dove i rappresenta l'i-esimo insieme di misure e j l'osservazione nell'i-esimo insieme.
Comunque la media delle medie anche conosciuta come media della popolazione puo' essere definita come:

$$\bar{\bar{X}} = \frac{\sum_i^m \bar{X}_i}{m} \simeq \frac{m\bar{x}}{m} \simeq \bar{x}$$

dove $\bar{X} \simeq \bar{X}_1 \dots = \bar{X}_m = \underset{\text{costanti}}{m}$
Se la deviazione in un insieme e' denotata con dij mentre quella tra le medie dei diversi insiemi e' denotata con Di, allora

$$\sigma^2(x) = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} d_{ij}^2$$

e

$$\sigma^2(\bar{X}_n) = \frac{1}{m} \sum_{i=1}^{m} D_i^2 \qquad (7)$$

dove $\sigma^2(x)$ e $\sigma^2(\bar{X}_n)$ rappresentano rispettivamente le variazioni delle singole misure e quella delle medie dei gruppi di misure. Si ha: $D_i = (\bar{X}_i - \bar{\bar{X}}) \simeq \frac{1}{n}\sum_{j=1}^{n} x_{ij} - \frac{n\bar{x}}{n}$ $(\because \bar{\bar{X}} \simeq \bar{x})$

$$\simeq \frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \bar{x}) \simeq \frac{1}{n}\sum_{j=1}^{n} d_{ij} \qquad (8)$$

Sostituendo il valore di Di dall'equazione (8) nella (7) si ottiene:

$$\sigma^2(\bar{X}_n) = \frac{1}{m}\sum_{i=1}^{m}\left\{\frac{1}{n}\sum_{j=1}^{n} d_{ij}\right\}^2 = \frac{1}{mn^2}\sum_{i=1}^{m}\left\{\sum_{j=1}^{n} d_{ij}\right\}^2 \qquad (9)$$

La doppia somma dell'equazione (9) contiene due differenti tipi di termini. Ci sono alcuni termini in cui dij e' elevato al quadrato mentre altri contengono il prodotto di due differenti dij. Poiche' c'e' una eguale probabilita' che i termini siano positivi e negativi, al limite, quando si considera un elevato numero di osservazioni nxm, la somma dei termini con il prodotto di due differenti dij tende a zero.

$$\sum_{i=1}^{m}\left\{\sum_{j=1}^{n} d_{ij}\right\}^2 \simeq \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}^2$$

La (9) diviene dunque

$$\sigma^2(\bar{X}_n) = \frac{1}{mn^2}\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}^2 = \frac{1}{n}\left\{\frac{1}{m\cdot n}\sum_{i}^{m}\sum_{j}^{n} d_{ij}^2\right\} = \frac{\sigma^2(x)}{n}$$

5. VALUTAZIONE DELLA MEDIA E DELLA DEVIAZIONE STANDARD MEDIANTE CODIFICHE INTERMEDIE

5.1 Metodo del cambio dell'origine dei dati

La procedura per operare mediante il cambiamento dell'origine dei dati e' anche chiamata "metodo della media assunta". Talvolta intuitivamente noi utilizziamo questo metodo di calcolo.
Si considerino, ad esempio 4 osservazioni:

100,20 - 100,30 - 100,15 - 100,25
e si desidera determinare la media. Invece di fare il totale 400,80, quindi dividere per 4 ottenendo 100,20, e' piu facile fare la media di 0,20 - 0,30 - 0,15 - 0,25, ottenendo 0,20, quindi sommare 100. In effetti stiamo operando su valori X' dove X'=X-100; in altre parole abbiamo spostato l'origine dei dati ad Xa=100.
In generale, noi possiamo scrivere i valori osservati nella seguente forma

$$Xi = X'i + Xa \qquad (10)$$

dove Xa e' la media assunta (pari ad un valore costante). La media valutata come nell'equazione precedente puo' essere ottenuta facendo la somma e dividendo per il numero totale di osservazioni:

$$\frac{x_i}{n} = \frac{1}{n}\sum(x'_i + X_A) \quad \text{oppure} \quad \bar{X}_n = \bar{X}'_n + X_A \qquad (11)$$

dove $\bar{X}'_n$ e' il valore medio dei valori ottenuti spostando l'origine dei dati. Similmente la varianza puo' essere ottenuta attraverso la sostituzione dei valori di x_i ed $\bar{X}_n$ dalle equazioni (10) e (11) nella seguente equazione:

$$\sigma^2(x) = \frac{1}{n}\sum(x_i - \bar{X}_n)^2 = \frac{1}{n}\sum\left\{(x'_i + X_A) - (\bar{X}'_n + X_A)\right\}^2 = \frac{1}{n}\sum(x'_i - \bar{X}'_n)^2 = \sigma^2(x')$$

$$\text{ovvero;} \quad \sigma(x) = \sigma(x')$$

5.2 Metodo del cambiamento dell'ordine di grandezza dei dati

Talvolta, un cambio nell'ordine di grandezza facilita il calcolo riducendo il numero di digits richiesto per le manipolazioni aritmetiche.
Ad esempio i valori di 4 osservazioni sono 400, 600, 200, 500 e si vuole determinare la loro media. Si potrebbero trascurare gli zeri e calcolare la media di 4,6 2 e 5 che e' 4,25, dopo di cio' si dovrebbero rimettere gli zeri e presentare la media come 425. In questo esempio e' stato eseguito un cambiamento dell'ordine di grandezza dei dati originali attraverso un appropriato fattore di moltiplicazione che nel caso suindicato e' 1/100. In tal modo possiamo scrivere i valori assegnati nella seguente forma generale
$Xi = a \cdot X'i$,
dove a e' un appropriato fattore di moltiplicazione. La media e la deviazione standard possono essere trovate in maniera simile a quella utilizzata nel paragrafo precedente, ed i risultati sono i seguenti:

$$\bar{X}_n = a \cdot \bar{X}'_n \qquad \sigma_n(x) = a \cdot \sigma_n(x')$$

6. LA DISTRIBUZIONE NORMALE

6.1 Introduzione

In generale agli strumenti di misura sono associati un numero di fattori che sono causa di errori casuali (random errors). Comunque, l'ampiezza degli errori, presi individualmente, e' usualmente piccola. Ora, se le misure fatte sono molte, i dati mostrano una distribuzione continua; questa potrebbe essere facilmente rappresentata mediante un istogramma normalizzato che presenta la frequenza relativa per ogni intervallo come ordinate, ed i valori misurati come ascissa. Se l'ampiezza dell'intervallo e' piccola e le ordinate dei vari punti medi delle classi sono congiunte da una curva regolare, la distribuzione risultante e' chiamata limite della distribuzione di frequenza. Tale distribuzione e' comunemente denominata "distribuzione normale o gaussiana".
Tale distribuzione e' rappresentata matematicamente dalla funzione :

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{(x-\bar{X})^2}{2\sigma^2}\right\}$$

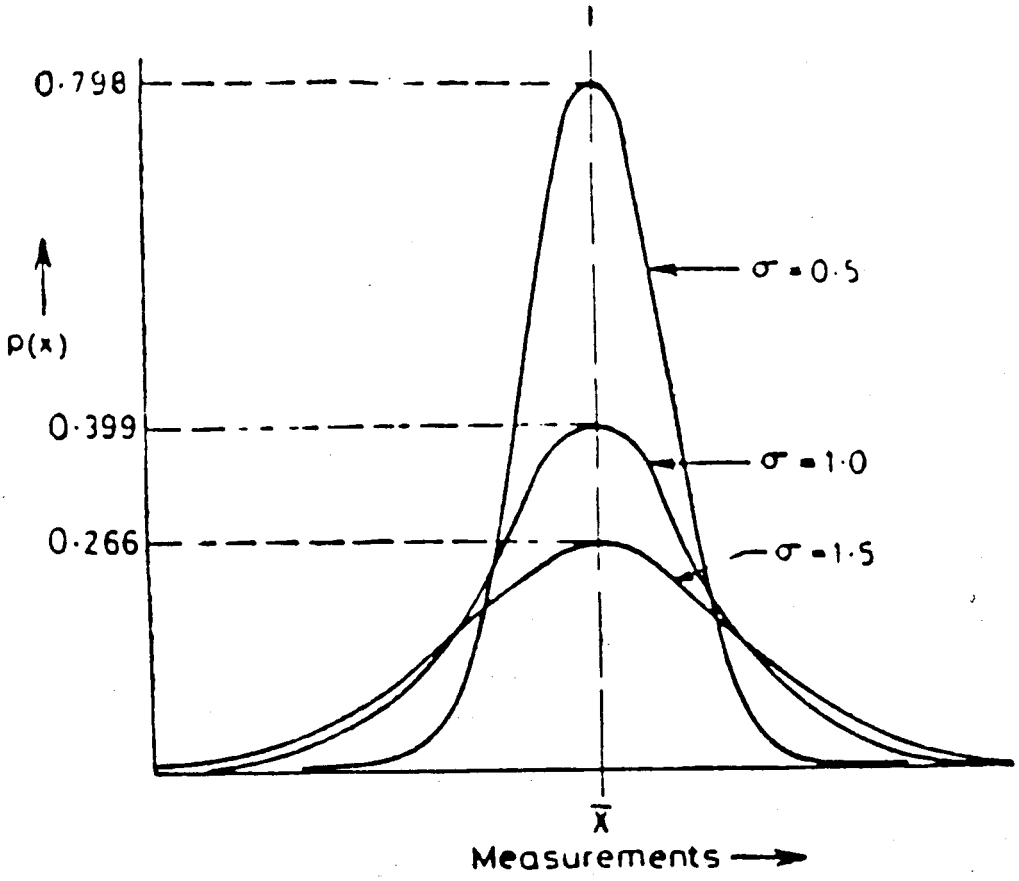
dove: p(x) e' la funzione densita' di probabilita' che, per un dato intervallo, rappresenta la frequenza relativa di occorrenza del valore misurato in quell'intervallo;
σ è la deviazione standard dei valori misurati;
$\bar{X}$ e' la media dei valori misurati.

Tale modello di distribuzione è utilizzato nel processo decisionale per valutare le probabilità che un valore misurato cada in un determinato intervallo. In alternativa, se il livello di probabilità (o livello di confidenza) è preassegnato, essa consente di valutare la dispersione permessa dei dati intorno al valore medio. Inoltre le proprietà di tale distribuzione sono utilizzate per confrontare diversi gruppi di misure attraverso criteri statistici noti come test di significatività. E' possibile infine determinare la "bonta'" o l'idoneita' dei valori misurati attraverso il confronto con i valori attesi dalla distribuzione normale. Questo viene fatto utilizzando il test del χ^2 (chi quadro). Tale criterio e' anche applicabile per valutare se una distribuzione qualunque, non normale, si adatta o meno a qualche altra distribuzione teoricamente conosciuta.

