King Abdulaziz University

Mechanical Engineering

MEP365

Thermal Measurements

Ch. 4 Probability and statistics

Feb. 2017

Ch. 4 Probability and statistics

Introduction

Concept of central value and probability

Probability density

Frequency distribution

Normal distribution

Infinite statistics

Finite statistics

Regression Analysis

Data Outlier Detection

Number of measurements data required

Introduction

Probability and statistics are used extensively in reducing and presenting measured data

Consider a person measuring the temperature in a room. How can the data be represented?

Consider a factory that manufacture a ball bearing. How can one represent the diameter of a sample of these bearings?

Variation in measured value is due:

- ❖Measurement system (Resolution and repeatability)
- ❖Measurement procedure and technique
- ❖Measured variable (Temporal variation, spatial variation)

We would like to represent the variation in measured variable x statistically by

$$x' = \bar{x} \pm u_x \quad (\mathrm{P\%})$$

Where

x' True value

$\bar{x}$ Mean value

u_x is the of uncertainty interval

P% = probability

Example of sample data

Table 4.1 Sample of Random Variable x

i	x_i	i	x_i
1	0.98	11	1.02
2	1.07	12	1.26
3	0.86	13	1.08
4	1.16	14	1.02
5	0.96	15	0.94
6	0.68	16	1.11
7	1.34	17	0.99
8	1.04	18	0.78
9	1.21	19	1.06
10	0.86	20	0.96

N = no of data points=20

How to represent this data by $x' = \bar{x} \pm u_x \ (\mathrm{P}\%)$

Concept of central value and probability

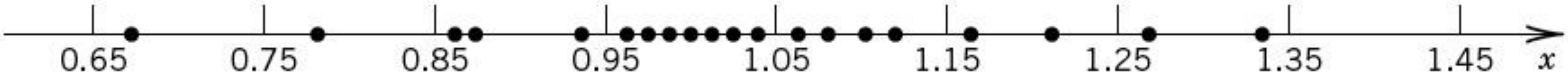


Figure 4.1 Concept of density in reference to a measured variable (from Example 4.1).

Frequency distribution

j	Interval	n_j	$f_j=n_j/N$
1	$0.65 \le x_j < 0.75$	1	0.05
2	$0.75 \le x_j < 0.85$	1	0.05
3	$0.85 \le x_j < 0.95$	3	0.15
4	$0.95 \le x_j < 1.05$	7	0.35
5	$1.05 \le x_j < 1.15$	4	0.20
6	$1.15 \le x_j < 1.25$	2	0.10
7	$1.25 \le x_j \le 1.35$	2	0.10

Table 4.1 Sample of Random Variable x

i	x_i	i	x_i
1	0.98	11	1.02
2	1.07	12	1.26
3	0.86	13	1.08
4	1.16	14	1.02
5	0.96	15	0.94
6	0.68	16	1.11
7	1.34	17	0.99
8	1.04	18	0.78
9	1.21	19	1.06
10	0.86	20	0.96

How to draw a histogram for the data

Divide the range into several intervals (K)

$$K = 1.87(N-1)^{0.4} + 1$$

For large values of N $K = \sqrt{N}$

N is number of data points

Provided that $n_j \geq 5$ **For at least one interval**

Histogram

Central tendency value at maximum frequency

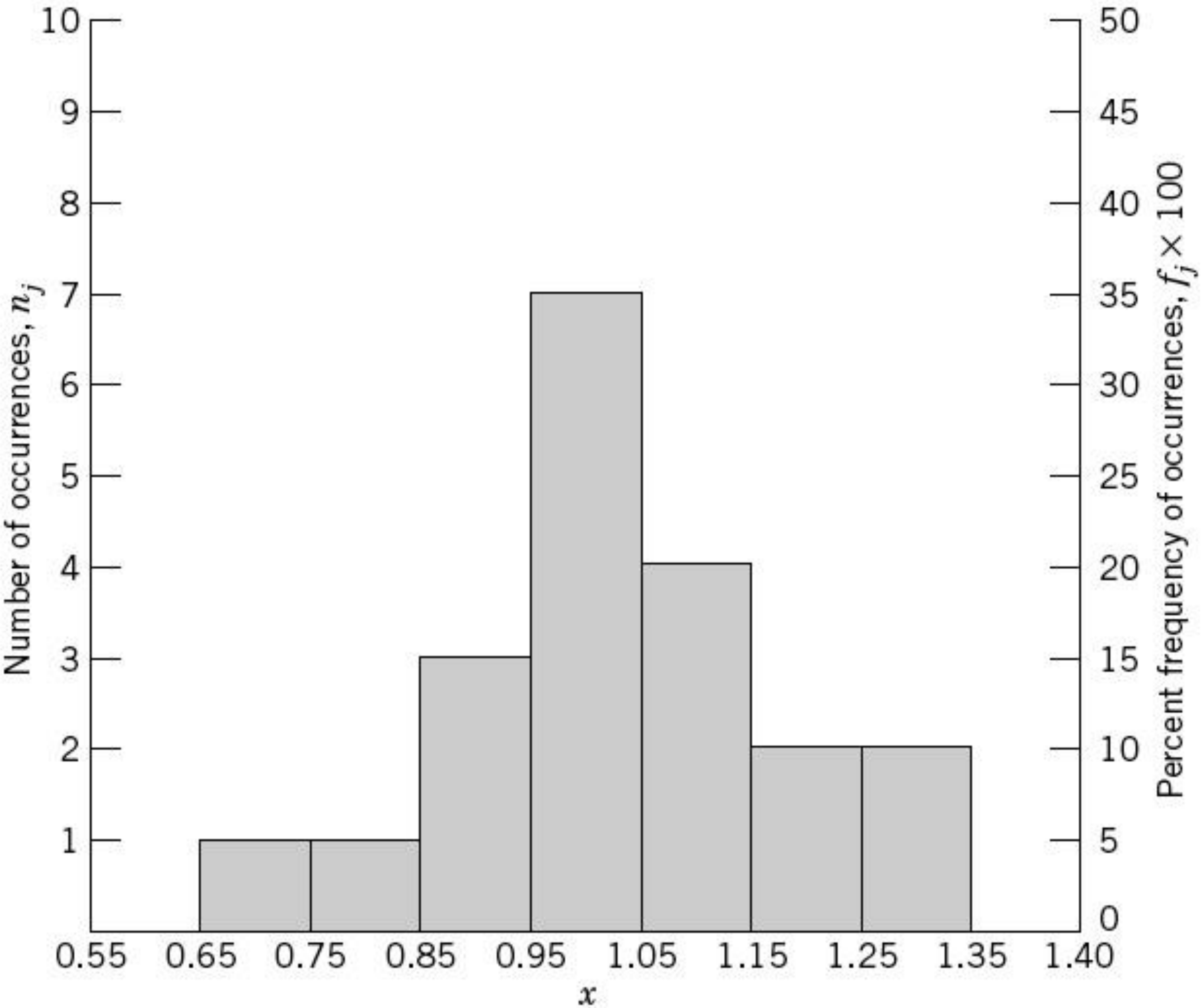


Figure 4.2 Histogram and frequency distribution for data in Table 4.1.

j	Interval	n_j	$f_j=n_j/N$
1	$0.65\le x_j \le 0.78$	1	0.05
2	$0.75\le x_j < 0.85$	1	0.05
3	$0.85\le x_j < 0.95$	3	0.15
4	$0.95 \le x_j < 1.05$	7	0.35
5	$1.05 \le x_j < 1.15$	4	0.20
6	$1.15 \le x_j < 1.25$	2	0.10
7	$1.25 \le x_j \le 1.35$	2	0.10

