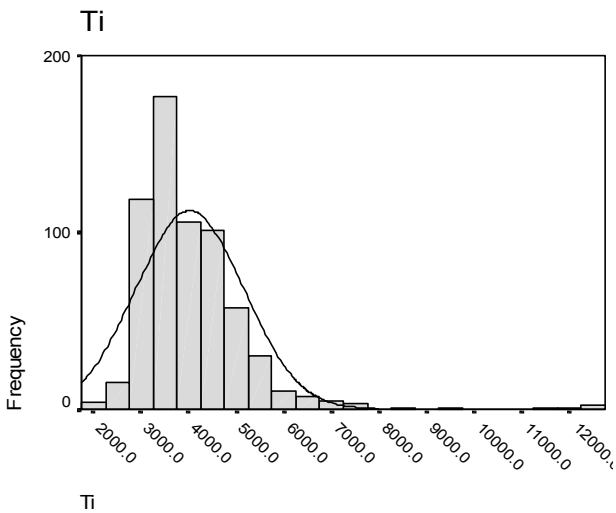
NORCU



**Fig(4-3):Statistical Parameters For Raw Data in Maksan**

**Statistics**

		Ti	LNTI	NORTI
N	Valid	640	640	640
	Missing	0	0	0
Mean		4025.641	8.2697	7.6056
Median		3785.000	8.2388	7.5922
Std. Deviation		1134.094	.2377	.4446
Skewness		2.897	.976	.000
Std. Error of Skewness		.097	.097	.097
Kurtosis		15.925	2.696	1.551
Std. Error of Kurtosis		.193	.193	.193
Minimum		2120.0	7.66	5.76
Maximum		12400.0	9.43	9.27

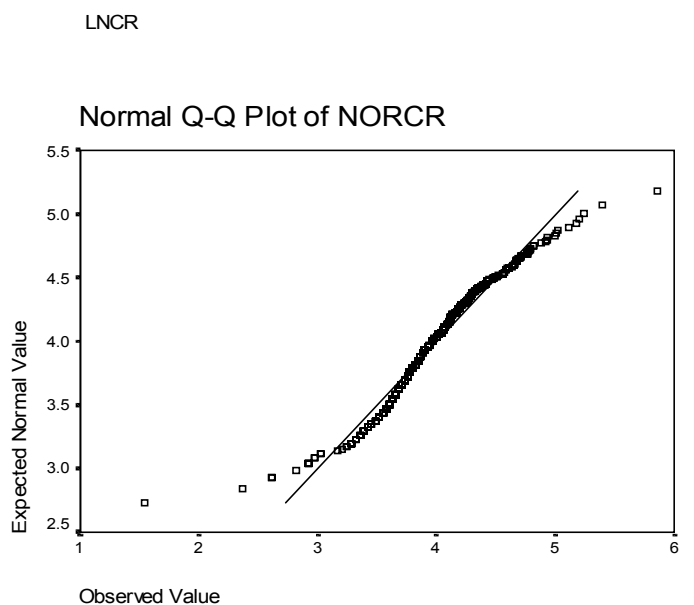
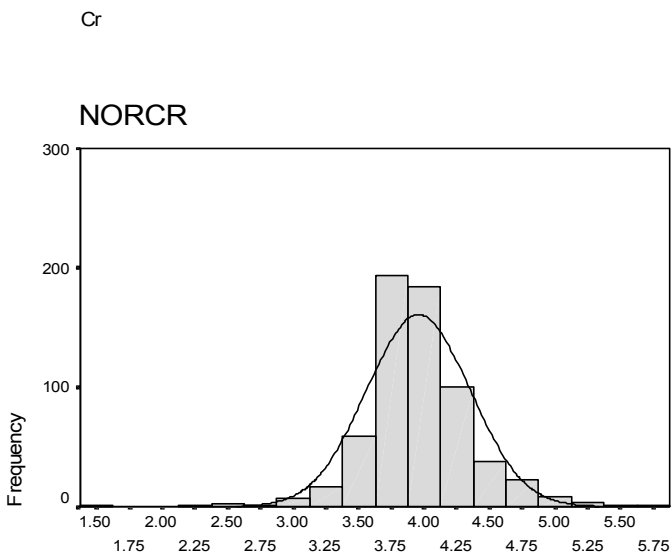
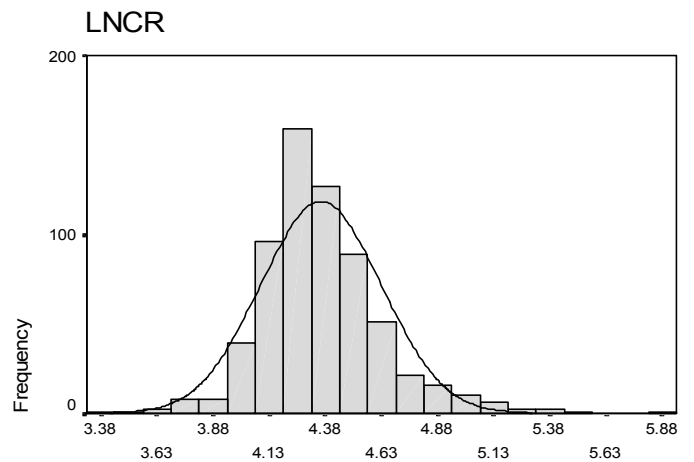
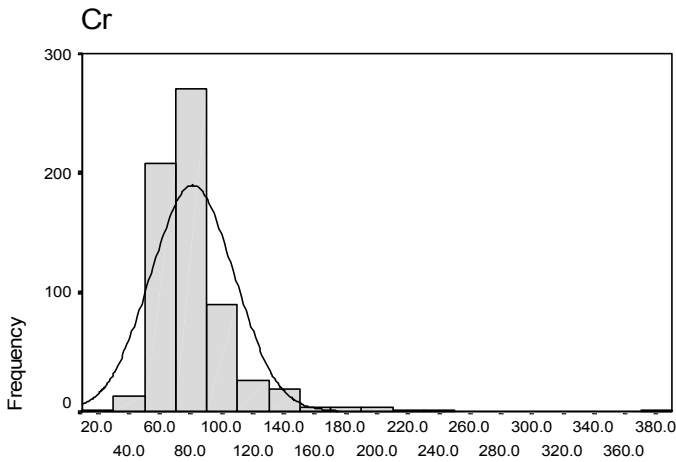




