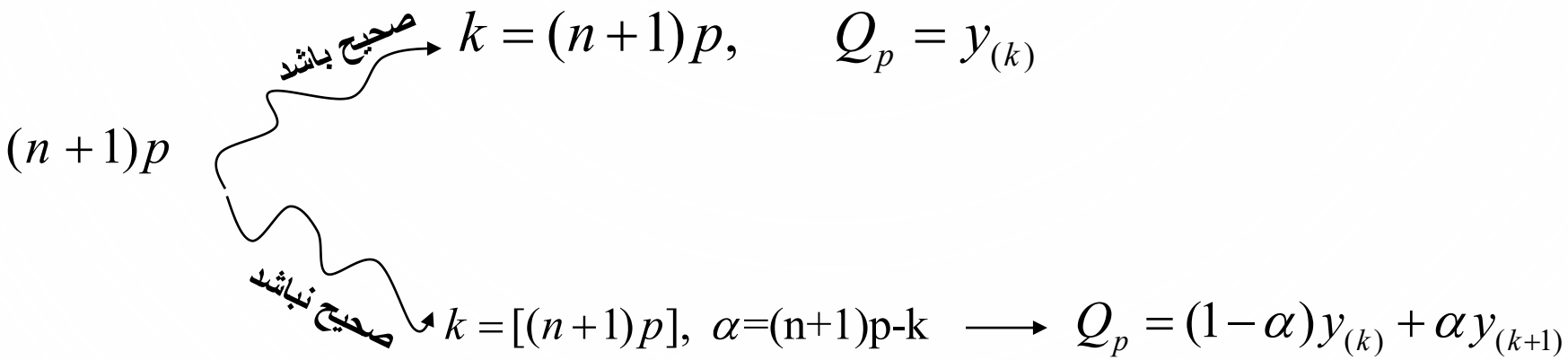
آمار توصیفی

ارائه دهنده: امیرحسین سبحانی

محاسبه چندك براي داده‌هاي گسسته

فرض كنيد $y_1, y_2, \cdots$ داده‌هاي ما باشند و شكل مرتب شده آنها را با $y_{(1)}, y_{(2)}, \cdots$ نمايش دهيم. براي محاسبه چندك



محاسبه چندك براي داده‌هاي پيوسته

با توجه به ستون فراواني تجمعي در جدول فراواني، رديفی را كه چندك (مثال p=0.25) در آن قرار دارد را مشخص مي‌كنيم.

رده	x_i	f_i	g_i
1/35-1/55	1/45	4	4
1/55-1/75	1/65	6	10
1/75-1/95	1/85	12	22
1/95-2/15	2/05	9	31
2/15-2/35	2/25	8	39
2/35-2/55	2/45	6	45
2/55-2/75	2/65	2	47
2/75-2/95	2/85	3	50
جمع		50	–

$$Q_p = L_{Q_p} + (\frac{np - g_{np}}{f_{Q_p}})I$$

$$(n)\times(p)$$
$$50\times 0.25 = 12.5$$

$Q_{0.25}$

L_{Q_p}: كران پايين رده چندك

g_{np} : فراواني تجمعي رده بلافاصله قبل از رده چندك

f_{Q_p} : فراواني رده چندك

I : طول رده

$$Q_{0.25} = 1.75 + (\frac{50\times 0.25 - 10}{12})\times 0.2 = 1.79$$

میانگین قدرمطلق انحراف ها

- میانگین قدر مطلق انحرافات: میانگین قدر مطلق انحراف که به آن انحراف متوسط می‌گویند به صورت زیر تعریف می‌شود.

$$AD = \frac{\sum_{i=1}^{k} f_i \left| x_i - \bar{x} \right|}{n}$$

- تذکر: گاهی ممکن است از میانه به جای میانگین برای محاسبه متوسط انحراف استفاد شود.

میانگین قدرمطلق انحراف ها

مثال: با توجه به جدول فراوانی زیر میانگین قدر مطلق انحراف ها را محاسبه نمایید:

x_i	f_i	$f_i x_i$	$f_i \lvert x_i - \bar{x} \rvert$
18.8	2	37.6	62.4
21.9	7	153.3	196.7
25	10	250	250
28.1	17	477.7	372.3
31.2	3	93.6	56.4
34.3	6	205.8	94.2
37.4	5	187	63
$\sum$	50	1405	1095

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{1405}{50} = 28.1$$

$$AD = \frac{\sum_{i=1}^{k} f_i \lvert x_i - \bar{x} \rvert}{n} = \frac{1095}{50} = 21.9$$

میانگین توان دوم انحراف ها (واریانس)

واریانس داده های یک جامعه به صورت زیر تعریف می گردد:

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

و واریانس نمونه ای به صورت زیر تعریف می گردد

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

همچنین انحراف معیار جامعه به صورت جذب مثبت واریانس و انحراف معیار نمونه ای به صورت جذر مثبت واریانس نمونه ای تعریف می شود:

$$\sigma = \sqrt{\sigma^2} \quad , \quad S = \sqrt{S^2}$$

میانگین توان دوم انحراف ها (واریانس)