Frequency distribution

Histogram

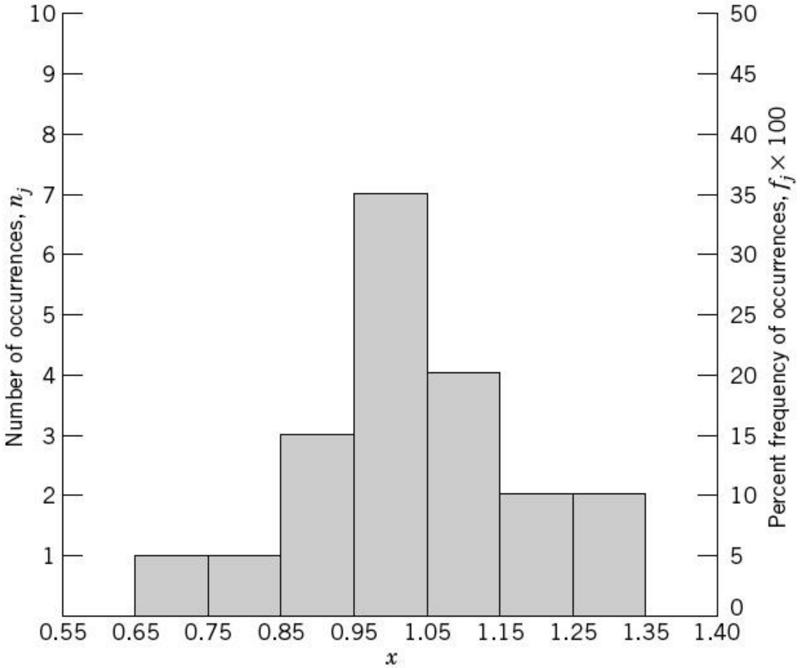
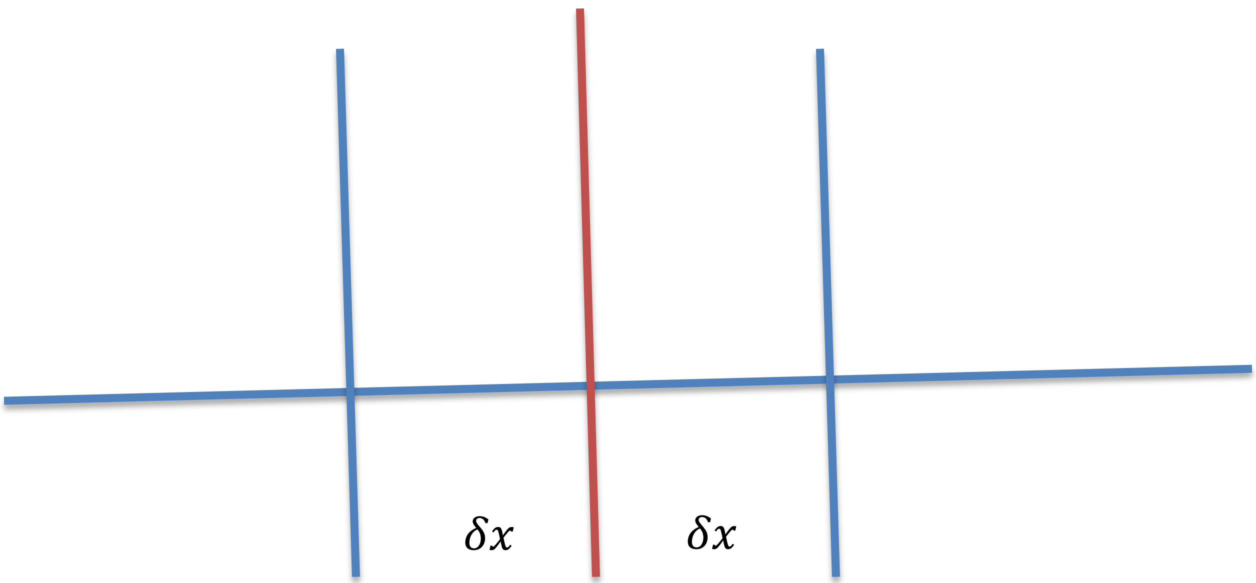


Figure 4.2 Histogram and frequency distribution for data in Table 4.1.

Probability density

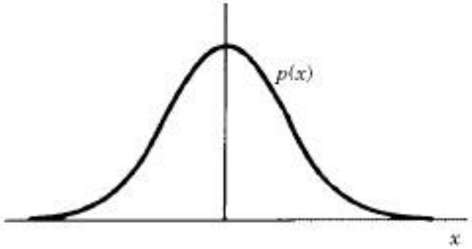
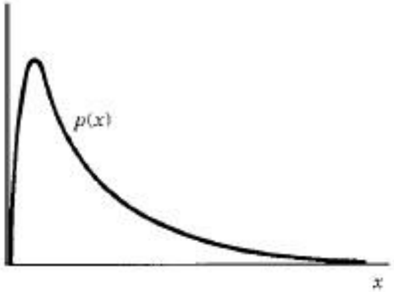
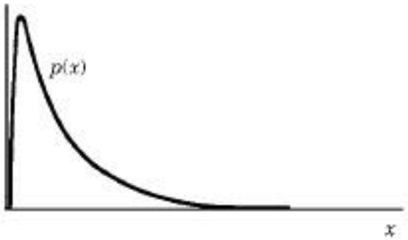
$$p(x) = \lim_{N\to\infty,\delta x\to 0} \frac{n_j}{N(2\delta x)}$$



Probability value changes from zero to maximum 1

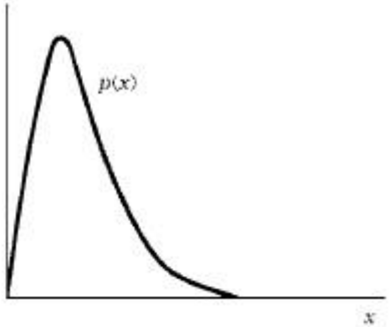
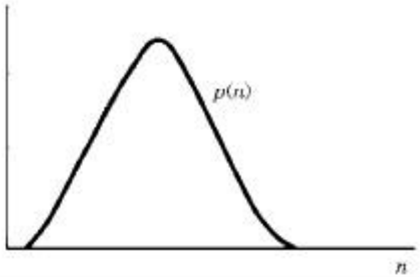
Samples of probability distributions

Table 4.2 Standard Statistical Distributions and Relations to Measurements

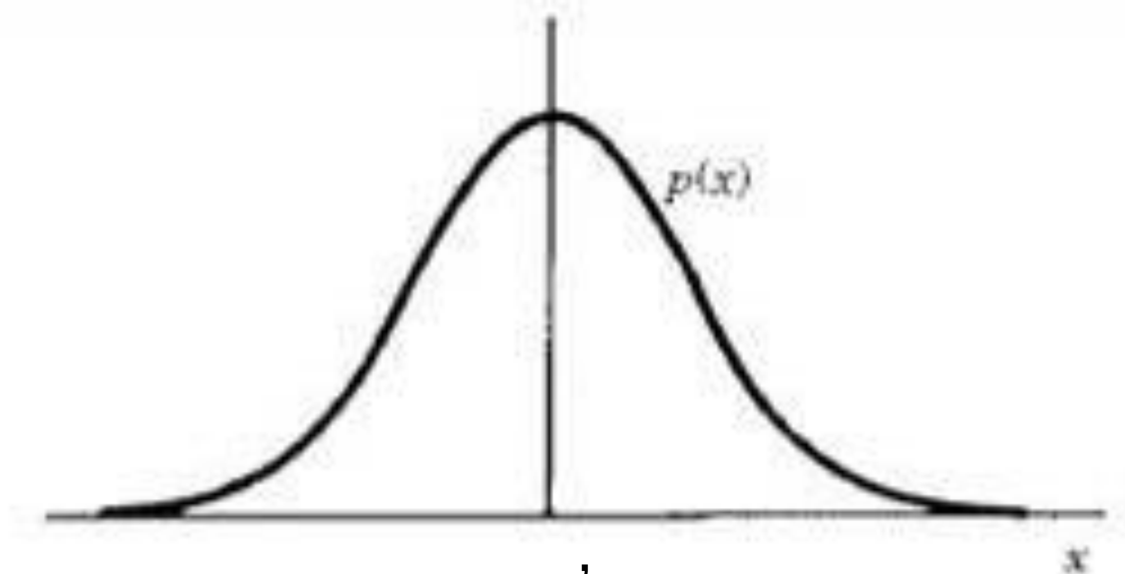
<table>
<tr><th>Distribution</th><th>Applications</th><th>Mathematical Representation</th><th>Shape</th></tr>
<tr><td>Normal</td><td>Most physical properties that are continuous or regular in time or space. Variations due to random error.</td><td>$p(x) = \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\frac{1}{2}\frac{(x-x')^2}{\sigma^2}\right]$</td><td>
</td></tr>
<tr><td>Log normal</td><td>Failure or durability projections; events whose outcomes tend to be skewed toward the extremity of the distribution.</td><td>$p(x) = \frac{1}{\pi\sigma(2\pi)^{1/2}} \exp\left[-\frac{1}{2}\ln\frac{(x-x')^2}{\sigma^2}\right]$</td><td>
</td></tr>
<tr><td>Poisson</td><td>Events randomly occurring in time; $p(x)$ refers to probability of observing x events in time t. Here λ refers to x'.</td><td>$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$</td><td>
</td></tr>
</table>

Samples of probability distributions [Continued]

Table 4.2 Standard Statistical Distributions and Relations to Measurements

Distribution	Applications	Mathematical Representation	Shape
Weibull	Fatigue tests; similar to log normal applications.	See [4]	
Binomial	Situations describing the number of occurrences, n, of a particular outcome during N independent tests where the probability of any outcome, P, is the same.	$p(n) = \left[\frac{N!}{(N-n)!n!}\right] P^n (1-P)^{N-n}$	