**Fig(4-4):Statistical Parameters For Raw Data in Maksan**

**Statistics**

		<b>Cr</b>	<b>LNCR</b>	<b>NORCR</b>
N	Valid	640	640	640
	Missing	0	0	0
Mean		81.020	4.3547	3.9555
Median		75.000	4.3175	3.9249
Std. Deviation		26.867	.2691	.3961
Skewness		3.579	.938	.000
Std. Error of Skewness		.097	.097	.097
Kurtosis		26.186	3.252	3.877
Std. Error of Kurtosis		.193	.193	.193
Minimum		29.0	3.37	1.54
Maximum		373.0	5.92	5.85



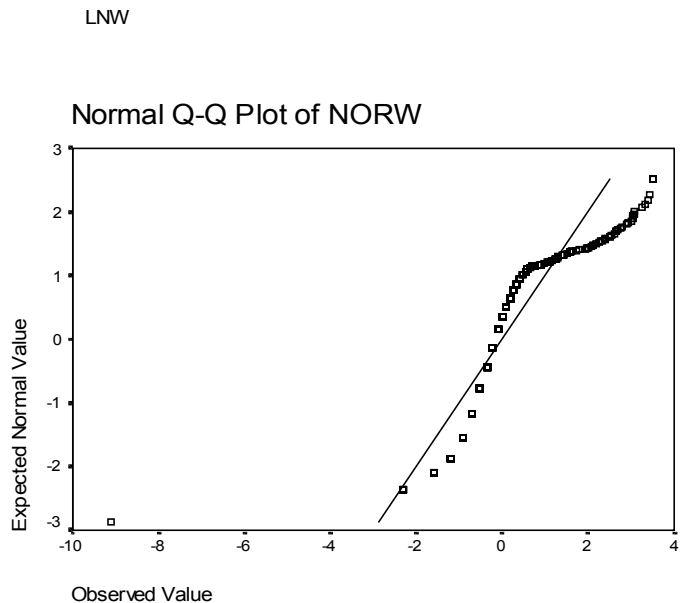
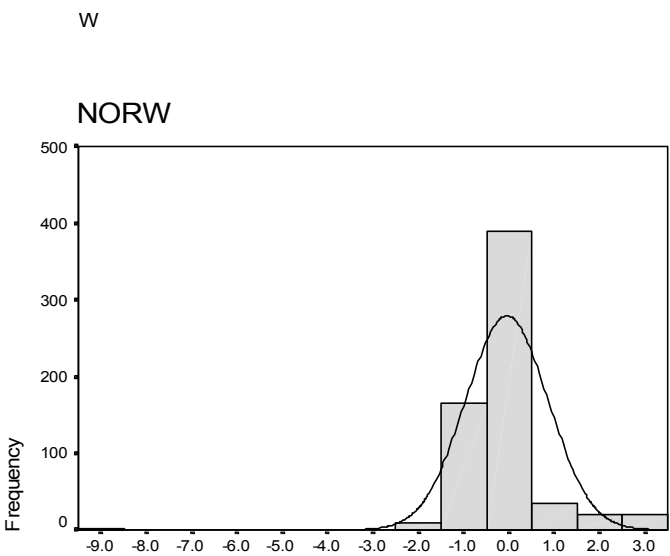
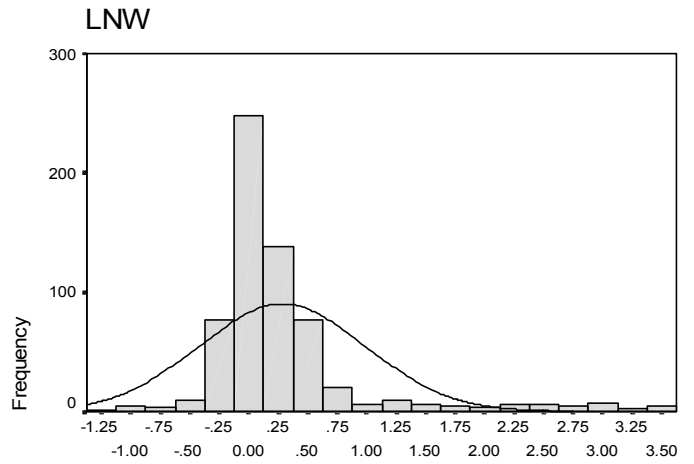
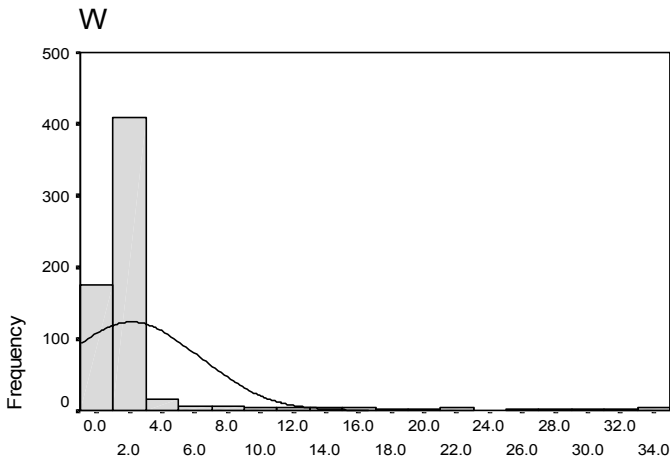
NORCR