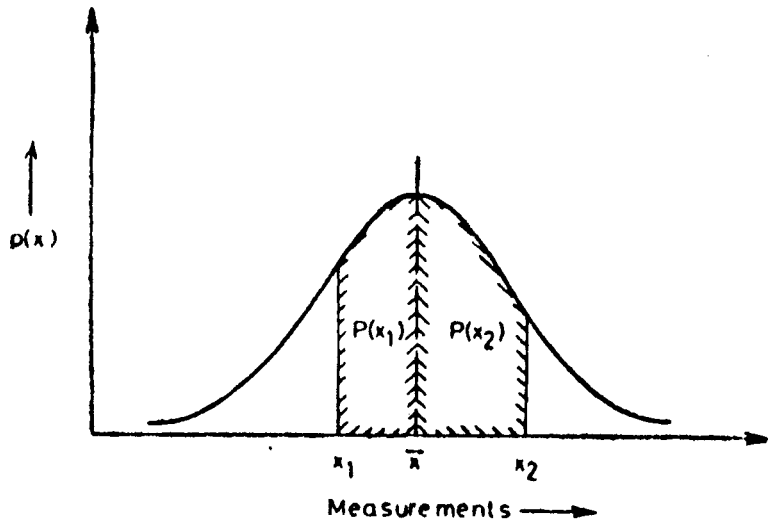
6.2 Proprieta' della distribuzione normale

Ogni tipica curva gaussiana ha le seguenti caratteristiche (fig.6):

1) essa ha un massimo per $x=\bar{X}$ (cioè in corrispondenza del valore medio);


0·798
0·399
0·266
P(x)
σ = 0·5
σ = 1·0
σ = 1·5
X̄
Measurements

– FIG. 6 –


p(x)
P(x1)
P(x2)
x1
X̄
x2
Measurements

– FIG. 7 –

2) i punti di flesso sono a $x=\bar{X}\pm\sigma$;

3) a causa della sua simmetria, la mediana e' uguale alla media. Inoltre, giacche' la media capita sul valore di picco della curva, essa rappresenta anche la moda;

4) l'asse delle ascisse e' un asintoto per la curva;

5) l'area sottesa alla curva e' unitaria, cioe':

$$\int_{-\infty}^{+\infty} P(x)\,dx = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{+\infty} \exp\left\{-\frac{(x-\bar{X})^2}{2\sigma^2}\right\}dx = 1$$

6) Per uno stesso valore di media, la curva assume un picco appuntito per piccoli valori di σ ed un picco appiattito per elevati valori di σ . Se σ è piccolo, ciò significa che c'è una piccola dispersione tra i dati e, di conseguenza, si hanno più valori concentrati intorno al valor medio e la curva assume un massimo più elevato. D'altra parte il massimo è inversamente proporzionale a σ:

$$\max(P(x)) = \frac{1}{\sqrt{2\pi}\,\sigma} = \frac{0{,}399}{\sigma}$$

7) La probabilita' che la media cada tra x_1 e x_2 e' uguale all'area della curva gaussiana compresa tra le ascisse X1 ed X2, come mostrato in fig. ; tale valore e' noto col nome di integrale gaussiano di probabilita' nell'intervallo (X1, X2) e si indica con $\left|P(x)\right|_{x_1}^{x_2}$. Svolgendo questo integrale si ottiene

$$\left|P(x)\right|_{x_1}^{x_2} = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{\frac{-(x-\bar{X})^2}{2\sigma^2}\right\}dx = \left|P(x)\right|_{x_1}^{\bar{X}} + \left|P(x)\right|_{\bar{X}}^{x_2} = P(x_1)+P(x_2)$$

dove $P(x_1)$ e $P(x_2)$ sono gli integrali gaussiani tra x_1 e $\bar{X}$ e tra $\bar{X}$ ed x_2 rispettivamente.

6.3 Area sottesa dalla curva normale

L'area sottesa alla curva di distribuzione normale tra i limiti $-\infty, +\infty$ e' l'integrale gaussiano della probabilita' con cui i valori misurati cadono nel suddetto intervallo. Di conseguenza, l'integrale gaussiano dovrebbe essere unitario poichè tutti i possibili valori misurati cadono tra $-\infty$ e $+\infty$. Per ottenere questo risultato, si integra l'equazione della distribuzione gaussiana, come segue:

l'area elementare dA della curva e' $dA = P(x)dx = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{\frac{-(x-\bar{X})^2}{2\sigma^2}\right\}dx$

integrando tra $(-\infty, +\infty)$ otteniamo l'area sottesa alla curva:

$$[A]_{norm} = \int_{-\infty}^{+\infty}\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{\frac{-(x-\bar{X})^2}{2\sigma^2}\right\}dx$$

operando la sostituzione $z=(x-\bar{X})/\sigma$ (da cui $dz=dx/\sigma$) si ha

$$[A]_{norm} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} exp\left(\frac{-z^2}{2}\right) dz$$

Il termine $\int_{-\infty}^{+\infty} exp(-z^2/2)dz$ e' conosciuto come funzione normale di errore ed il suo valore, calcolato attraverso integrazione grafica oppure in maniera differente, e' di $\sqrt{2\pi}$.
Sostituendo il valore della funzione d'errore cosi' ottenuta nell'equazione precedente, abbiamo come risultato che l'area sottesa alla curva normale e' unitaria.