Normal Gaussian distribution

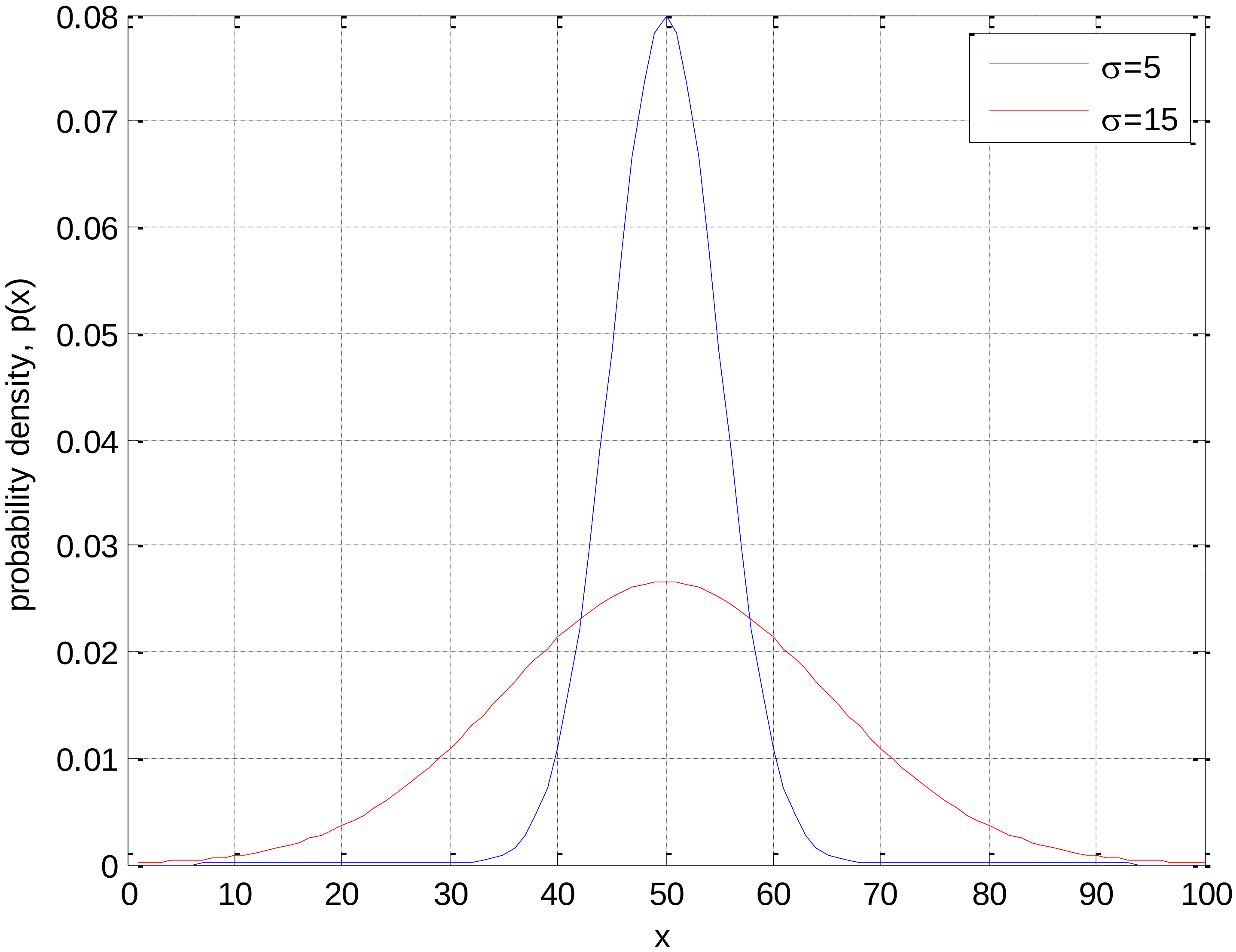


x'

$$p(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\frac{(x-x')^2}{\sigma^2}\right]$$

X' is the true mean, σ is the standard of deviation

Gaussian Probability Function Distribution



Continues data

True mean value

$$x' = \int_{-\infty}^{+\infty} x p(x) dx$$

True variance

$$\sigma^2 = \int_{-\infty}^{+\infty} (x - x')^2 p(x) dx$$

Discrete data

True mean value

$$x' = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} x_i$$

True variance

$$\sigma^2 = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} (x - x')^2$$

Standard of deviation is σ

$$\sigma = \sqrt{(Variance)}$$

Normal Gaussian distribution function

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(x - x')^2}{\sigma^2} \right]$$

Infinite statistics (N→∞)

Define:

$$\beta = \frac{(x - x')}{\sigma} \qquad z_1 = \frac{(x_1 - x')}{\sigma}$$

The probability of x to have a value between

$$x' - \delta x \le x \le x' + \delta x$$

$$P(x' - \delta x \le x \le x' + \delta x) = \int_{x' - \delta x}^{x' + \delta x} p(x)\,dx$$

$$\beta = \frac{(x - x')}{\sigma} \qquad z_1 = \frac{(x_1 - x')}{\sigma}$$

$$P(-z_1 \le \beta \le z_1) = \frac{1}{\sqrt{2\pi}} \int_{-z_1}^{z_1} e^{-\beta^2/2}\,d\beta = 2\left[\frac{1}{\sqrt{2\pi}} \int_{0}^{z_1} e^{-\beta^2/2}\,d\beta\right]$$