Observed Value

**Fig(4-5):Statistical Parameters For Raw Data in Maksan**

**Statistics**

		W	LNW	NORW
N	Valid	640	640	640
	Missing	0	0	0
Mean		2.121	.2820	-4.84E-02
Median		1.100	9.531E-02	-.2230
Std. Deviation		4.127	.7063	.9128
Skewness		5.244	2.521	.000
Std. Error of Skewness		.097	.097	.097
Kurtosis		29.999	7.223	18.336
Std. Error of Kurtosis		.193	.193	.193
Minimum		.3	-1.20	-9.10
Maximum		33.3	3.51	3.50



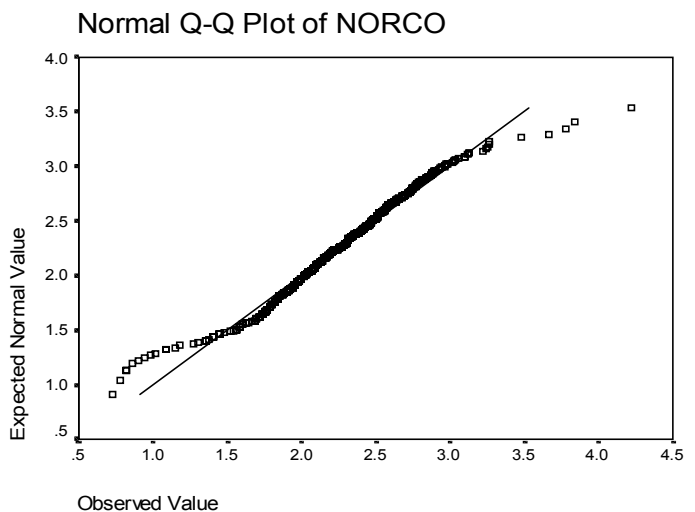
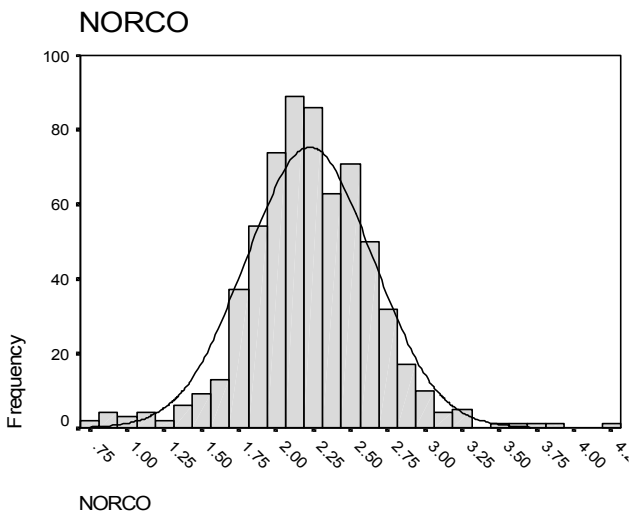
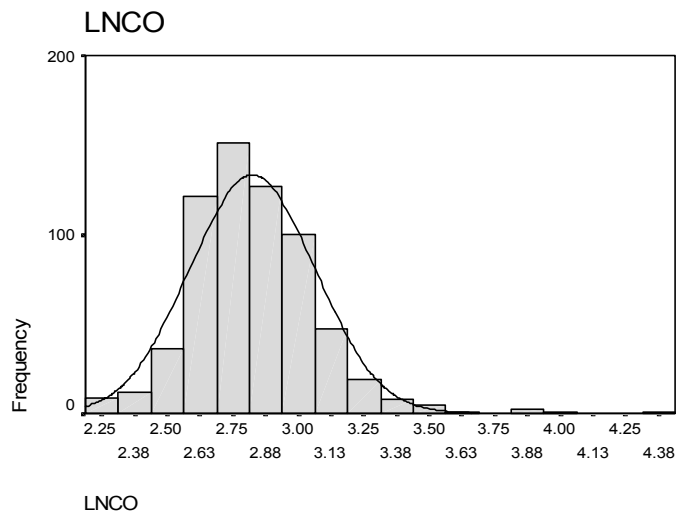
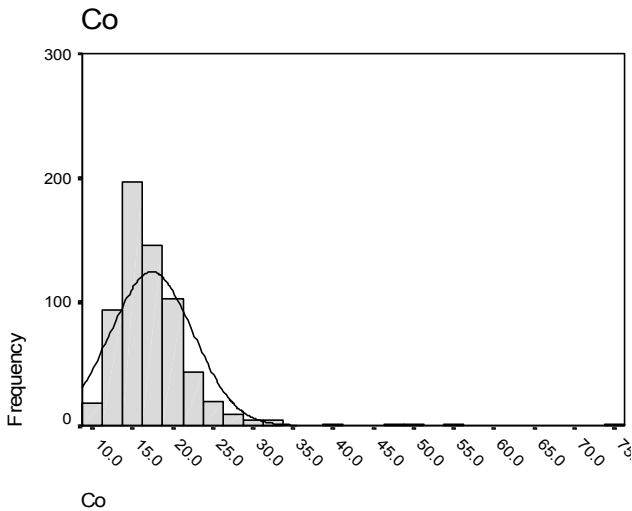