- قضیه: واریانس را می‌توان به کمک فرمول زیر محاسبه نمود:

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

x_i	f_i	$f_i x_i$	$f_i x_i^2$
18.8	2	37.6	706.88
21.9	7	153.3	3357.27
25	10	250	6250
28.1	17	477.7	13423.37
31.2	3	93.6	2920.32
34.3	6	205.8	7058.94
37.4	5	187	6993.8
$\sum$	50	1405	40710.58

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{1405}{50} = 28.1$$

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} f_i x_i^2 - n\bar{x}^2\right) = \frac{1}{49}(40710.58 - 50(28.1)^2) = 25.103$$

میانگین توان دوم انحراف ها (واریانس)

- اگر به کمک تبدیل $y_i = ax_i + b$ داده‌های $x_1, x_2, \ldots, x_n$ را به $y_1, y_2, \ldots, y_n$ تبدیل کنیم رابطه زیر بین واریانس x_i ها و y_i برقرار است:

$$S^2{}_y = a^2 S^2{}_x \quad , S_y = a S_x \quad a > 0$$

$$\bar{y} = a\bar{x} + b$$

روش تبدیل داده ها جهت محاسبه واریانس

روش تبدیل داده ها جهت محاسبه واریانس: برای ساده تر کردن محاسبات جهت محاسبه میانگین و واریانس خصوصاً وقتی داده ها بزرگ باشند می توان از روش تبدیل داده ها استفاده نمود. در این روش به کمک تبدیل

$$y_i = \frac{x_i - a}{b} \quad i = 1,2,...,n$$

داده‌های $x_1, x_2, ..., x_n$ را به $y_1, y_2, ..., y_n$ تبدیل می‌کنیم. در اینجا a و b به ترتیب برای تغییر مبدا و تغییر واحد اندازه گیری استفاده میشود. معمولاً برای داده های پیوسته a را نماینده رده نمایی و b را طول رده یعنی I می گیرند. سپس میانگین و واریانس برای داده های جدید $y_1, y_2, ..., y_n$ محاسبه می شود و به کمک فرمول های زیر میانگین و واریانس داده های قبلی محاسبه می گردد:

$$S^2{}_x = b^2 S^2{}_y \quad , S_x = b S_y$$

$$\bar{x} = b\,\bar{y} + a$$

روش تبدیل داده ها جهت محاسبه واریانس

رده	x_i	f_i	y_i	$f_i y_i$	$f_i y_i^2$
17.25-20.35	18.8	2	-3	-6	18
20.35-23.45	21.9	7	-2	-14	28
23.45-26.55	25	10	-1	-10	10
26.55-29.65	28.1	17	0	0	0
29.65-32.75	31.2	3	1	3	3
32.75-35.85	34.3	6	2	12	24
35.85-38.95	37.4	5	3	15	45
$\sum$		50		0	128

$$y_i = \frac{x_i - 28.1}{3.1} \quad i = 1,2,...,7$$

$$\bar{y} = \frac{0}{50} = 0$$

$$S_y^2 = \frac{1}{49}(128 - 0) = 2.612$$

$$\bar{x} = 28.1 + 3.1(0) = 28.1$$

$$S_x^2 = (3.1)^2(2.012) = 25.103$$

$$S_x = \sqrt{S_x^2} = 5.01$$

ضریب تغییرات

ضریب تغییر: واریانس و انحراف استاندارد به واحد اندازه گیری بستگی دارد. برای مقایسه دو سری داده باید از شاخص هایی استفاده کنیم که به واحد اندازه گیری بستگی نداشته باشد یکی از این شاخص ها ضریب تغییر است که به کمک فرمول زیر تعریف می شود:

$$CV = \frac{S}{\bar{x}} \ , CV = \frac{\sigma}{\mu}$$

مثال: کارخانه ای دو نوع لامپ تولید می‌کند، لامپ نوع اول دارای میانگین طول عمر 200 ساعت و انحراف معیار 11 ساعت است و لامپ دوم دارای میانگین کارکرد 240 و انحراف معیار 12 ساعت است. کدام نوع لامپ بهتر است؟

جواب: لامپ دوم مناسب تر است. چون میانگین بیشتر و ضریب تغییر کمتری دارد.

$$CV_1 = \frac{S_1}{\bar{x}_1} = \frac{11}{200} = 0.055$$

$$CV_2 = \frac{S_2}{\bar{x}_2} = \frac{12}{240} = 0.05$$

Z-نمره (Z-SCORE)

Z-نمره (مقدار استاندارد): تعداد انحراف استاندارد هایی است که یک مقدار معین بالاتر و یا پایین تر از میانگین است.