Table 4.3 → **Error function**

Probability for z to be between 0 and any value z_1

Area =(1/2) P $(-z_1 \leq \beta \leq z_1)$

OR

It can be directly found from table 4.3 in your text book

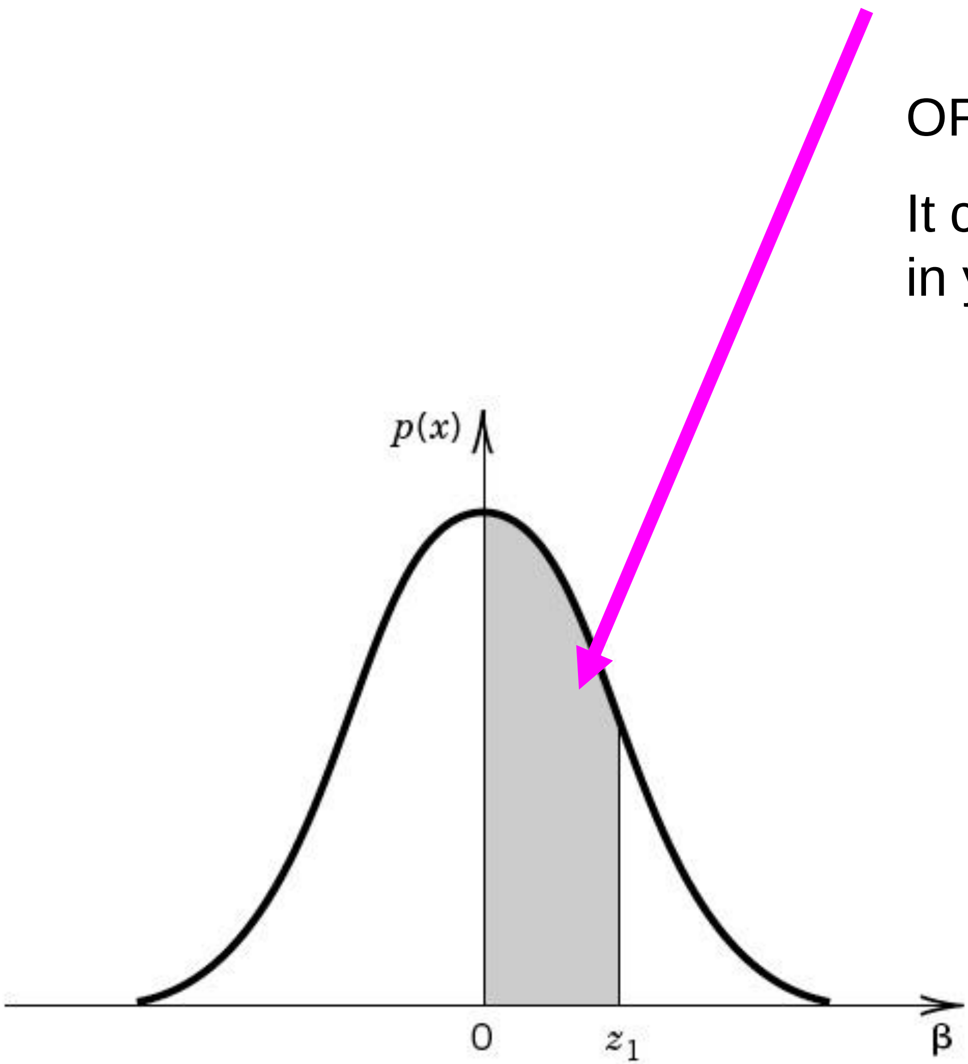


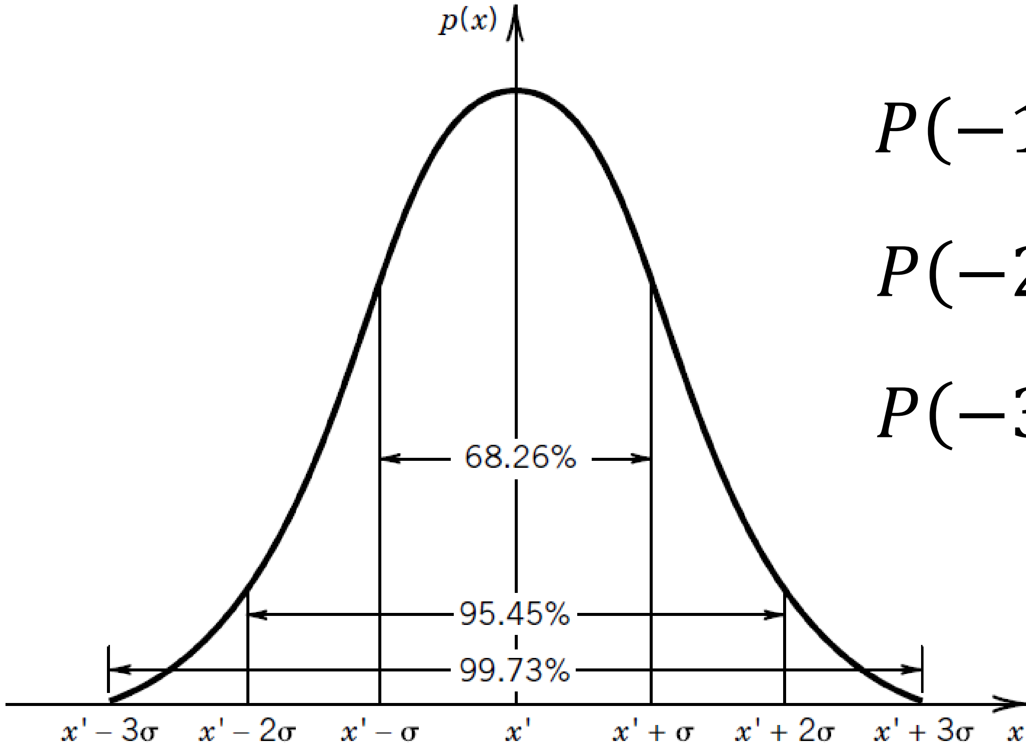
Figure 4.3 Integration terminology for the normal error function.

Table 4.3 Probability Values for Normal Error Function

One-Sided Integral Solutions for $p(z_1) = \frac{1}{(2\pi)^{1/2}} \int_0^{z_1} e^{-\beta^2/2} d\beta$

$z_1 = \frac{x_1 - x'}{\sigma}$	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	0.1809	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	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	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	0.2995	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	0.4758	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4799	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	0.49865	0.4987	0.4987	0.4988	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(x - x')^2}{\sigma^2} \right]$$



$$P(-1 \leq z \leq 1) = 0.6826$$

$$P(-2 \leq z \leq 2) = 0.9545$$

$$P(-3 \leq z \leq 3) = 0.9973$$

$z_1 = 1$, 68.26% of the area under $p(x)$ lies within $\pm z_1\sigma$ of x'.
$z_1 = 2$, 95.45% of the area under $p(x)$ lies within $\pm z_1\sigma$ of x'.
$z_1 = 3$, 99.73% of the area under $p(x)$ lies within $\pm z_1\sigma$ of x'.

Normal-Gaussian Distribution (cont.)

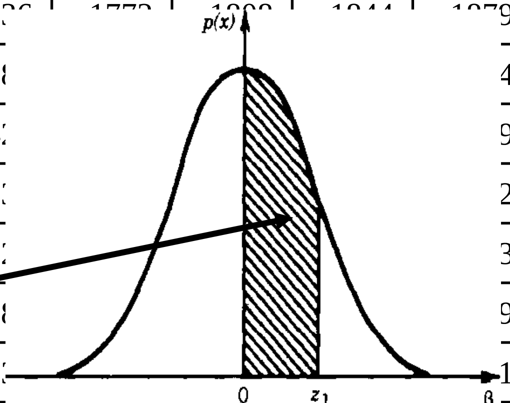
Table 4.3 Probability values for normal error function, one-sided integral solutions for

$$p(z_1) = \left[\frac{1}{(2\pi)^{1/2}} \int_0^{z_1} e^{-\beta^2/2} d\beta \right]$$

$P(0 \le z_1 \le 1.02) = ?$

$z_j = \frac{(x - x')}{\sigma}$	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
0.5	.1915	.1950	.1985	.2019	.2054	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
0.6	.2257	.2291	.2324	.2357	.2389	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
0.7	.2580	.2611	.2642	.2673	.2704	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
0.8	.2881	.2910	.2939	.2967	.2995	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
0.9	.3159	.3186	.3212	.3238	.3264	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1.0	.3413	.3438	.3461	.3485	.3508	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1.1	.3643	.3665	.3686	[illegible]	[illegible]	.3749	.3770	.3790	.3810	.3830

$Z_1 = 1.02$



$P(z_1 = 1.02) = 34.61\%$

Also, $Z_1(P = 0.3461) = 1.02$

Example on using Gaussian normal distribution

Assume a normal distribution. Using table 4.3 find the probability that the value of x be In the range x'±σ

since

$$z_1 = \frac{(x_1 - x')}{\sigma}$$

$$z_1 = \frac{(x' + \sigma - x')}{\sigma} = 1$$

from table 4.3 with z=1, the half side probability is 0.3413. Therefore for the full sided probability is

P=2*0.3413=0.6826 or 68.26 %

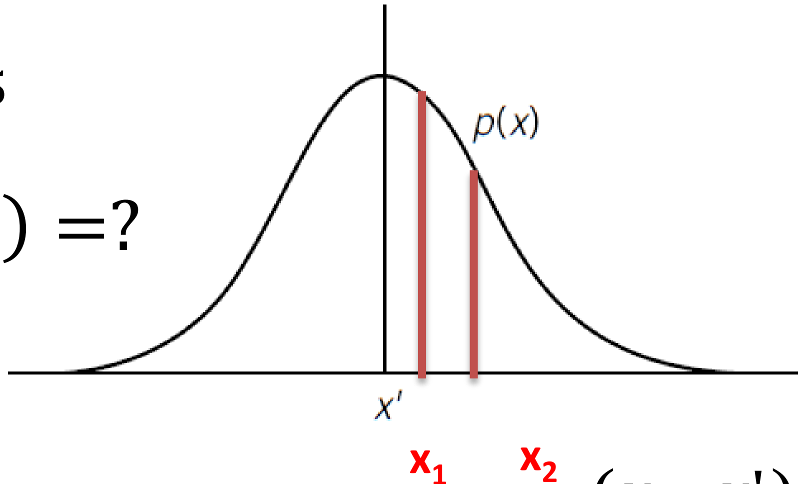
Example 4.3

The statistics of a well-defined varying voltage signal are given by $x' = 8.5$ V and $\sigma^2 = 2.25$ V^2. If a single measurement of the voltage signal is made, determine the probability that the measured value indicated will be between 10.0 and 11.5 V.

KNOWN $x' = 8.5$ V
$\sigma^2 = 2.25$ V^2

$$\sigma = \sqrt{2.25} = 1.5$$

$X_1=10.0$
$X_2=11.5$



$$P(10.0 \le x \le 11.5) = ?$$

$$z = \frac{(x - x')}{\sigma}$$

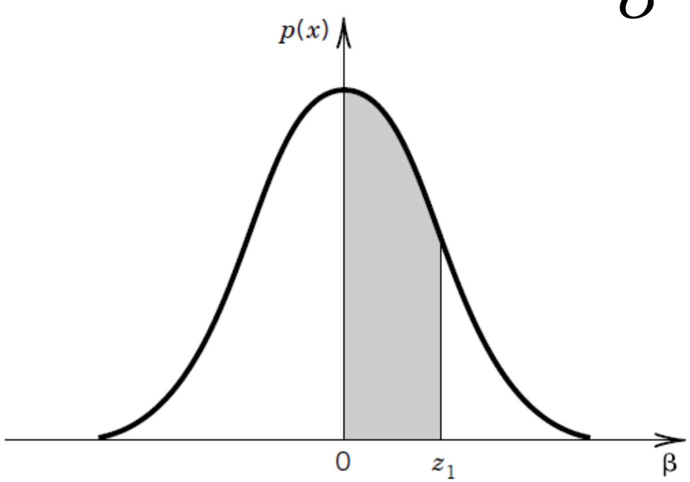
$$z_1 = \frac{10.0 - 8.5}{1.5} = 1 \qquad z_2 = \frac{11.5 - 8.5}{1.5} = 2$$