NORW

Observed Value

**Fig(4-6):Statistical Parameters For Raw Data in Maksan**

**Statistics**

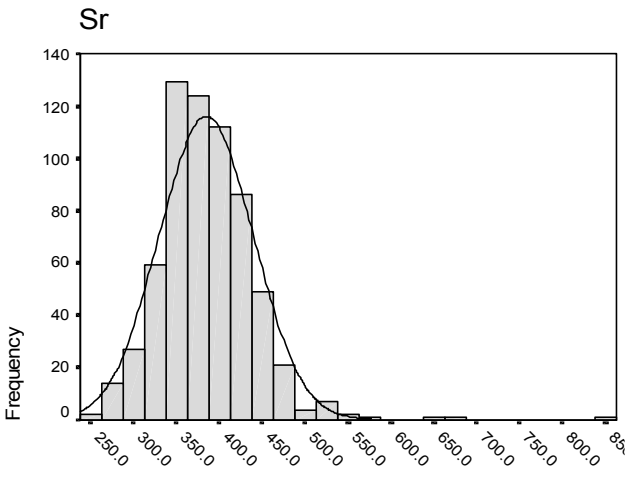
		Co	LNCO	NORCO
N	Valid	640	640	640
	Missing	0	0	0
Mean		17.391	2.8244	2.2263
Median		16.400	2.7973	2.2159
Std. Deviation		5.134	.2390	.4227
Skewness		4.067	1.039	.000
Std. Error of Skewness		.097	.097	.097
Kurtosis		33.611	4.102	2.056
Std. Error of Kurtosis		.193	.193	.193
Minimum		9.3	2.23	.73
Maximum		75.5	4.32	4.22



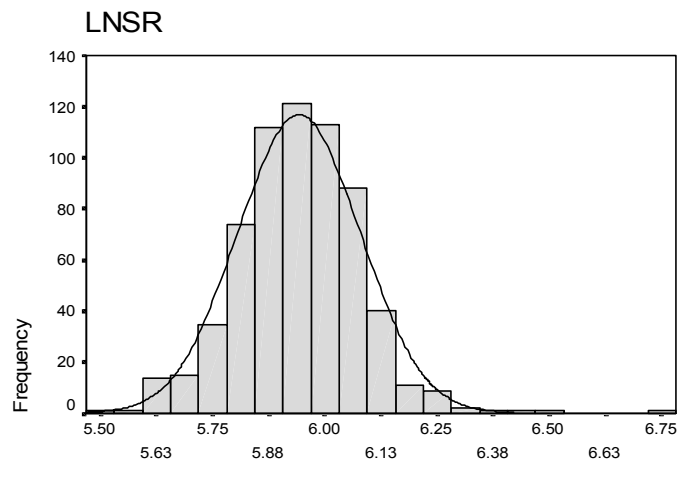
**Fig(4-7):Statistical Parameters For Raw Data in Maksan**