Z-نمره به کمک فرمول زیر قابل محاسبه است:

$$z = \frac{x - \bar{x}}{S} \quad , \quad z = \frac{x - \mu}{\sigma}$$

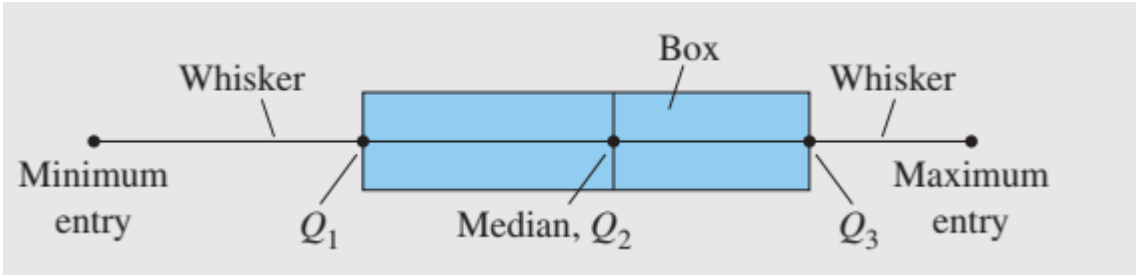
- ✓ Z- نمره فاقد مقیاس اندازه گیری است.
- ✓ یک داده، داده پرت است اگر Z-نمره آن بیش از 3 یا کمتر از -3 شود.
- ✓ اگر Z-نمره یک داده بیش از 2 یا کمتر از -2 شود غیر عادی است.

مثال: داده های زیر مربوط به تعداد ضربان های قلب یک نمونه است. مقدار پرت را به کمک Z-نمره بیابید.

AGE	GENDER (1=M)	PULSE	Z-score		
43	0	80	0.475079	Mean	74.04082
38	0	94	1.59119	Std	12.54356
69	0	58	-1.27881		
44	0	66	-0.64103		
72	0	56	-1.43825		
37	0	82	0.634524		
29	0	78	0.315635		
44	0	86	0.953413		
51	0	88	1.112857		
66	0	56	-1.43825		
46	0	36	-3.0327		
31	0	66	-0.64103		
35	0	84	0.793968		
55	0	76	0.15619		
22	0	78	0.315635		
51	0	64	-0.80048		
26	0	66	-0.64103		
62	0	78	0.315635		
43	0	60	-1.11936		
56	0	64	-0.80048		
54	0	84	0.793968		
20	0	82	0.634524		
63	0	70	-0.32214		

نمودار جعبه ای

از نمودار جعبه ای برای نمایش پراکندگی داده ها استفاده می شود. برای رسم این نمودار از پنج شاخص استفاده می کنیم: چارک اول، چارک سوم، میانه، مینیموم و ماکسیموم داده ها. برای پیدا کردن مینیموم و ماکسیموم، داده‌های پرت را کنار میگذاریم. مقدار ماکسیموم را بزرگترین داده‌ای که بیشتر از $Q_3 + 1/5IQR$ نباشد و مقدار مینیمم را کوچکترین داده‌ای که کمتر از $Q_1 - 1/5IQR$ نباشد در نظر می گیریم. برای رسم این نمودار، مستطیلی بین چارک اول و سوم رسم کرده و میانه را با خطی در جعبه مشخص می‌کنیم. همچنین مینیموم و ماکسیموم را با پاره خطی نمایش می دهیم واین پاره خط ها را به مستطیل مطابق شکل زیر وصل می‌کنیم. همچنین داده های پرت را نیز با دایره یا ستاره مشخص می کنیم.

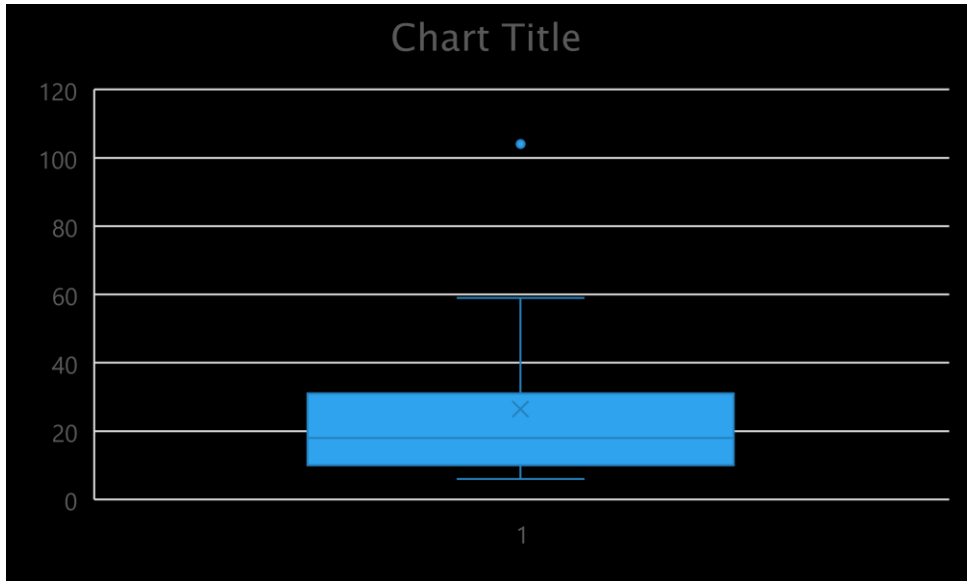


7 18 11 6 59 17 18 54 104 20 31 8 10 15 19

Lower half | Upper half

6 7 8 10 11 15 17 18 18 19 20 31 54 59 104

Q_1 Q_2 Q_3



$$IQR = Q_3 - Q_1 = 31 - 10 = 21$$

$$QD = \frac{1}{2} IQR = 10.5$$

$$Q_1 - 1.5\,IQR = -21.5$$

$$Q_3 + 1.5\,IQR = 62.5$$

داده های روبرو مربوط به وزن هنگام تولد نوزادان دچار مشکل ریه می‌باشد. این داده ها به دو گروه متفاوت بر اساس نوزادان نجات یافته و نوزادان از دست رفته تقسیم شده اند. به کمک نمودار جعبه ای تحلیل کنید که آیا وزن نوزادان رابطه ای با مرگ آنها داشته است؟