$$\beta = \frac{(x - x')}{\sigma}$$

$$P(1 \le z \le 2) = ?$$

Use Table 4.3 to find

$$P(0 \le \beta \le z_1) = ?$$



Example 4.3 continue

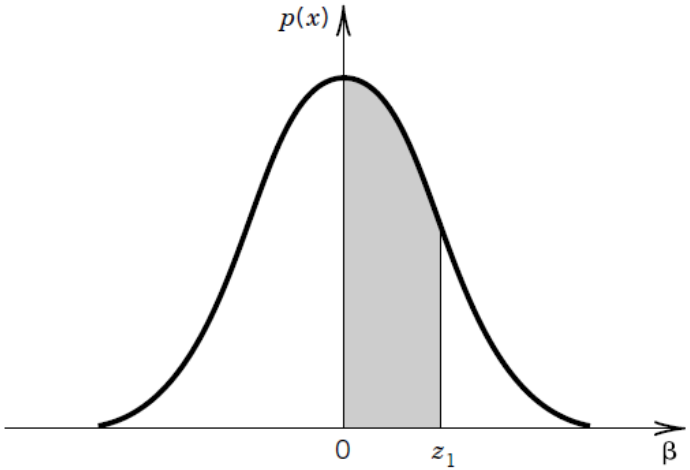


Table 4.3 Probability Values for Normal Error Function

One-Sided Integral Solutions for $p(z_1) = \frac{1}{(2\pi)^{1/2}} \int_0^{z_1} e^{-\beta^2/2} d\beta$

$z_1 = \frac{x_1 - x'}{\sigma}$	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	0.1809	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	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	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	0.2995	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	0.4758	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4799	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	0.49865	0.4987	0.4987	0.4988	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990

$P(0 \leq z_1 \leq 1)$ =0.3413

$P(0 \leq z_2 \leq 2)$ =0.4772

$P(1 \leq z \leq 2)$ =0.4772-0.3413=0.1359

The probability that x is between 10 and 11.5 is 13.59 %

Finite statistics

Sample mean

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Sample variance

$$s_x{}^2 = \left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]$$

Sample standard of deviation

$$s_x = \left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{1/2}$$

ν=N-1 = degree of freedom

Finite statistics (t-distrbuition)

Range of values of x

$$x_i = \bar{x} \pm t_{v,P} s_x \qquad (P\%)$$

$\pm t_{v,P} s_x$ Uncertainty interval

$$s_x = \left[\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 \right]^{1/2}$$

ν is the degree of freedom= N-1

$t_{v,P}$ is t estimator (student distribution) from table 4.4 as a function of ν and P(%)

As $N \rightarrow \infty$, $t_{v,p} = z_1$, $s_x = \sigma$

50, 90, 95 and 99 are the probabilities

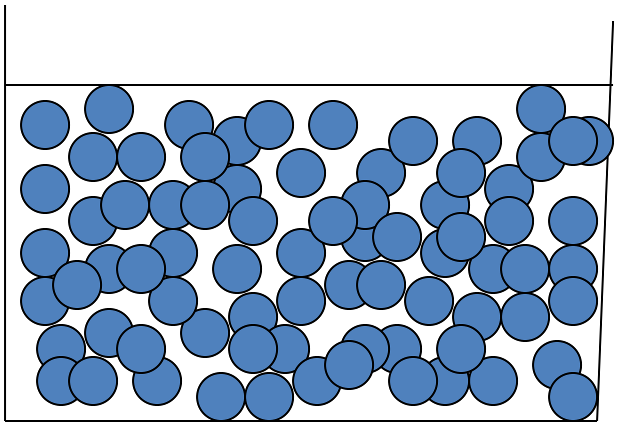
Evaluating $t_{\nu,P}$

ν is the degree of freedom= N-1

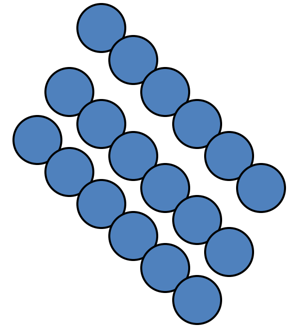
Table 4.4 Student-*t* Distribution

ν	t_{50}	t_{90}	t_{95}	t_{99}
1	1.000	6.314	12.706	63.657
2	0.816	2.920	4.303	9.925
3	0.765	2.353	3.182	5.841
4	0.741	2.132	2.770	4.604
5	0.727	2.015	2.571	4.032
6	0.718	1.943	2.447	3.707
7	0.711	1.895	2.365	3.499
8	0.706	1.860	2.306	3.355
9	0.703	1.833	2.262	3.250
10	0.700	1.812	2.228	3.169
11	0.697	1.796	2.201	3.106
12	0.695	1.782	2.179	3.055
13	0.694	1.771	2.160	3.012
14	0.692	1.761	2.145	2.977
15	0.691	1.753	2.131	2.947
16	0.690	1.746	2.120	2.921
17	0.689	1.740	2.110	2.898
18	0.688	1.734	2.101	2.878
19	0.688	1.729	2.093	2.861
20	0.687	1.725	2.086	2.845
21	0.686	1.721	2.080	2.831
30	0.683	1.697	2.042	2.750
40	0.681	1.684	2.021	2.704
50	0.680	1.679	2.010	2.679
60	0.679	1.671	2.000	2.660
∞	0.674	1.645	1.960	2.576

Standard deviation of the means

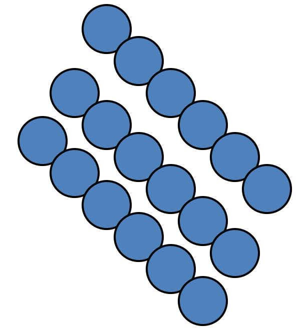


Population



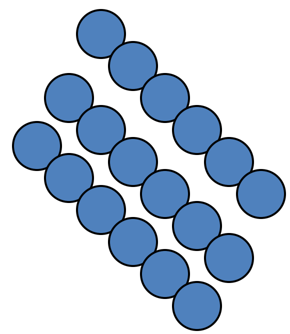
Sample 1

$N_{1,}\ S_1$



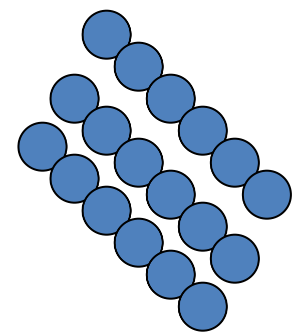
Sample 2

$N_{2,}\ S_2$



Sample 3

$N_{3,}\ S_3$



Sample 4

$N_{m,}\ S_m$

Standard deviation of the mean

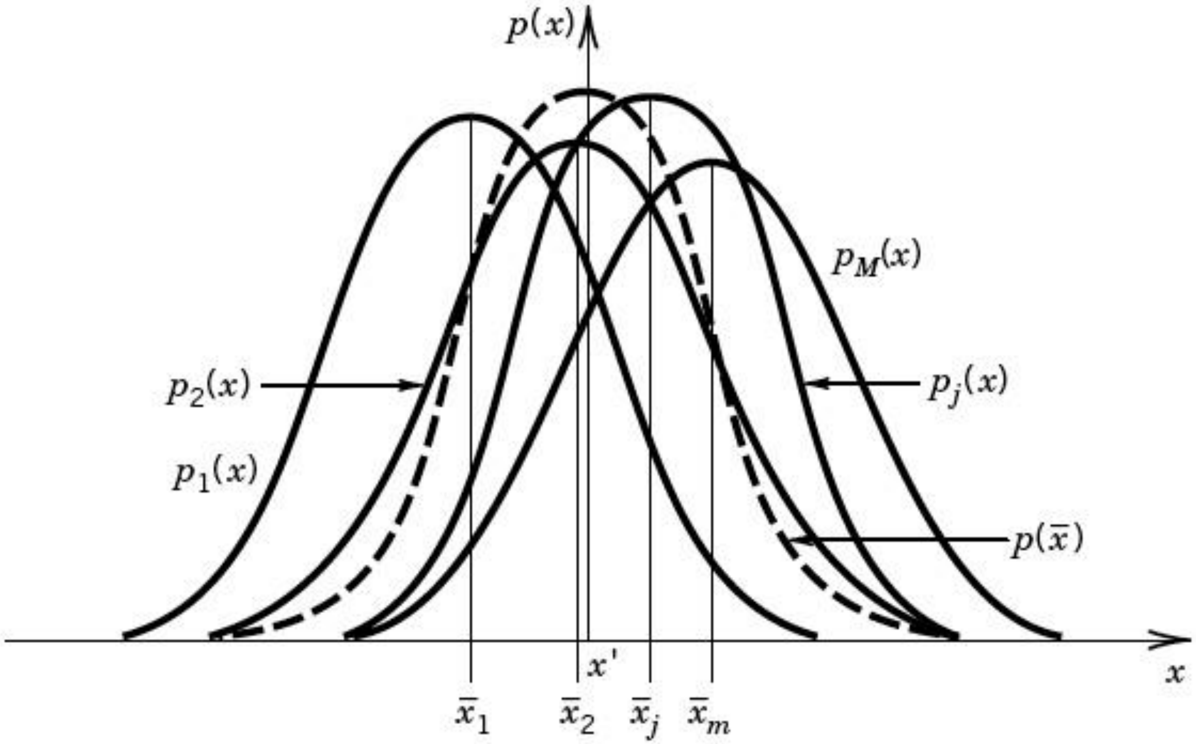


Figure 4.5 The normal distribution tendency of the sample means about a true value in the absence of systematic error.