**Statistics**

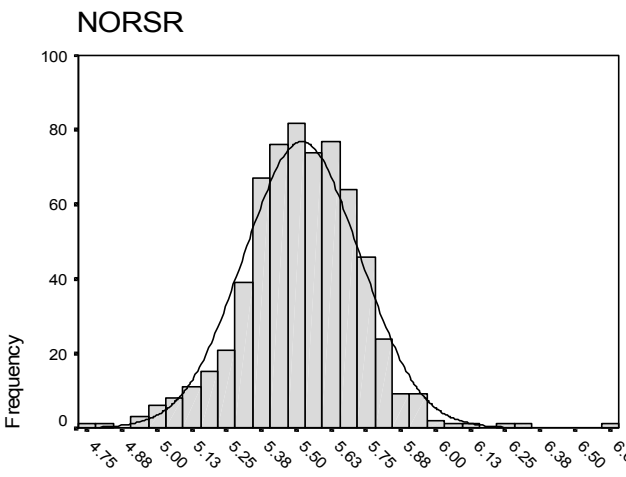
		<b>Sr</b>	<b>LNSR</b>	<b>NORSR</b>
N	Valid	640	640	640
	Missing	0	0	0
Mean		384.306	5.9420	5.5201
Median		380.000	5.9402	5.5247
Std. Deviation		54.897	.1364	.2073
Skewness		1.571	.381	.000
Std. Error of Skewness		.097	.097	.097
Kurtosis		10.084	2.448	1.651
Std. Error of Kurtosis		.193	.193	.193
Minimum		244.0	5.50	4.74
Maximum		860.0	6.76	6.59



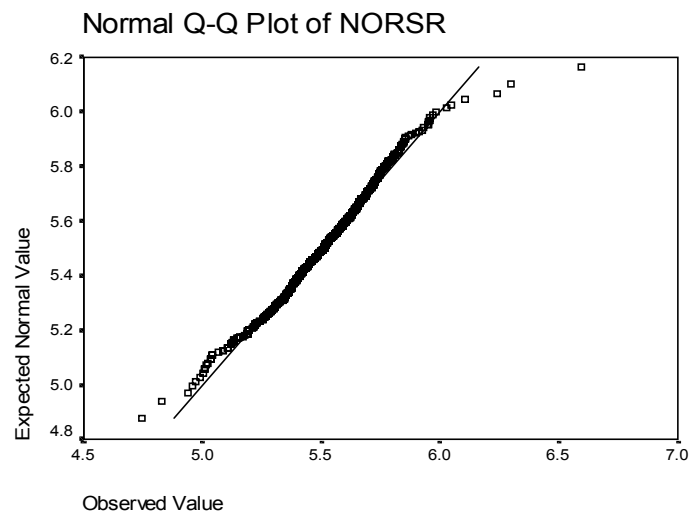
Sr



LNSR



NORSR



### تعیین ضریب همبستگی:

برای تعیین اینکه آیا ارتباط معنی‌داری میان تغییرات متغیرهای آماری وجود دارد، ضرایب همبستگی میان آنها محاسبه می‌شود. این عمل به دو منظور کشف همبستگی بین متغیرها و تخمین مقدار یک یا چند متغیر دیگر صورت می‌گیرد. برای بررسی، دو نوع ضریب همبستگی پیرسون و اسپیرمن به صورت ماتریس ضرایب همبستگی محاسبه شده‌اند که در جداول (۲-۴) و (۳-۴) آمده است شرط محاسبه ضریب همبستگی پیرسون، نرمال بودن تابع توزیع متغیرها می‌باشد. در این جداول، **Sig(2-Tailed)** میزان معنی‌دار بودن ضرایب همبستگی طبق آزمون فرض مساوی صفر بودن ضریب همبستگی می‌باشد.

برای محاسبه ضریب همبستگی پیرسون به علت تاثیرپذیری این پارامتر از آستانه‌های بالا و پایین حتماً باید داده‌های خام نرمال شوند تا ضریب همبستگی محاسبه شوند. جدول (۲-۴) مقادیر این ضرایب را نشان می‌دهد.