6.4 Distribuzione normale standardizzata

Allo scopo di ridurre differenti distribuzioni normali in una forma generale, e' usuale standardizzarle spostando l'origine delle coordinate della media e scegliere una scala sull'asse delle ascisse in funzione di σ.
Una conveniente variabile e' $z=(x-X)/\sigma$, detta variabile normale standardizzata.
La probabilita' di occorrenza nell'intervallo $[x_1, x_2]$ ricavabile dall'equazione della distribuzione normale e':

$$P(x)\{x_1 \leq x \leq x_2\} = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{(x-\bar{X})^2}{2\sigma^2}\right\} dx \qquad (12)$$

Questa equazione puo' essere trasformata secondo il cambiamento di variabili proposto ottenendo:

$$z \triangleq \frac{x-\bar{X}}{\sigma}\ ;\ z_1 = \frac{x_1-\bar{X}}{\sigma}\ ;\ z_2 = \frac{x_2-\bar{X}}{\sigma}\ ;\ z_3 = \frac{dx}{\sigma}$$

da cui

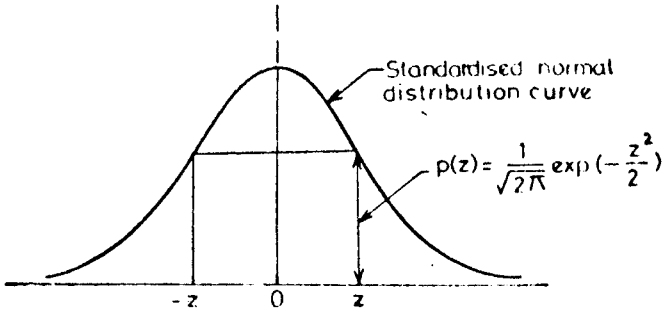
$$P(z)\{z_1 \leq z \leq z_2\} = \int_{z_1}^{z_2} \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{z^2}{2}\right\} dz \qquad (13)$$

Tale formula e' denominata equazione normale standardizzata.
Confrontando le (12) e (13) e' ovvio che ogni qualvolta e' richiesta la probabilita' nell'intervallo $[x_1, x_2]$, è richiesto il calcolo dell'integrale in (12). Quest'ultimo e' piuttosto complesso comparato all'espressione (13) in quanto nel secondo caso, l'uso di valori tabulati della funzione di errore, cioe' $\int exp(-z^2/2)dz$ semplifica notevolmente i calcoli.
In secondo luogo, la forma della curva (12) dipende dai valori di $\bar{X}$ e σ ed e' differente da caso a caso. Al contrario l'equazione (13) mostra che tutte le distribuzioni normali in x, con qualsiasi valore di X e σ, si riducono alla stessa forma di distribuzione normale standardizzata in termini di variabile normale standardizzata. Essa e' caratterizzata dall'avere valore medio uguale a zero e deviazione standard uguale ad 1.
L'integrale gaussiano di probabilità $P(z)=\int \rho(z)dz$ e' dunque tabulato rispetto ai valori di Z (vedi tab. I e tab II).

Table 18.1 Normal Probability Density Function $p(z)$

$$p(z)=\frac{1}{\sqrt{2\pi}}\exp(-z^2/2)$$

Each entry in the table indicates normal probability density function $p(z)$ corresponding to $\pm z$

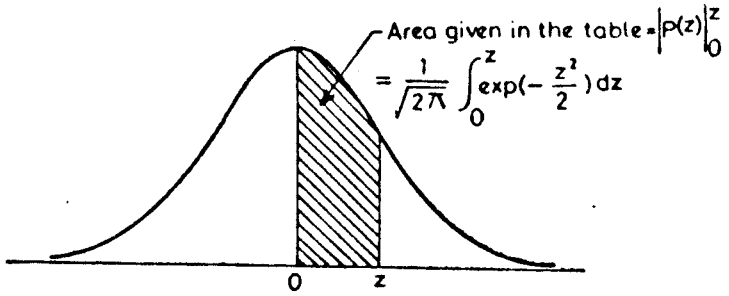


To illustrate: the ordinate $p(z)$ of the standardised normal distribution curve corresponding to $z=\pm 1.0$ is 0.2420

$\pm z$	0.00	0.02	0.04	0.06	0.08
0.0	0.3989	0.3989	0.3986	0.3982	0.3977
0.1	0.3970	0.3961	0.3951	0.3939	0.3925
0.2	0.3910	0.3894	0.3876	0.3857	0.3836
0.3	0.3814	0.3790	0.3765	0.3739	0.3712
0.4	0.3683	0.3658	0.3621	0.3589	0.3555
0.5	0.3521	0.3485	0.3448	0.3410	0.3372
0.6	0.3332	0.3292	0.3251	0.3209	0.3166
0.7	0.3123	0.3079	0.3034	0.2989	0.2943
0.8	0.2897	0.2850	0.2803	0.2756	0.2709
0.9	0.2661	0.2613	0.2565	0.2516	0.2468
1.0	0.2420	0.2371	0.2323	0.2275	0.2227
1.1	0.2179	0.2131	0.2083	0.2036	0.1989
1.2	0.1942	0.1895	0.1849	0.1804	0.1758
1.3	0.1714	0.1669	0.1626	0.1582	0.1539
1.4	0.1497	0.1456	0.1415	0.1374	0.1334
1.5	0.1295	0.1257	0.1219	0.1182	0.1145
1.6	0.1109	0.1074	0.1040	0.1006	0.0973
1.7	0.0940	0.0909	0.0878	0.0848	0.0818
1.8	0.0790	0.0761	0.0734	0.0707	0.0681
1.9	0.0656	0.0632	0.0608	0.0584	0.0562
2.0	0.0540	0.0519	0.0498	0.0478	0.0459
2.1	0.0440	0.0422	0.0404	0.0387	0.0371
2.2	0.0355	0.0339	0.0325	0.0310	0.0297
2.3	0.0283	0.0270	0.0258	0.0246	0.0235
2.4	0.0224	0.0213	0.0203	0.0194	0.0184
2.5	0.0175	0.0167	0.0158	0.0151	0.0143
2.6	0.0136	0.0129	0.0122	0.0116	0.0110
2.7	0.0104	0.0099	0.0093	0.0088	0.0084
2.8	0.0079	0.0075	0.0071	0.0067	0.0063
2.9	0.0060	0.0056	0.0053	0.0050	0.0047
3.0	0.0044	0.0042	0.0039	0.0037	0.0035