For several measurements, the means will have a normal distribution

Standard deviation of the mean

What is the mean if M replications were done?
Each time with number of measurements =N

By definition

Standard deviation of the mean

$$S_{\bar{x}} = \frac{S_x}{\sqrt{N}}$$

True mean

$$x' = \bar{x} \pm t_{v,P} S_{\bar{x}}$$

$$t_{v,p} S_{\bar{x}}$$

Represents the confidence interval of the mean value around the mean

Distribution of x and distribution of the mean of x

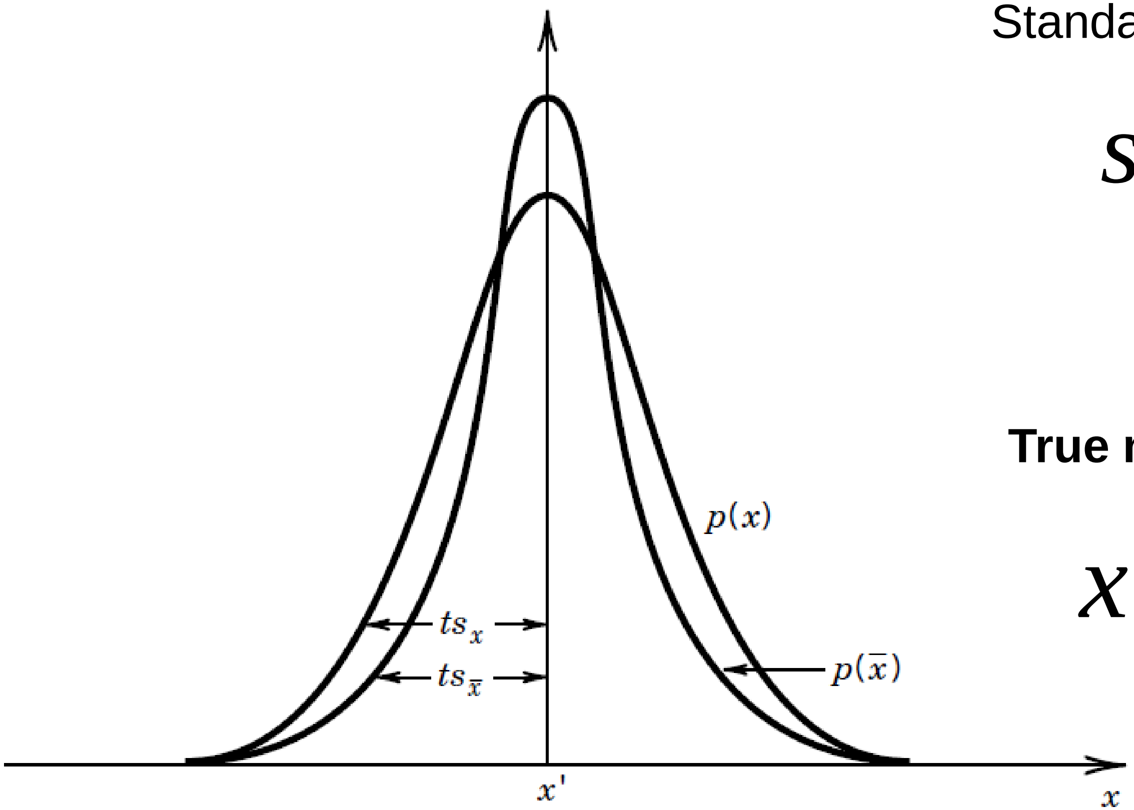


Figure 4.6 Relationships between s_x and the distribution of x and between $s_{\bar{x}}$ and the true value x'.

Standard deviation of the mean

$$s_{\bar{x}} = \frac{s_x}{\sqrt{N}}$$

True mean

$$x' = \bar{x} \pm t_{\nu,P} s_{\bar{x}}$$

Example 4.4

Find

a) Compute the sample statistics (sample mean and standard deviation s_x)

b) Estimate the interval of values for 95 % probability

c) Estimate the true mean

Table 4.1 Sample of Random Variable x

i	x_i	i	x_i
1	0.98	11	1.02
2	1.07	12	1.26
3	0.86	13	1.08
4	1.16	14	1.02
5	0.96	15	0.94
6	0.68	16	1.11
7	1.34	17	0.99
8	1.04	18	0.78
9	1.21	19	1.06
10	0.86	20	0.96

Part a: Sample statistics

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{20}\sum_{i=1}^{20} x_i = 1.02$$

$$s_x = \left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{1/2} = 0.16$$

Part b: Interval of values if P=95%

$$x_i = \bar{x} \pm t_{v,P} s_x \qquad (P\%)$$

Degree of freedom v=20-1=19

From table 4.4 with v=19, P=95%, $t_{19,95}$=2.095

Range of values of x_i within 95 % probability

$$x_i = 1.02 \pm (2.093 * .16) = 1.02 \pm 0.33 \qquad (95\%)$$

If one to pick one more ball the diameter will be between 0.69 and 1.35 with 95% probability

c) true mean

Standard of deviation for the mean

$$s_{\bar{x}} = \frac{s_x}{\sqrt{N}} = \frac{0.16}{\sqrt{20}} = 0.04$$

The range of the true mean with confidence 95 % is

$$x' = \bar{x} \pm t_{v,P} s_{\bar{x}} = 1.02 \pm 2.093 * 0.04 = 1.02 \pm 0.08$$

$P = 95\%$

Table 4.4 Student-t Distribution

ν	t_{50}	t_{90}	t_{95}	t_{99}
1	1.000	6.314	12.706	63.657
2	0.816	2.920	4.303	9.925
3	0.765	2.353	3.182	5.841
4	0.741	2.132	2.770	4.604
5	0.727	2.015	2.571	4.032
6	0.718	1.943	2.447	3.707
7	0.711	1.895	2.365	3.499
8	0.706	1.860	2.306	3.355
9	0.703	1.833	2.262	3.250
10	0.700	1.812	2.228	3.169
11	0.697	1.796	2.201	3.106
12	0.695	1.782	2.179	3.055
13	0.694	1.771	2.160	3.012
14	0.692	1.761	2.145	2.977
15	0.691	1.753	2.131	2.947
16	0.690	1.746	2.120	2.921
17	0.689	1.740	2.110	2.898
18	0.688	1.734	2.101	2.878
19	0.688	1.729	2.093	2.861
20	0.687	1.725	2.086	2.845
21	0.686	1.721	2.080	2.831
30	0.683	1.697	2.042	2.750
40	0.681	1.684	2.021	2.704
50	0.680	1.679	2.010	2.679
60	0.679	1.671	2.000	2.660
∞	0.674	1.645	1.960	2.576