بر پایه جدول ضریب همبستگی پیرسون بین جفت متغیرهای **Zn,Co(0.815)** و **Mn,Co(0.684)** و **Ti,Co(0.817)** و **Ti,Cr(0.635)** و **Zn,Mn(0.869)** و **Ti,Mn(0.855)** و **Ti,Zn(0.888)** و **Be,Sn(0.657)** و **As,Sb(0.620)** در سطح اعتماد مطلوب ۹۹٪ می‌باشد که بیشترین ارتباط همبستگی بین عناصر **Ti,Zn(0.888)** وجود دارد. این ضرایب بیانگر ارتباط پارائزی بین عناصر می‌باشند.

برای محاسبه ضریب همبستگی اسپیرمن از داده‌های خام استفاده شده است و همانطور که مشاهده می‌شود، در بعضی مواقع وضعیت متفاوتی نسبت به ضریب همبستگی پیرسون دارد. این اختلاف بیشتر زمانی بروز می‌کند که مقدار داده‌های خارج از رده زیاد باشد. اما مقایسه دقیق آنها، این نکته را بیان می‌کند که اختلاف این دو

ضریب همبستگی خیلی زیاد نیست ، این امر نشان دهنده تاثیرپذیری کم داده‌ها از مقادیر خارج از رده است. جدول (۳-۴) مقادیر این ضرایب را نشان می‌دهد.







بر پایه این جدول ضریب همبستگی مشاهده شده بین عناصر  $\text{Mn,Co}(0.705)$  و  $\text{Zn,Co}(0.825)$  و  $\text{Ti,Co}(0.829)$  و  $\text{Sn,Mo}(0.617)$  و  $\text{W,Sn}(0.641)$  و  $\text{Zn,Mn}(0.877)$  و  $\text{Ti,Mn}(0.862)$  و  $\text{Ti,Zn}(0.878)$  و  $\text{Sn,Be}(0.707)$  در سطح اعتماد ۹۹٪ می‌باشد که بیشترین ارتباط همبستگی بین عناصر  $\text{Ti,Zn}(0.878)$  وجود دارد. ضریب همبستگی بین جفت متغیرها به روش پیرسون و اسپیرمن بیانگر اختلاف تقریباً کم بین ضرایب همبستگی عناصر متناظر می‌باشد که حکایت از توزیع نسبتاً نرمال عناصر و همین‌طور عدم تاثیر نمونه‌های دور افتاده دارد.