Table 18.2 Integral Gaussian Probability $p(z)$

Each entry in the table indicates the area under the standard normal curve from 0 to z



To illustrate: area under the standard normal curve between the maximum ordinates and a point 1.96 standard deviations away is 0.4750

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0 1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2258	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2518	0.2549
0.7	0.2580	0.2612	0.2642	0.2673	0.2704	0.2734	0 2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2996	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0 4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0 4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.5000	0.5000	0.5000
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

E' da notare che mentre si usano le tavole per valutare le probabilita' riferite alle varie distribuzioni normali, e' consigliabile disegnare l'area cui corrisponde la probabilita' richiesta. Inoltre, bisogna stare attenti nell'uso delle tabelle, in quanto non tutte danno gli stessi valori di area. Alcune tabelle danno l'area tra 0 e z, altre la danno tra -z e z, altre ancora da z a ∞ o da $-\infty$ a z.

6.5 Livelli di confidenza

Si e' gia' detto che l'area sottesa alla distribuzione normale tra$(-\infty,+\infty)$ e' unitaria. In realta' dovrebbe essere cosi' in quanto la probabilita' che tutti i possibili valori misurati cadano tra$(-\infty,+\infty)$ deve essere unitaria. Nella pratica, si specifica generalmente un certo range di valori ammessi intorno al valore medio e si determina la probabilita' con cui il valore misurato cade in tale range. Tale probabilita' puo' essere valutata calcolando l'area sottesa dalla curva normale di distribuzione in quell'intervallo prefissato. Quando questa probabilita' viene espressa in percentuale essa prende il nome di livello di confidenza.
Ad esempio, si vuole valutare qual'e' la probabilita' che un valore misurato cada nell'intervallo $[X-\sigma, X+\sigma]$; usando la tabella si trova che questa probabilita' e' uguale a 0,6826. Dunque si puo' dire che le probabilita' che un valore cada nell'intervallo prefissato siano di 2:1, (in realtà 68,26:(100-68.26)). Si puo' dire che il livello di confidenza in tal caso e' del 68,26%.
Generalmente sia le tabelle sia i livelli di confidenza per casi diversi possono essere determinati. Alcuni valori tipici di livelli di confidenza sono indicati in figura 18.6 (pag. 451). Usualmente, in ogni esperimento individuiamo un insieme di livelli di confidenza all'interno dei quali ci aspettiamo che cadano i valori misurati con una certa probabilita' (pag. 451); per fare questo dobbiamo decidere le probabilita' di errore che siamo disposti ad accettare. La percentuale di probabilita' di errore e' definita come uguale a 100-livello di confidenza.
Di solito, accettiamo errori del 5%, cioe' 95% e' il livello di confidenza per il quale z=1.96 (dalla tabella di pag. 449) considerando entrambi i lati del diagramma e z=1,645 considerando solo un lato. Comunque, in ogni luogo in cui la vita umana si e' sviluppata, si considera come bassa probabilita' di errore, quella dell'ordine del 1%; questo assicura un livello di confidenza del 99% per il quale z=2,326 per un solo lato e z=2,575 per entrambi i lati.
Nella pratica, possiamo incontrare due tipi di problemi: problema diretto nel quale i limiti di X sono assegnati ed e' richiesta la determinazione della probabilita' di occorrenza in quel dato campo; l'altro e' un problema di tipo inverso nel quale e' assegnata la probabilita' (il livello di confidenza assegnato e' usualmente il 95%) e sono richiesti i limiti di X.