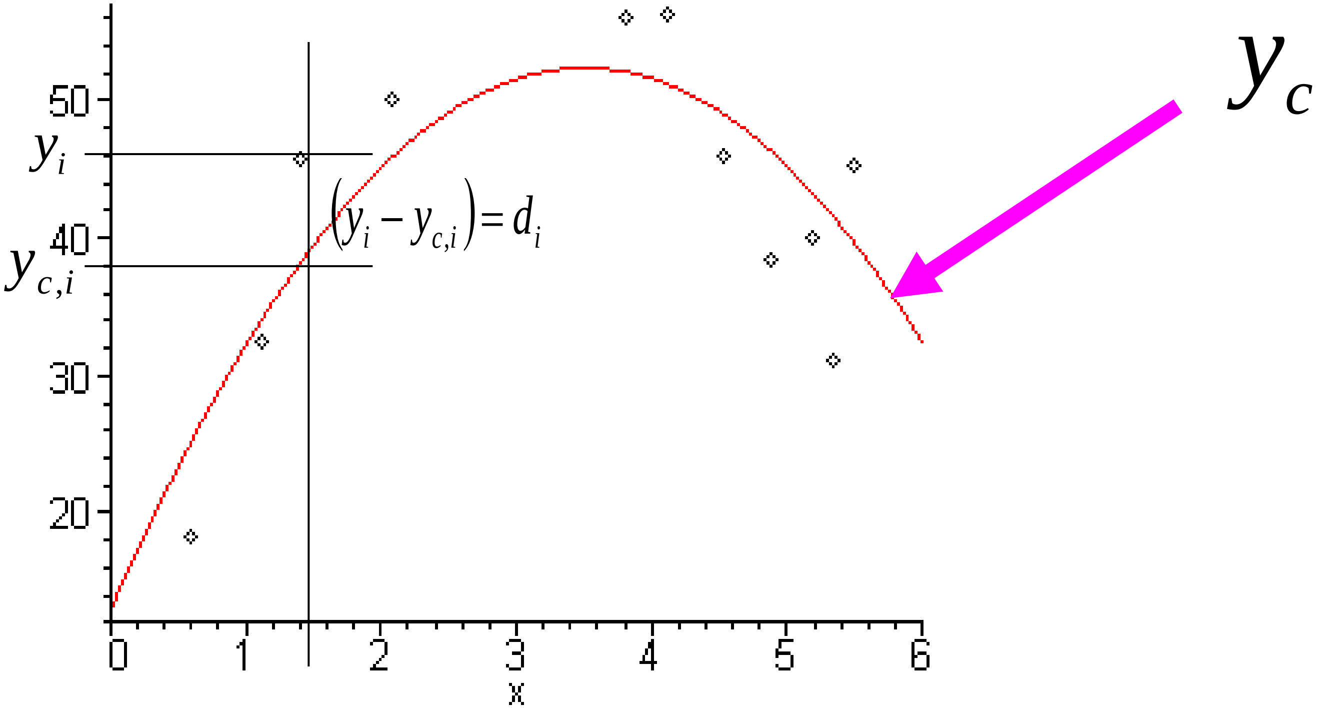
$v = 19$

$$\therefore\ t_{v,P} = t_{19,95} = 2.093$$

Pool statistics and
Sec. 4.5 CHI-squared distribution is omitted

Regression Analysis

Given data points: (x_i, y_i)



Regression Analysis

A procedure to get a relation between dependent and independent variables

For each value of x, there are n values of y (scattered)

Total number of data points is N

Regression Analysis

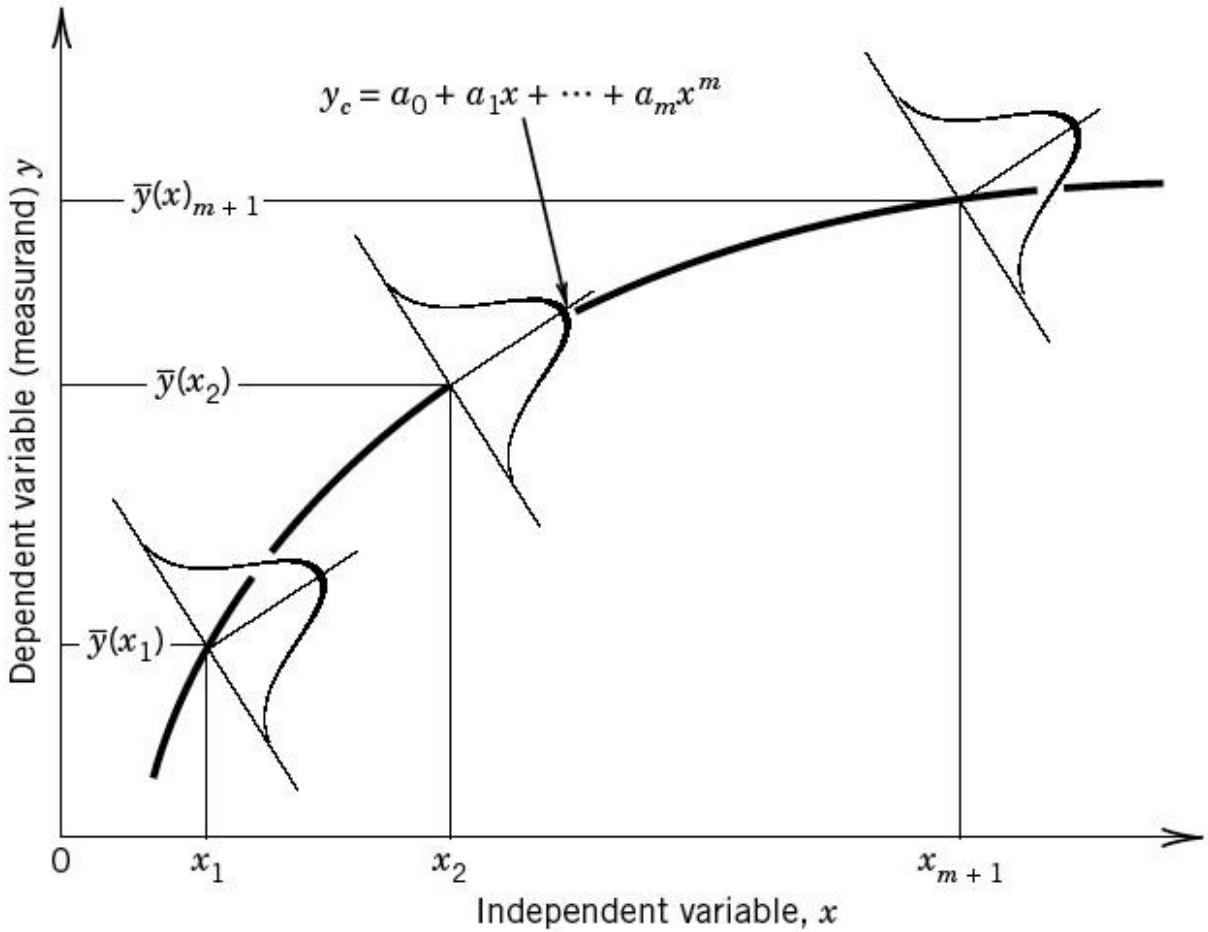


Figure 4.9 Distribution of measured value y about each fixed value of independent variable x. The curve y_c represents a possible functional relationship.

Regression Analysis

Least squares method

$$y_c = a_0 + a_1 x + a_2 x^2 + \ldots\ldots + a_m x^m$$

Number of constants to be found is m+1

Sum of square of deviations $$D = \sum_{i=1}^{N} (y_i - y_{ci})^2$$

$$D = \sum_{i=1}^{N} \left[y_i - \left(a_0 + a_1 x + a_2 x^2 + \ldots a_m x^m \right) \right]^2$$

Requirement: Reduce D. i.e. D$\rightarrow$0

Least squares method

Objective: Minimize the sum of squares of deviations

$$dD = \frac{\partial D}{\partial a_0} da_0 + \frac{\partial D}{\partial a_1} da_1 + \frac{\partial D}{\partial a_2} da_2 + \ldots . \frac{\partial D}{\partial a_m} da_m$$

$$\frac{\partial D}{\partial a_0} = 0 = \frac{\partial}{\partial a_0}\left\{ \sum_{i=1}^{N} \left[y_i - (a_0 + a_1 x + a_2 x^2 + \ldots a_m x^m) \right]^2 \right\}$$

$$\frac{\partial D}{\partial a_1} = 0 = \frac{\partial}{\partial a_1}\left\{ \sum_{i=1}^{N} \left[y_i - (a_0 + a_1 x + a_2 x^2 + \ldots a_m x^m) \right]^2 \right\}$$

$$\frac{\partial D}{\partial a_2} = 0 = \frac{\partial}{\partial a_2}\left\{ \sum_{i=1}^{N} \left[y_i - (a_0 + a_1 x + a_2 x^2 + \ldots a_m x^m) \right]^2 \right\}$$

Least squares method

$$\frac{\partial D}{\partial a_0} = 0 = \frac{\partial}{\partial a_0}\left\{\sum_{i=1}^{N}\left[y_i - (a_0 + a_1 x + a_2 x^2 + \ldots a_m x^m)\right]^2\right\}$$

$$2 * \left[\sum_{i=1}^{N}\left[y_i - (a_0 + a_1 x + a_2 x^2 + \ldots a_m x^m\right] * -1\right] = 0$$

$$\sum_{i=1}^{N} a_0 + a_1 \sum_{i=1}^{N} x_i + a_2 \sum_{i=1}^{N} x_i^{\,2} + \ldots.. = \sum_{i=1}^{N} y_i$$

$$\frac{\partial D}{\partial a_1} = 0$$ **Will give**

$$\sum_{i=1}^{N} a_0 x_i + a_1 \sum_{i=1}^{N} x_i^{\,2} + a_2 \sum_{i=1}^{N} x_i^{\,3} + \ldots.. = \sum_{i=1}^{N} y_i x_i$$