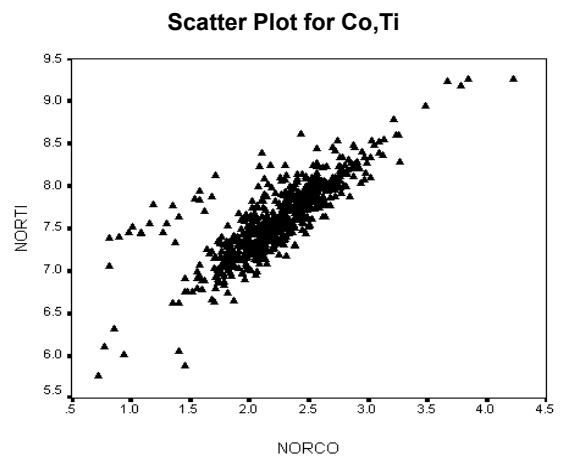
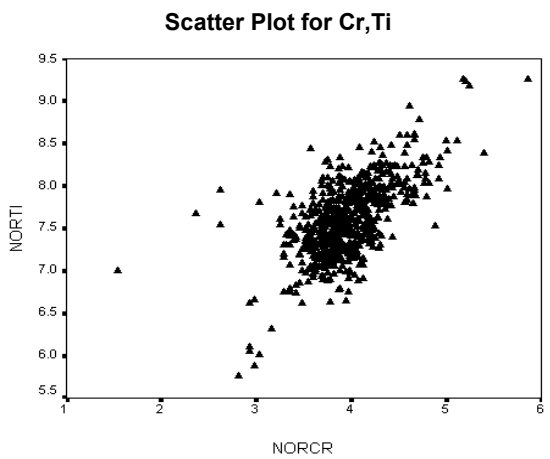
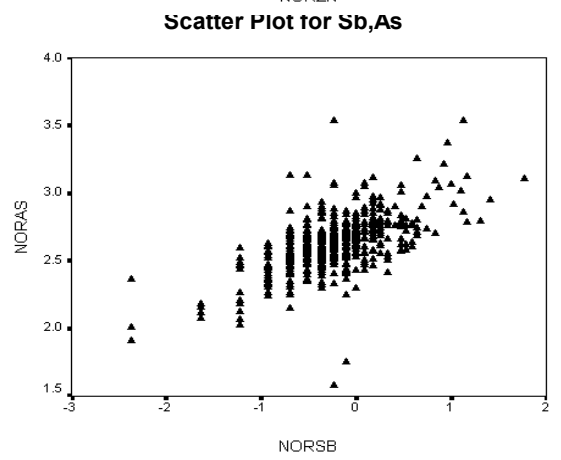
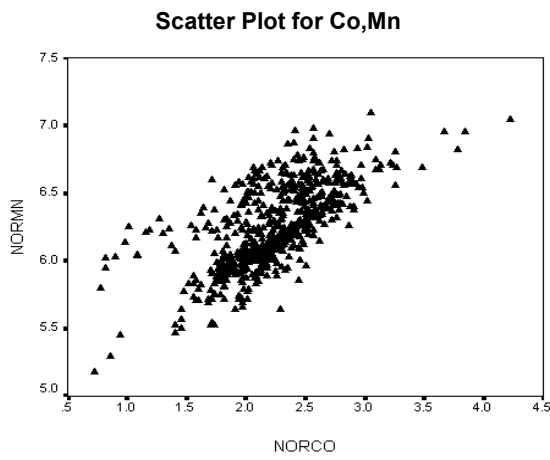
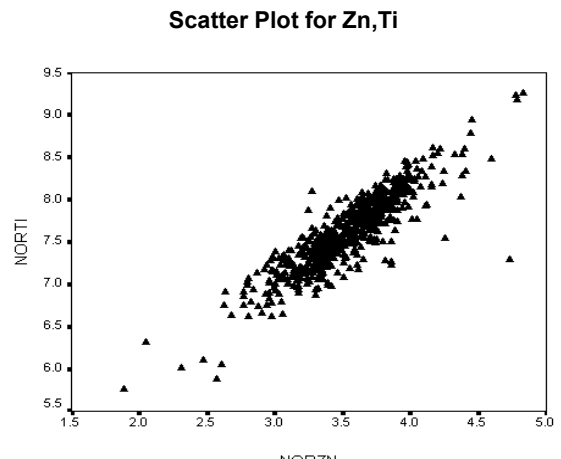
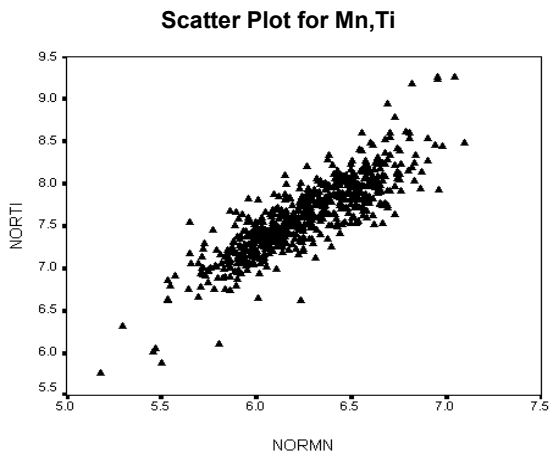
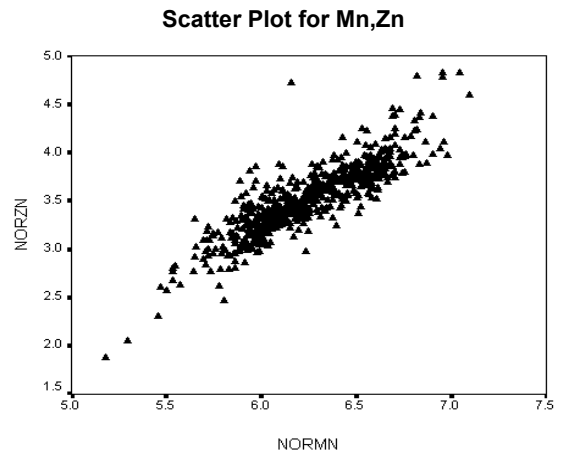
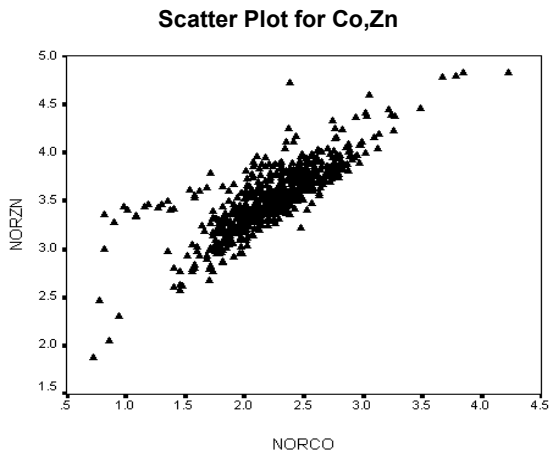
یکی دیگر از راههای بررسی ارتباط تغییرات عناصر با یکدیگر، رسم نمودار پراکنش (**Scatter Plot**) می‌باشد. زوج مرتب‌هایی از مقادیر دو متغیر که دارای توزیع دو متغیره یکسان باشند بر روی نمودار دو بعدی ترسیم می‌گردند. هر چه پراکندگی نقاط در نمودارهای پراکنش بیشتر باشد پیوند بین متغیرها ضعیف‌تر است. شکل (۴-۸) پراکنش مقادیر داده‌های خام نرمال شده برای چند زوج عنصری است که بیشترین ارتباط را نشان می‌دهد. در این نمودارها زوج عنصر  $(\text{Ti,Zn})$  بیشترین همبستگی را با یکدیگر نشان می‌دهد.