died	Survived
1.03	1.13
1.05	1.41
1.1	1.575
1.175	1.68
1.185	1.715
1.23	1.72
1.23	1.76
1.26	1.93
1.30	2.015
1.30	2.04
1.31	2.09
1.5	2.2
1.55	2.4
1.6	2.55
1.72	2.57
1.75	2.6
1.77	2.7
1.82	2.83
1.89	2.95
1.94	3.005
2.20	3.16
2.27	3.4
2.275	3.64
2.44	
2.5	
2.56	
2.73	

died	Survived
1.175	1.13
1.23	1.575
1.31	1.68
1.5	1.76
1.6	1.93
1.72	2.015
1.75	2.09
1.77	2.6
2.275	2.7
2.5	2.95
1.03	3.16
1.1	3.4
1.185	3.64
1.23	2.83
1.26	1.41
1.30	1.715
1.30	1.72
1.55	2.04
1.82	2.2
1.89	2.4
1.94	2.55
2.20	2.57
2.27	3.005
2.44	
2.56	
2.73	
1.05	

DIED

$$n = 27, p = 0.25, (n+1)p = 7$$

$$Q_1 = x_7 = 1.23$$

$$n = 27, p = 0.5, (n+1)p = 14$$

$$Q_2 = x_{14} = 1.6$$

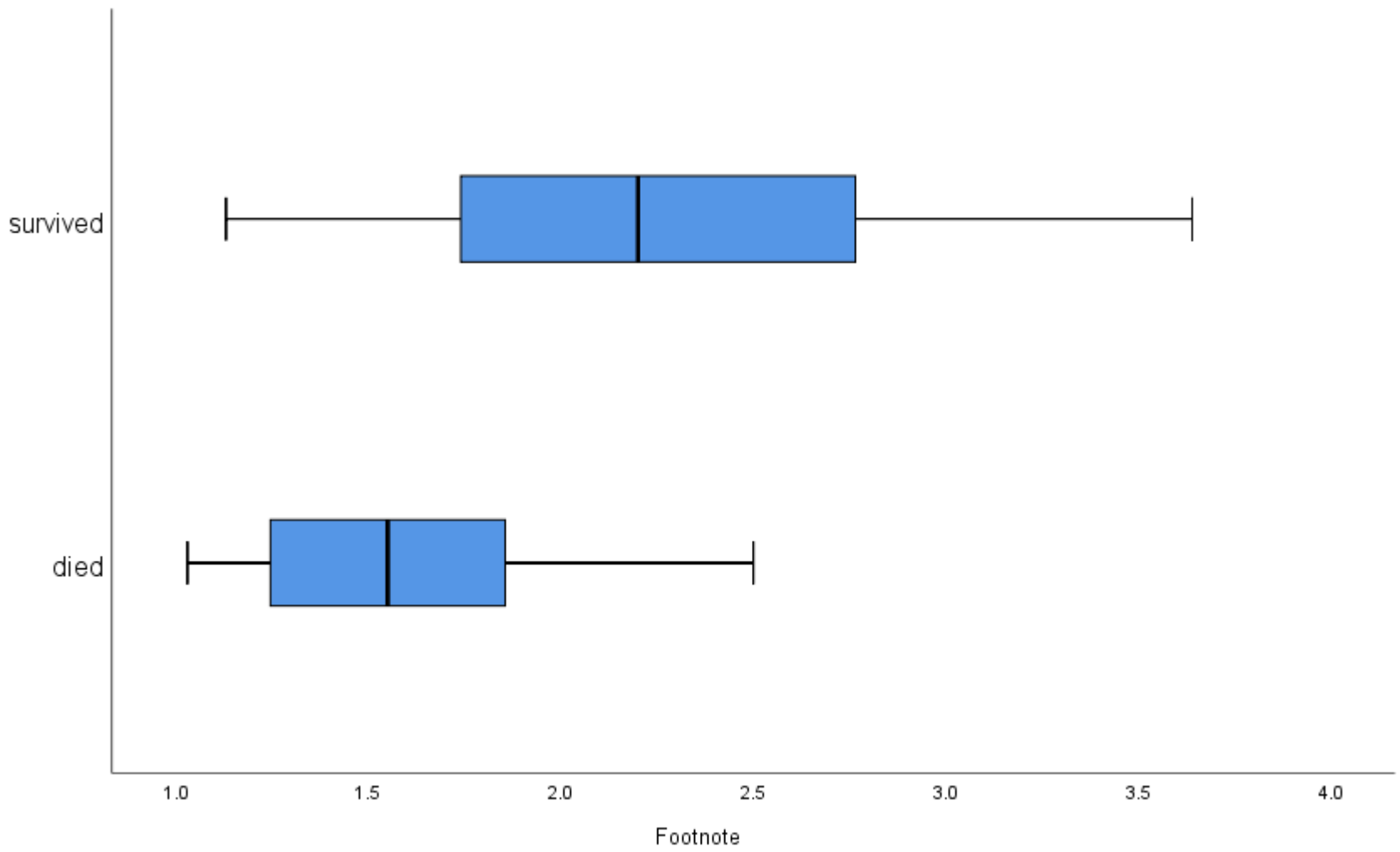
$$n = 27, p = 0.75, (n+1)p = 21$$

$$Q_3 = x_{21} = 2.2$$