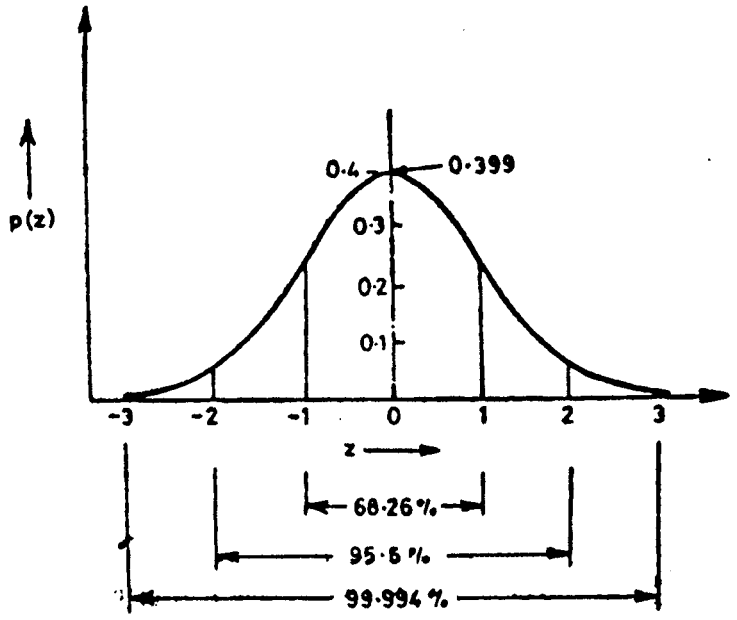


Fig. 18.6 Confidence levels for different ranges of measured values

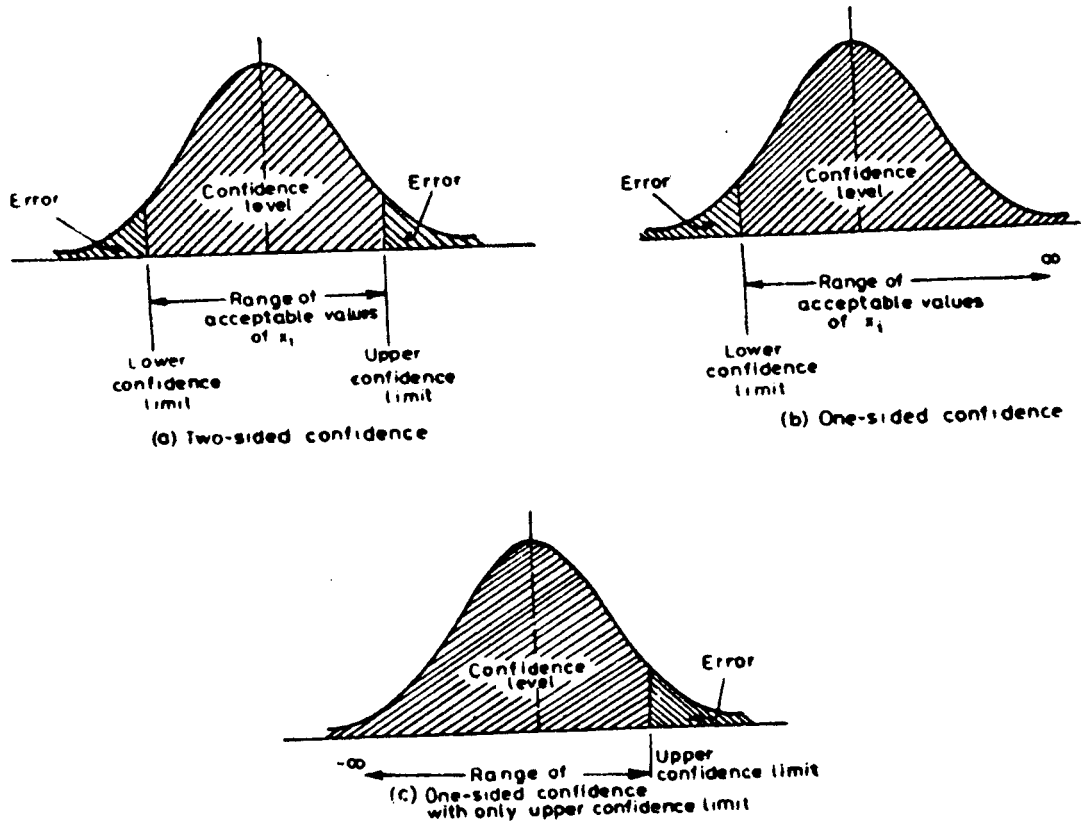


Fig. 18.7 Pictorial representation of Gaussian distribution with given confidence limits

TEOREMA DEL LIMITE CENTRALE

Il teorema del limite centrale e' molto importante nel campo statistico. Esso asserisce che il campione di un popolazione segue l'andamento gaussiano, qualunque sia la distribuzione delle misure individuali. La variabile normale standardizzata e' definita come

$$z = \frac{\bar{x} - \bar{\bar{x}}}{\sigma_m}$$

dove $\bar{x}$ e' il valore medio di ogni campione, $\bar{\bar{x}}$ e' la media della popolazione e σ_m e' deviazione media standard della media che e' circa uguale alla stima interna di incertezza (oppure errore standard interno) Vi, quando il numero di osservazioni e' abbastanza alto.
Puo' essere notato che la distribuzione normale delle misure di un campione ha minor precisione se comparata alla corrispondente distribuzione normale dal campione medio della popolazione. In altre parole, l'ultima distribuzione ha un picco piu' appuntito rispetto alla prima considerata. E cio' avviene a causa della relazione tra $\sigma_m = \frac{\sigma}{\sqrt{n}}$

Nella maggior parte dei casi siamo interessati all'analisi dell'errore standard interno dei dati e nell'impiego della variabile standardizzata Z invece che alla media della popolazione ed alla deviazione standard delle medie. Comunque, in alcuni problemi, siamo interessati alla deviazione standard del solo campione J; in casi del genere il valore di Z e' calcolato usando la media e la deviazione standard dal campione.

TEST DI SIGNIFICATIVITA'

Se la distribuzione normale di una particolare grandezza e' nota e tale grandezza e' misurata di nuovo sotto condizione diverse da quella precedente, il valor medio sarà probabilmente diverso dalla distribuzione originale. Se la differenza delle medie e' piccola, sarebbe ragionevole assumere che la distribuzione e' proveniente dalla stessa popolazione.
Invece, se la differenza e' considerevole, allora sarebbe ragionevole assumere che le circostanze diverse in cui e' stato effettuato il secondo rilevamento hanno alterato i valori ed il risultato e' da prendere in considerazione. In altre parole, possiamo dire che i dati originali ed i dati susseguenti non sono derivati dalla stessa popolazione.
Per testare se i due rilevamenti hanno riportato una differenza significativa, noi usiamo un test di significativita' basato sulla differenza delle medie. Il criterio statistico di tali medie e' quello secondo cui se la differenza dei valori medi dei due campioni devia di 1,96 volte l'errore standard interno della differenza delle medie, allora il cambio dei dati e' significativo.
Diciamo che un particolare campione e' individuato da n_1 valori ed ha $\bar{x}_1$ come valor medio e σ_1 come deviazione standard

..; analogamente, un altro campione rilevato sotto condizioni diverse ha n_2 valori, ed $\bar{X}_2$ come media e deviazione standard σ_2. Si ricorda che l'errore standard interno (o stima interna di incertezza) di un campione di n valori con deviazione standard σ. e' dato da

$$U = \frac{\sigma}{\sqrt{n-1}}$$

Tenendo questo a mente si procede per determinare se il cambio di dati e' significativo o meno:

1) facciamo la differenza tra i valori medi che da' il campo di variazione di essi, cioe' $R(\bar{x}) = \bar{X}_1 - \bar{X}_2$

2) determiniamo l'errore standard medio di $R(\bar{X})$ cioe':

$$\{(U^2)\}_{R(\bar{X})} = \left[\frac{\partial R(\bar{X})}{\partial(\bar{X}_1)}\right]^2 \cdot U_1^2 + \left[\frac{\partial R(\bar{x})}{\partial \bar{X}_2}\right]^2 \cdot U_2^2 = U_1^2 + U_2^2$$

Sostituendo i valori di U_1 ed U_2 in termini delle corrispondenti deviazioni standard dei campioni, abbiamo:

$$(U)_{R(\bar{x})} = \sqrt{U_1^2 + U_2^2} = \left[\left\{\frac{\sigma_1}{\sqrt{n_1 - 1}}\right\}^2 + \left\{\frac{\sigma_2}{\sqrt{n_2 - 2}}\right\}^2\right]^{1/2}$$

3) Noti i valori di $R(\bar{X})$ e $(U)_{R(\bar{x})}$ dalle equazioni precedenti possiamo adesso impiegare il test di significatività che e' il seguente:

$$R(\bar{X}) > \left|1.96 \cdot (U)_{R(\bar{x})}\right|$$

allora la differenza delle medie e' significativa, altrimenti possiamo assumere che i dati originari e quelli seguenti sono appartenenti alla stessa popolazione.
In altre parole possiamo dire che sopra il 95% di livello di confidenza, le medie dei due campioni non sono della stessa popolazione se il test di significatività e' positivo.
Al contrario se la differenza delle medie $R(\bar{X})$ e' minore di 1,96 volte l'errore standard medio allora il risultato non e' significativo e possiamo dire che la probabilita' di errore e' minore del 5% considerando entrambi i campioni appartenenti alla stessa popolazione. Invece se $R(\bar{X})$ si allontana rispetto a 2,58 volte l'errore combinato interno standard (U) $.R(\bar{X})$ allora il risultato e' altamente significativo ed in questo caso il livello di confidenza considerato e' del 99%.

TEST CHI-QUADRATO PER LA BONTA' E L'ADATTABILITA'

Quando e' stato fatto un certo insieme di misure, si ritiene che esse sono un campione di una distribuzione teorica nota. Per confrontare le differenti parti della distribuzione

osservata, noi suddividiamo i dati in n classi e determiniamo le frequenze osservate in ogni classe; quindi stimiamo le frequenze ottenute di ogni classe per assumere la distribuzione adottata alla particolare distribuzione teorica. Ad esempio, se la distribuzione assunta e' quella gaussiana allora la seguente procedura puo' essere adottata per calcolare i valori attesi di frequenza per un certo insieme di dati:

1) calcolare media e dati standard dei dati;

2) per ogni intervallo calcolare z_u e z_l cioe' il piu' basso ed il piu' alto valor limite rispettivamente;

3) dalla tabella dell'integrale gaussiano determinare la probabilita' tra 0 e z_u e tra 0 e z_l;

4) la differenza dei suddetti valori di probabilita' da' l'integrale gaussiano nel dato intervallo se entrambi i limiti cadono tra 0 ed ∞ oppure tra $-\infty$ e zero;
La somma di questi valori da' l'integrale gaussiano di probabilita' se il piu' alto limite cade tra 0 ed ∞ ed il piu' basso cade tra $-\infty$ e zero, e viceversa;

5) Moltiplicando l'integrale gaussiano di probabilita' in un dato intervallo per il numero totale di osservazione, si ottiene la frequenza attesa il quel dato intervallo;

6) la somma delle frequenze attese di tutte le classi non e' talvolta uguale al numero totale di osservazioni; cio' e' causato dall'arrotondamento nelle interpolazioni eseguite sulla tabella.

Quindi, le frequenze attese negli intervalli sono moltiplicate per un conveniente fattore di correzione in modo tale che la somma della frequenza eguaglia il numero di osservazioni.
Dopo la determinazione delle frequenze attese nelle varie classi, determiniamo il χ^2 (si pronuncia chi-quadr) parametro come segue: diciamo che vi sono n classi e le frequenze attese ed osservate sono rispettivamente:

$$f_{e1}, f_{e2}, \ldots f_{en} \quad ; \quad f_{o1}, f_{o2}, \ldots, f_{on}$$

Adesso, il nostro scopo sara' quello di determinare se le frequenze osservate e quelle attese sono tali per concludere se esse provengano o meno dalla stessa distribuzione di probabilita'.

→ vedi pag. 462-463

CRITERIO PER LA BONTA' E L'ADATTABILITA'

Il criterio statistico per la bonta' e l'adattabilita', cioe' come un insieme di dati osservatisi adatti ad una certa distribuzione teorica, sono i seguenti:

1) se il valore di probabilita' nel test-χ^2 cade tra 0.1 e 0.9 allora la distribuzione osservata e' considerata come proveniente dalla distribuzione teorica assunta. In alcuni casi, il piu' basso valore della probabilita' χ^2 (definito come livello di significativita' o semplicemente livello) puo' essere ridotta a 0,05.

2) Se il valore di probabilita' nel test-χ^2 e' al di sotto del piu' basso limite prescritto, allora, il risultato e' significativo ed il campione di dati e' considerato completamente differente dalla distribuzione assunta.