### بررسی‌های آماری چند متغیره:

هر تجزیه و تحلیل چند متغیره که بر روی بیش از دو متغیر انجام گیرد، می‌تواند در قالب آنالیزهای چند متغیره بیان شود. غالب تکنیکهای چند متغیره در اصل بسط و توسعه آنالیزهای تک متغیره می‌باشند و البته بعضی از روشهای چند متغیره تنها برای

پاسخگویی به مقاصد چند متغیره طراحی شده‌اند که از جمله این روشها می‌توان به آنالیز فاکتوری اشاره کرد.

**Fig (4-8) : Pearson Scatter Plot For Normal Raw Data**



تجربه نشان داده است که چنانچه ترکیبی از متغیرها به جای یک متغیر به کار گرفته شوند و از نتایج ترکیبی آنها استفاده شود امکان تشخیص هاله‌های مرکب ژئوشیمیایی در اطراف توده‌های کانساری به مراتب افزایش می‌یابد. و از طرفی اثرات خطاهای تصادفی در بکارگیری ترکیبی متغیرها نسبتاً کاهش می‌یابد. از دیگر مزایای استفاده از روشهای چند متغیره، کاهش تعداد متغیرها در مباحث داده‌پردازی و در نتیجه کاستن از تعداد نقشه‌هاست. با استفاده از این روشها امکان مقایسه متغیرها و کسب نتایج راحت‌تر خواهد بود. البته استفاده بهینه از روشهای چند متغیره در حالتی صادق خواهد بود که در پردازش داده‌ها با تعداد زیادی متغیر روبرو باشیم و تا حدودی امکان اخذ نتیجه از متغیرها به گونه منفرد غیر ممکن و یا توأم با خطای زیاد باشد. در این گزارش از روشهای چند متغیره مانند روشهای آنالیز خوشه‌ای و آنالیز فاکتوری و ... استفاده شده است.

### **آنالیز خوشه‌ای و تفسیر آن:**

به دلیل اینکه هر گروه از عناصر نسبت به یکسری از شرایط محیطی کم و بیش به طور مشابه حساسیت نشان می‌دهند، شناخت ارتباط و همبستگی ژنتیکی متقابل بین عناصر مختلف می‌تواند در شناخت دقیق‌تر تغییرات موجود در محیطهای ژئوشیمیایی به کار گرفته شود. ضمناً تجمع ژنتیکی بعضی از عناصر ممکن است به عنوان راهنمای مستقیم در تفسیر نوع نهشته‌ای که احتمالاً در ناحیه وجود دارد، به کار رود. در کل شناخت همبستگی ژنتیکی که در بین عناصر وجود دارد اطلاعات لازم را برای تفسیر هر چه صحیح‌تر داده‌های ژئوشیمیایی در اختیار می‌گذارد.

آنالیز خوشه‌ای یک روش آماری چند متغیره است که عناصر را بر اساس شباهت تغییرپذیری بین آنها در قالب دسته‌ها یا گروه‌هایی طبقه‌بندی می‌کند. دلایل زیادی برای ارزشمند بودن آنالیز خوشه‌ای وجود دارد، از جمله اینکه آنالیز خوشه‌ای می‌تواند در یافتن گروه‌های واقعی کمک کند و همچنین باعث کاهش تراکم داده‌ها شود. البته باید توجه داشت که آنالیز خوشه‌ای می‌تواند گروه‌های غیر قابل انتظاری را نیز ایجاد نماید که بیانگر روابط جدیدی خواهند بود و باید مورد بررسی قرار گیرند. در روش آنالیز خوشه‌ای از داده‌های خام نرمال شده استفاده شده است تا اثر مقادیر غیر همساز از جامعه اصلی و نیز اثر تغییر مقیاس داده‌ها از میان برود. نتایج حاصل از آنالیز خوشه‌ای عناصر مورد مطالعه در شکل (۴-۹) آورده شده است. با توجه به شکل می‌توان سه گروه اصلی را جدا نمود که بیانگر ارتباط پاراژنزی بین متغیرها باشد.

گروه اول: شامل عناصر **Zn, Ti, Mn, Co, Cr, B, Be, Sn, Bi, Mo, W**

گروه دوم: شامل عناصر **As, Sb, Se, Au, Hg**

گروه سوم: شامل عناصر **Sr, Ag, Pb, Ba, Cu, Ni**

*Fig(4-9): Cluster Analyse for Normal Raw Data*

Dendrogram using Complete Linkage