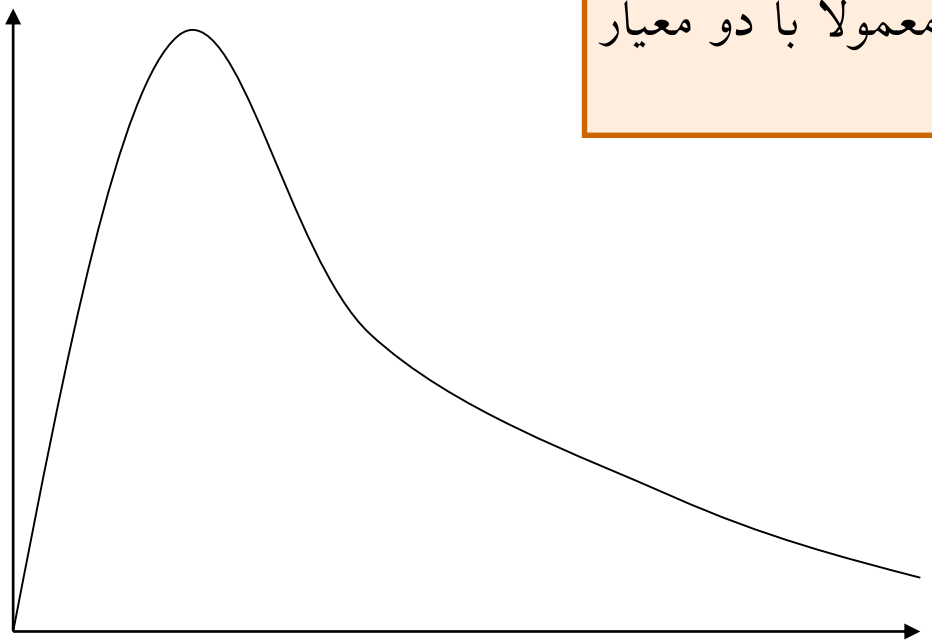
$$IQR = Q_3 - Q_1 = 2.2 - 1.23 = 0.97$$

$$QD = \frac{1}{2} IQR = 0.485$$

died	Survived
1.175	1.13
1.23	1.575
1.31	1.68
1.5	1.76
1.6	1.93
1.72	2.015
1.75	2.09
1.77	2.6
2.275	2.7
2.5	2.95
1.03	3.16
1.1	3.4
1.185	3.64
1.23	2.83
1.26	1.41
1.30	1.715
1.30	1.72
1.55	2.04
1.82	2.2
1.89	2.4
1.94	2.55
2.20	2.57
2.27	3.005
2.44	
2.56	
2.73	
1.05	

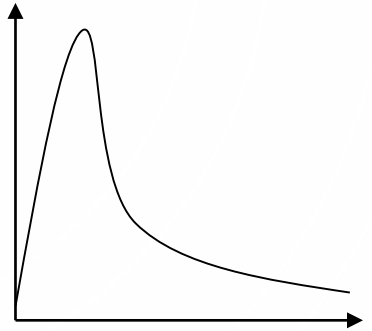


منحنيهاي فراواني در طبيعت تنوع زيادي دارند، اما بسياري از منحنيهاي فراواني يا متقارن هستند يا چوله و يا برجسته و يا پخ. ايده آلترين منحني فراواني متقارن، منحني فراواني نرمال استاندارد است. در طبيعت، عموما منحني فراواني متقارن ايده آل كمتر يافت مي‌شود ميزان انحراف از تقارن ايده آل را معمولا با دو معيار چولگي و برجستگي مي‌سنجند.