3) Se il valore del parametro χ^2 e' vicino allo zero o molto piu' piccolo, allora la probabilita' puo oltrepassare il limite superiore di 0.9. Simili casi sono affrontati nella pratica; e se e' cosi', allora normalmente consideriamo i dati "buoni" ma con una certa diffidenza.

18.9 CHI-SQUARE TEST FOR GOODNESS OF FIT

When a set of measurements is obtained, it is believed that the measurements are a sample of some known theoretical distribution; say normal frequency distribution which is generally hypothesised in cases involving experimental statistics. For comparing the different parts of the observed distribution, we subdivide the data into a number of classes say n and determine the observed frequency in each class. Then we estimate the expected frequency of each class by assuming that the distribution conforms to a particular theoretical distribution.

For example, if the assumed distribution considered is Gaussian, then the following procedure may be adopted to calculate the expected values of frequencies for a given set of data:

1. Calculate the mean value and staddard deviation of the data.
2. For each class interval, calculate the standard normal variates z_u and z_l for the upper and lower boundary values respectively.
3. From the integral Gaussian table, determine the integral Gaussian probability between 0 and z_u and 0 and z_l.
4. The difference in the above values gives the integral Gaussian probability in the given interval if both the upper and lower boundaries lie either between 0 and ∞ or 0 and $-\infty$. The sum of these values gives the integral Gaussian probability if the upper boundary lies between 0 and ∞ and the lower boundary lies between 0 and $-\infty$ and vice versa.
5. Multiplying the integral Gaussian probability in a given class interval by the total number of observations gives the expected frequency of occurrence of the variable in that interval.
6. The summation of expected frequencies in all classes sometimes does not equal the total number of observations. The slight difference is caused by small rounding-off errors due to interpolations in the integral Gaussian table. Therefore, the expected frequencies in step (5) are multiplied by a suitable correction factor so as to make the sum of expected frequencies equal to the number of observations.

After determining the expected frequencies in the various classes, we determine the χ^2 (pronounced chi-square) parameter as follows:

Let us say that there are n classes ($n > 1$) and the expected and observed frequencies in the various classes are denoted by

$$f_{e1}, f_{e2}, f_{e3}, \cdots f_{en} \quad \text{and} \quad f_{o1}, f_{o2}, f_{o3}, \cdots f_{on}.$$

Now, our aim is to determine whether the observed frequencies and the expected frequencies are close enough for us to conclude that they come from the same probability distribution. To do so, we define the χ^2 parameter as:

$$\chi^2_{(n-m)df} = \sum_{i=1}^{n} \left\{ \frac{(f_{oi} - f_{ei})^2}{f_{ei}} \right\} \tag{18.29}$$

where n is the number of values that are summed up to produce the values of χ^2
m is the number of constants used in the calculation of expected frequencies
$n - m$ is the degrees of freedom and
subscript df stands for the degrees of freedom

The values of the numerator in the χ^2 (chi square) expression represent the squares of deviations between the expected and observed frequencies in various classes which is always positive. These values are normalised in each class by dividing them by the respective expected frequency of each class. It may be noted that the same order of deviation in the expected and observed frequencies causes relatively larger contribution in the χ^2-parameter at the tail portions of the normally distributed data, as compared to the values close to the mean value of the data. This is because of relatively large values of the expected frequencies near the mean value of the data which are in the denominator of the χ^2-parameter. In order to restrict the unusually large contributions in χ^2 parameter when the expected frequencies are small, the empirical criterion commonly used in practice is to regroup the various classes in such a way so that expected frequency in each class is not less than 5.0.

Further, a correction is sometimes applied to the chi-square values when the degree of freedom F, i.e. $(n-m)$ is of the order of 1.0. This is termed *Yate's* correction and accounts for the inaccuracies involved when the results of continuous distributions are applied to discrete data. The correction consists of writing Eq. (18.29) in the following form:

$$\chi^2_{F \approx 1.0} = \sum_{i=1}^{n} \left[\frac{\{ | (f_{o_i} - f_{e_i}) | - 0.5 \}^2}{f_{e_i}} \right] \tag{18.30}$$

If the sample distribution agrees with the assumed theoretical distribution then $\chi^2 = 0$. This is of course very unlikely because even if the sample is taken from the parent distribution, one would not expect *exact* agreement in every interval. But, the larger the value of χ^2, the more the disagreement between the assumed distribution and the observed values. In other words, in such a case, the smaller is the probability that the observed distribution matches the expected distribution. Thus, the chi-square parameter is quite useful in statistical analysis of data as it helps to test a particular hypothesis in the given data.

In applying the chi-square test, we first determine the value of χ^2 for the given data. Then, we determine the values of degrees of freedom F which is equal to $(n-m)$. Knowing the values of χ^2 and F, we determine the probability that the actual measurements match the expected distribution from either the chi-square tables (Table 18.3) or from the χ^2-F diagram (Fig. 18.14) which gives cross-plots of chi-square probability, $P(\chi^2)$, for various values of χ^2 and F.

18.10 CRITERIA FOR GOODNESS OF FIT

The statistical criteria for the goodness of fit, i.e. how well a set of observed data fit the assumed theoretical distributions are as follows:

1. If the value of probability in the χ^2-test lies between 0.1 and 0.9, then the observed distribution is considered to follow the assumed distribution. In other words, there is no reason to suspect the hypothesis. In certain cases, the lower limit of chi-square probability (also termed *significance level* or simply *level*) may be reduced to 0.05.
2. If the value of the probability in the χ^2-test is below the lower prescribed limit, then the result is significant and the sample data is considered to be entirely different from the assumed distribution. In such cases, the value of the χ^2-parameter is usually quite large.