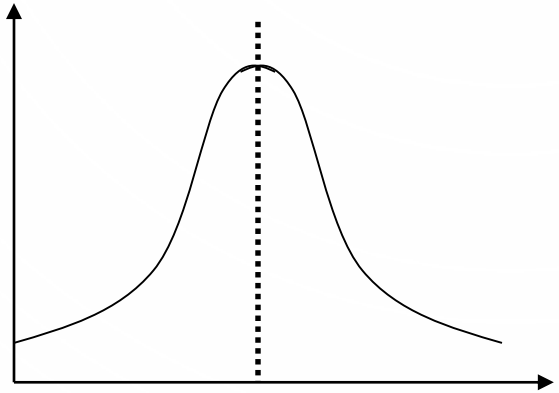


منحنيهاي فراواني:

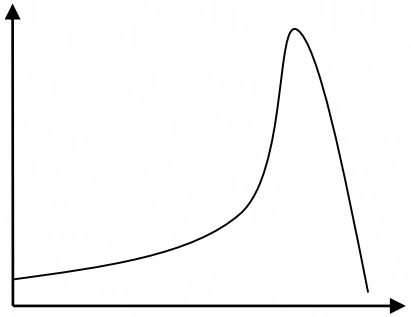
نامتقارن/ چوله به راست
Skewed to right



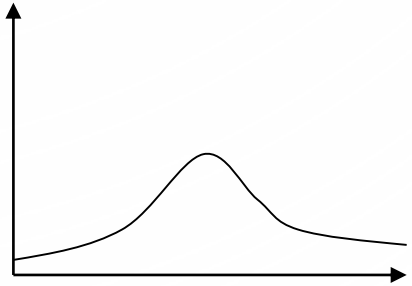
منحني نرمال استاندارد
Standard normal



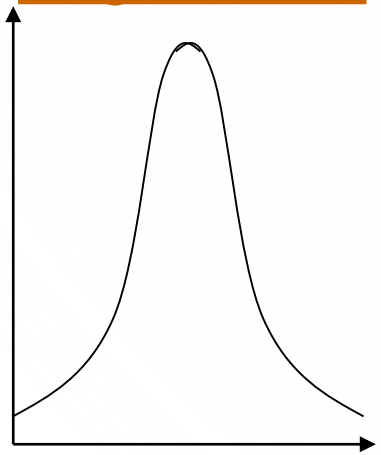
نامتقارن/ چوله به چپ
Skewed to left



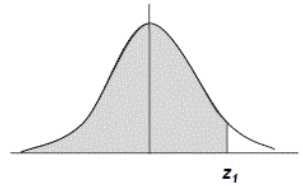
پخ
low kurtosis



برجسته
High kurtosis



Standard Normal Distribution

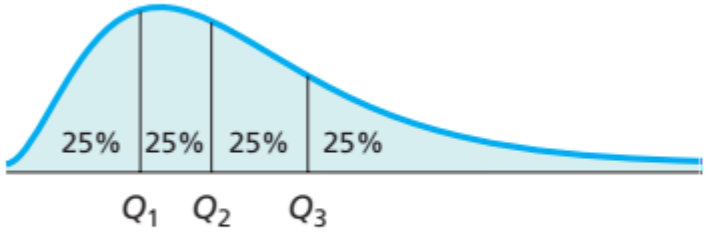


$$p(z \le z_1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_1} e^{-\frac{1}{2}z^2} dz$$

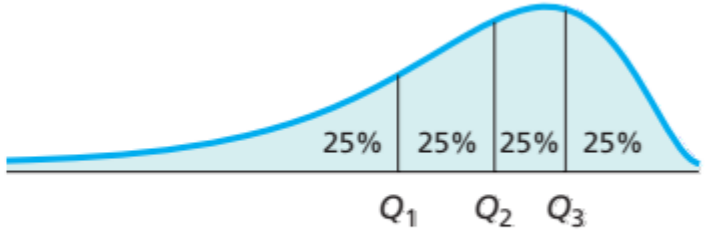
z_1	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

چولگی

چولگی معیاری از سنجش تقارن و یا عدم تقارن توزیع داده ها در منحنی فراوانی به حساب می آید. توزیع داده ها را چوله گوییم اگر متقارن نباشد و در یک سمت بیش از سمت دیگر گسترش پیدا کرده باشد.



(c) Right skewed



(d) Left skewed

چولگی منفی (چولگی به چپ): اگر داده‌های با فراوانی بالا در سمت راست منحنی فراوانی باشند و دم منحنی به سمت چپ کشیده شده باشد منحنی را چوله به چپ میگویند

چولگی مثبت (چولگی راست): اگر اگر داده های با فراوانی بالا در سمت چپ منحنی فراوانی قرار داشته باشند و دم منحنی فراوانی به سمت راست کشیده شده باشد و گسترش پیدا کرده باشد در این صورت منحنی راچوله به راست میگویند

تذکر: در توزیع متقارن میانه نما و میانگین بر هم منطبق است

چولگی

تذکر: توزیع های چوله به راست ار توزیع های چوله به چپ بیشتر ظاهر می‌شوند زیرا معمولا داده های غیر عادی بزرگ هستند.

باتوجه به اینکه میانگین به سادگی تحت تاثیر داده های پرت قرار می‌گیرد، در توزیع های چوله به چپ مد از میانگین بزرگتر است و در توزیع های چوله به راست، مد از میانگین کوچکتر است.

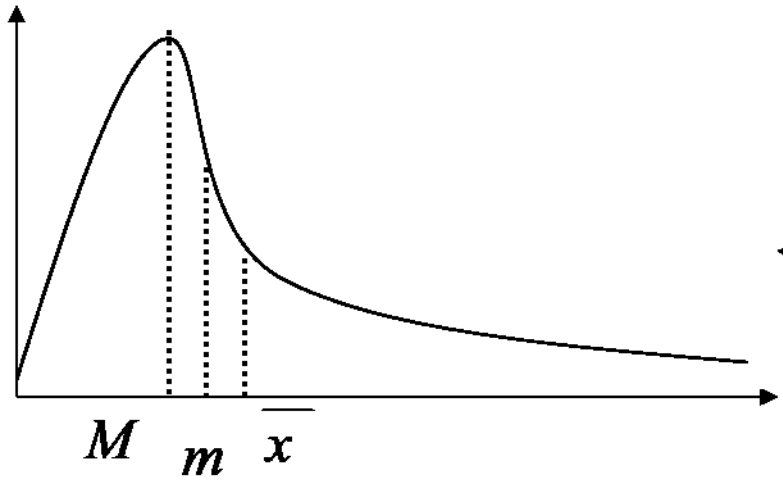
در حالت کلی در بیشتر مواقع اگر توزیع چوله به راست باشد:

$$M < m < \overline{x}$$

,و اگر توزیع چوله به چپ باشد معمولا داریم:

$$\overline{x} < m < M$$

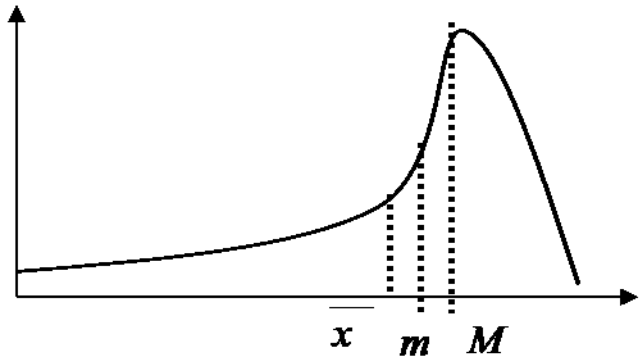
چولگی



$M < m < \bar{x}$

به راست یا مثبت

چولگي:



$M > m > \bar{x}$

به چپ یا منفی

چولگی

معیار چولگی کارل پیرسون: ضریب چولگی کارل پیرسون بر اساس این واقعیت تعریف شده است که در توزیع چوله به راست مد کمتر از میانگین و در توزیع چوله به چپ مد بیشتر از میانگین است:

$$SK_1 = \frac{\bar{x} - M}{S}$$

وقتی از مقدار مد اطلاعی نداشته باشیم، از فرمول دوم پیرسون استفاده می‌کنیم:

$$SK_2 = \frac{3(\bar{x} - m)}{S}$$

اگر ضریب پیرسون منفی باشد توزیع چولگی منفی (به چپ) و اگر مثبت باشد چولگی به راست دارد.

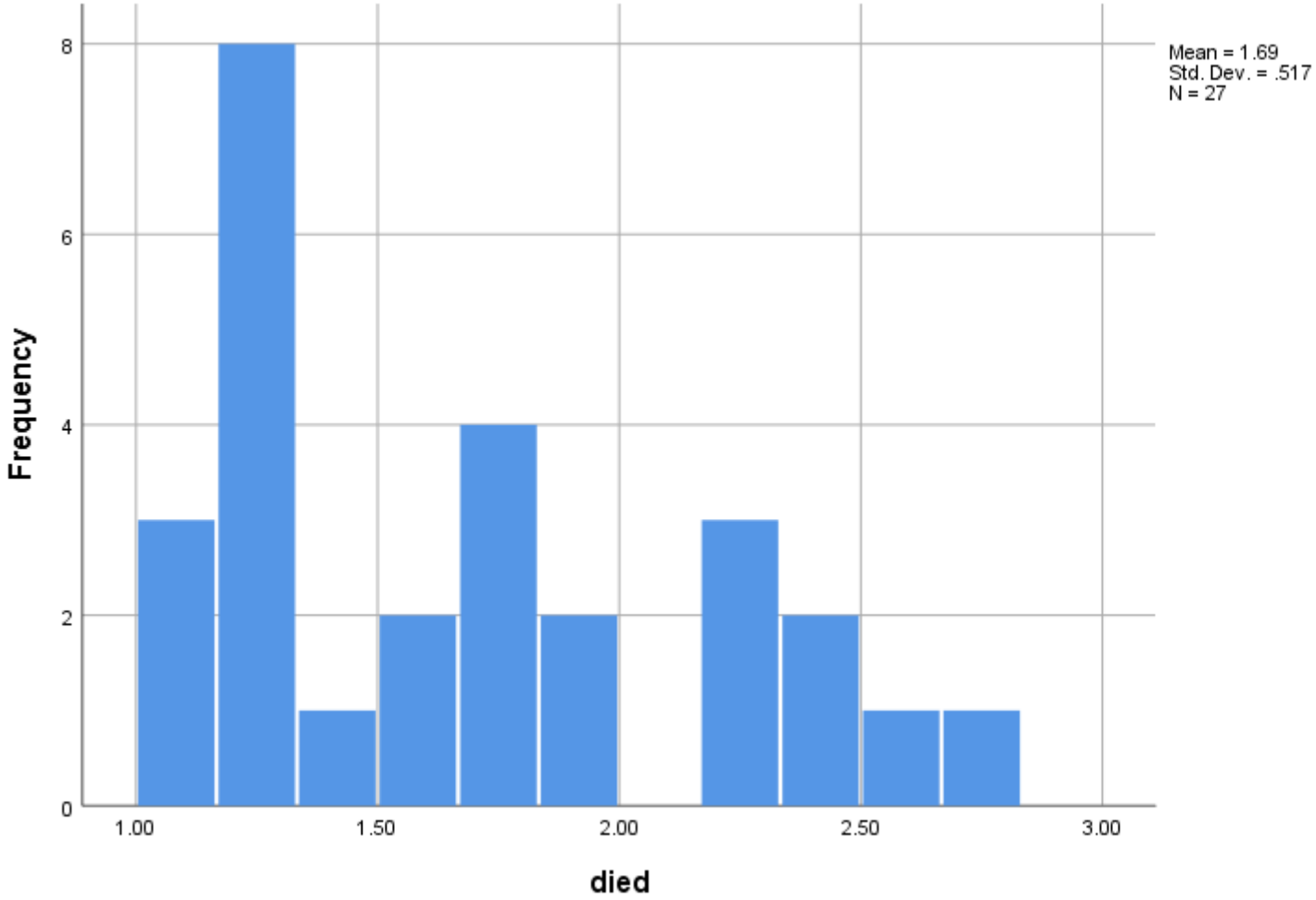
چولگی

ضریب چولگی گشتاوری:

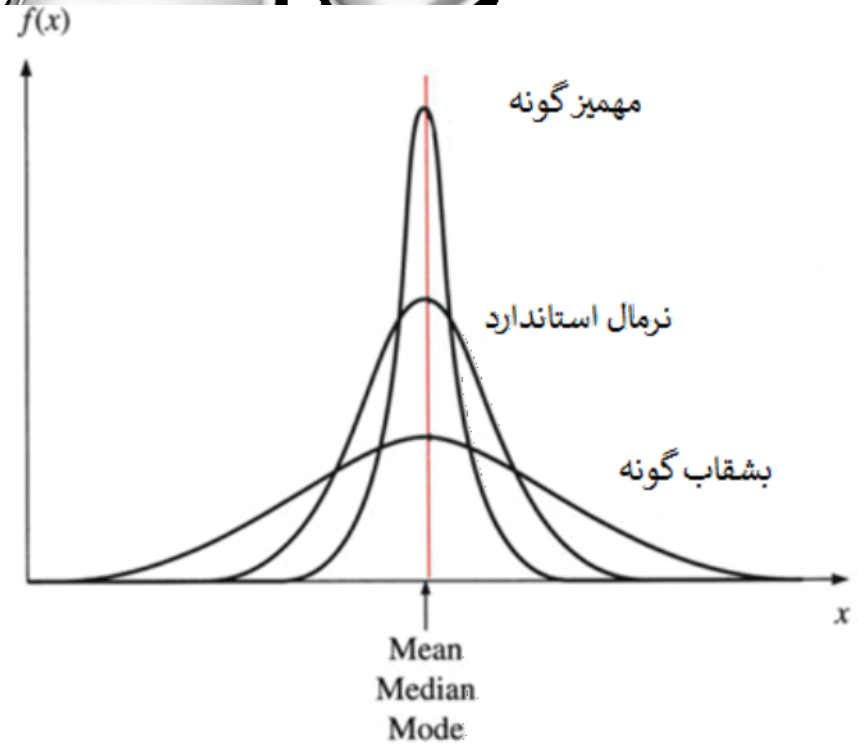
$$\alpha_3 = \frac{\mu_3}{\sigma} \, , \, \mu_3 = \frac{\sum f_i (x_i - \mu)^3}{n}$$

ضریب چولگی چارکی:

$$SQ = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$


Mean = 1.69
Std. Dev. = .517
N = 27
Frequency
0
2
4
6
8
1.00
1.50
2.00
2.50
3.00
died

کشیدگی



کشیدگی یک توزیع معیار سنجش برآمدگی توزیع نسبت به توزیع نرمال استاندا

توزیع ها از نظر کشیدگی سه نوع هستند:

توزیع نرمال استاندارد: که کشیدگی آن را معیار قرار داده و توزیع های دیگر را با آن می‌سنجیم.

توزیع با کشیدگی بالا (مهمیزگونه): که کشیدگی آن از کشیدگی توزیع نرمال استاندارد بیشتر است.

توزیع با کشیدگی منفی (بشقاب گونه): که کشیدگی آن از توزیع نرمال استاندارد کمتر است.

ضرایب کشیدگی

ضریب کشیدگی گشتاوری:

$$\alpha_4 = \frac{\mu_4}{\sigma^4} - 3$$

ضریب کشیدگی چندکی:

$$E = \frac{QD}{Q_{0.9} - Q_{0.1}} - 0.263$$