Least squares method

$$\frac{\partial D}{\partial a_0} = 0 \quad \rightarrow \quad \sum_{i=1}^{N} a_0 + a_1 \sum_{i=1}^{N} x_i + a_2 \sum_{i=1}^{N} x_i^{\;2} + ... = \sum_{i=1}^{N} y_i$$

$$\frac{\partial D}{\partial a_1} = 0 \quad \rightarrow \quad \sum_{i=1}^{N} a_0 x_i + a_1 \sum_{i=1}^{N} x_i^{\;2} + a_2 \sum_{i=1}^{N} x_i^{\;3} + ... = \sum_{i=1}^{N} y_i x_i$$

$$\frac{\partial D}{\partial a_2} = 0 \quad \rightarrow \quad \sum_{i=1}^{N} a_0 x_i^2 + a_1 \sum_{i=1}^{N} x_i^{\;3} + a_2 \sum_{i=1}^{N} x_i^{\;4} + ... = \sum_{i=1}^{N} y_i x_i^2$$

Least squares method

Least squares method for 2nd order curve fit

$$y_c = a_0 + a_1 x + a_2 x^2$$

$$\begin{bmatrix} N & \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^{\,2} \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^{\,2} & \sum_{i=1}^{N} x_i^{\,3} \\ \sum_{i=1}^{N} x_i^{\,2} & \sum_{i=1}^{N} x_i^{\,3} & \sum_{i=1}^{N} x_i^{\,4} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} x_i^{\,2} y_i \end{bmatrix}$$

Statistics of the fit

Standard error of the fit $$s_{yx} = \sqrt{\frac{\sum_i^N (y_i - y_{ci})^2}{\nu}}$$

ν is the degree of the freedom $$\nu = N - (m+1)$$

Considering the variation of both dependent and independent variables, the confidence interval

$$\pm t_{\nu,P} s_{yx} \left[\frac{1}{N} + \frac{(x-\bar{x})^2}{\sum_{i=1}^N (x_i - \bar{x})^2} \right]^{1/2} \quad (P\%)$$

If only y variation is considered (common in measurement) then the curve fit is statistically described by:

$$y_c \pm t_{\nu,P} \frac{s_{yx}}{\sqrt{N}} \quad (P\%)$$

Linear Polynomial

Correlation coefficient

$$r = \sqrt{1 - \frac{S_{yx}{}^2}{S_y{}^2}}$$

Coefficient of determination, r^2

Where

$$s_y{}^2 = \frac{1}{N-1}\sum_i^N (y_i - \bar{y})^2$$

When $\pm 0.9 < r < \pm 1.0$ **Good or reliable fit**

R^2 is called the coefficient of determination. Excel Tendline can show this factor on the curve

r and r^2 are not effective estimators of the random error in y_c

Linear Curve fit

$$y_c = a_0 + a_1 x$$ **With N data points**

$$\begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^{\,2} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{bmatrix}$$

Examples 4.8 & 4.9

$$\begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^{\ 2} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{bmatrix} \qquad \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 15.7 \\ 57.5 \end{bmatrix}$$

	x_i	y_i
	1	1.2
	2	1.9
	3	3.2
	4	4.1
	5	5.3
N =	5	
Σ =	15	15.7

$$5a_0 + 15a_1 = 15.7$$

$$15a_0 + 55a_1 = 57.5$$

Two equations in two unknowns

a_0 =0.02, a_1=1.04, r=0.9965 (correlation coefficient)

Σyx=57.5, Σx^2=55

$$y_c = 0.02 + 1.04x \quad V$$

If you have CASIO 880P use 6510 LIB

Examples 4.8 & 4.9 Continue

$$v = N - (m+1) = 5 - (2) = 3$$

$$s_{yx} = \sqrt{\frac{\sum_i^N (y_i - y_{ci})^2}{v}} = 0.16$$

From table 4.4 $t_{v,P} = 3.18 \qquad (P = 95\%)$

Uncertainty interval for probability of 95%

$$\pm t_{v,P} \frac{s_{xy}}{\sqrt{N}} \quad (P\%) \qquad \pm 3.18 \frac{0.16}{\sqrt{5}} = \pm 0.23 \ (95\%)$$

$$1.04x + 0.02 \pm 0.23 \quad (95\%)$$

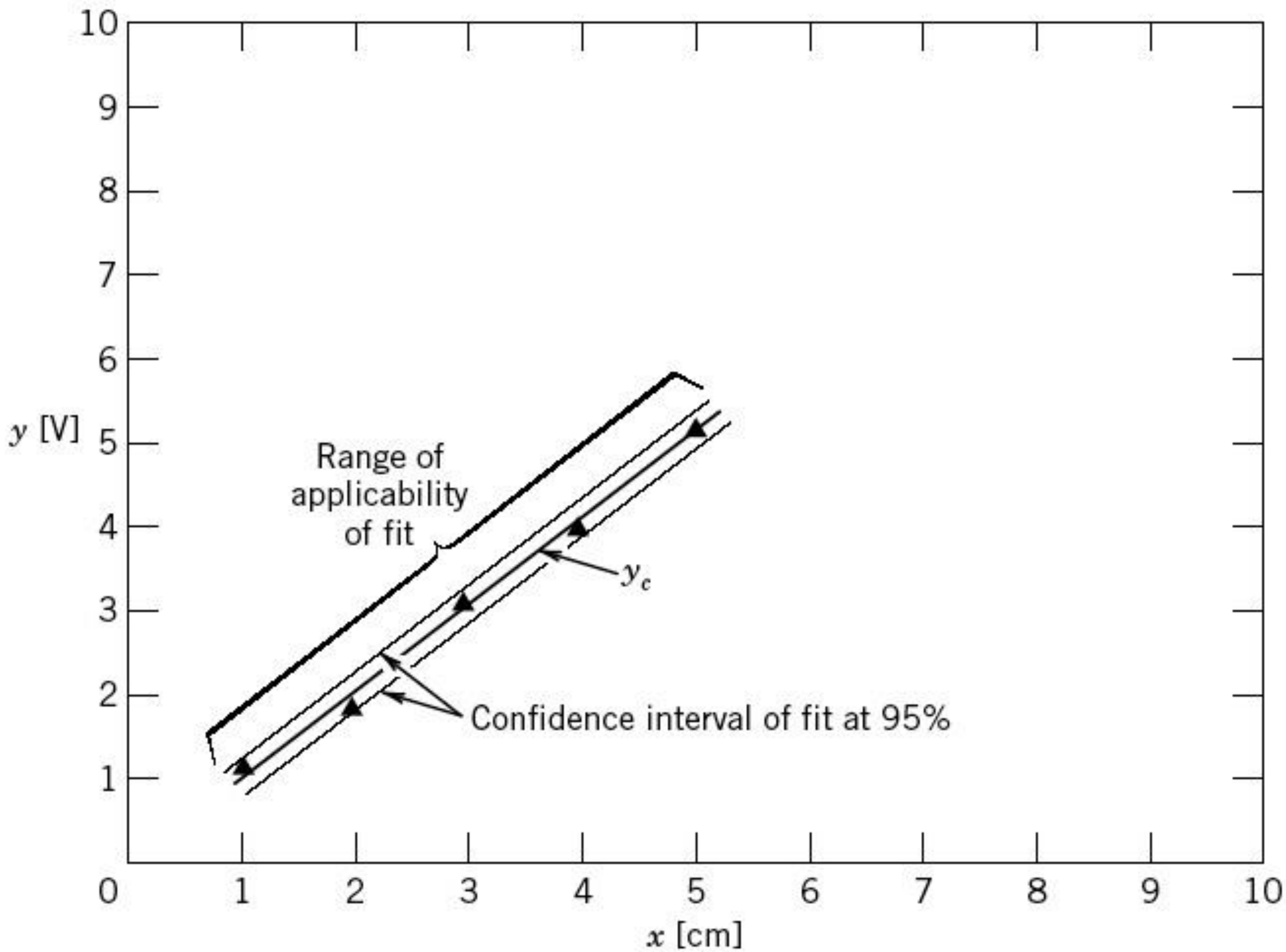


Figure 4.10 Results of the regression analysis of Example 4.9.

Summary of relations

Table 4.7 Summary Table for a Sample of N Data Points

Sample mean	$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
Sample standard deviation	$s_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$
Standard deviation of the means [a]	$s_{\bar{x}} = \frac{s_x}{\sqrt{N}}$
Precision interval for a single data point, x_i	$\pm t_{\nu,P} s_x \quad (P\%)$
Confidence interval [b,c] for a mean value, $\bar{x}$	$\pm t_{\nu,P} s_{\bar{x}} \quad (P\%)$
Confidence interval[b,d] for curve fit, $y = f(x)$	$\pm t_{\nu,P} \frac{s_{yx}}{\sqrt{N}} \quad (P\%)$

[a]Measure of random standard uncertainty in x.

[b]In the absence of systematic errors.

[c]Measure of random uncertainty in $\bar{x}$.

[d]Measure of random uncertainty in curve fit (see conditions of Eqs. 4.37–4.39).

Example 4.10

See your textbook

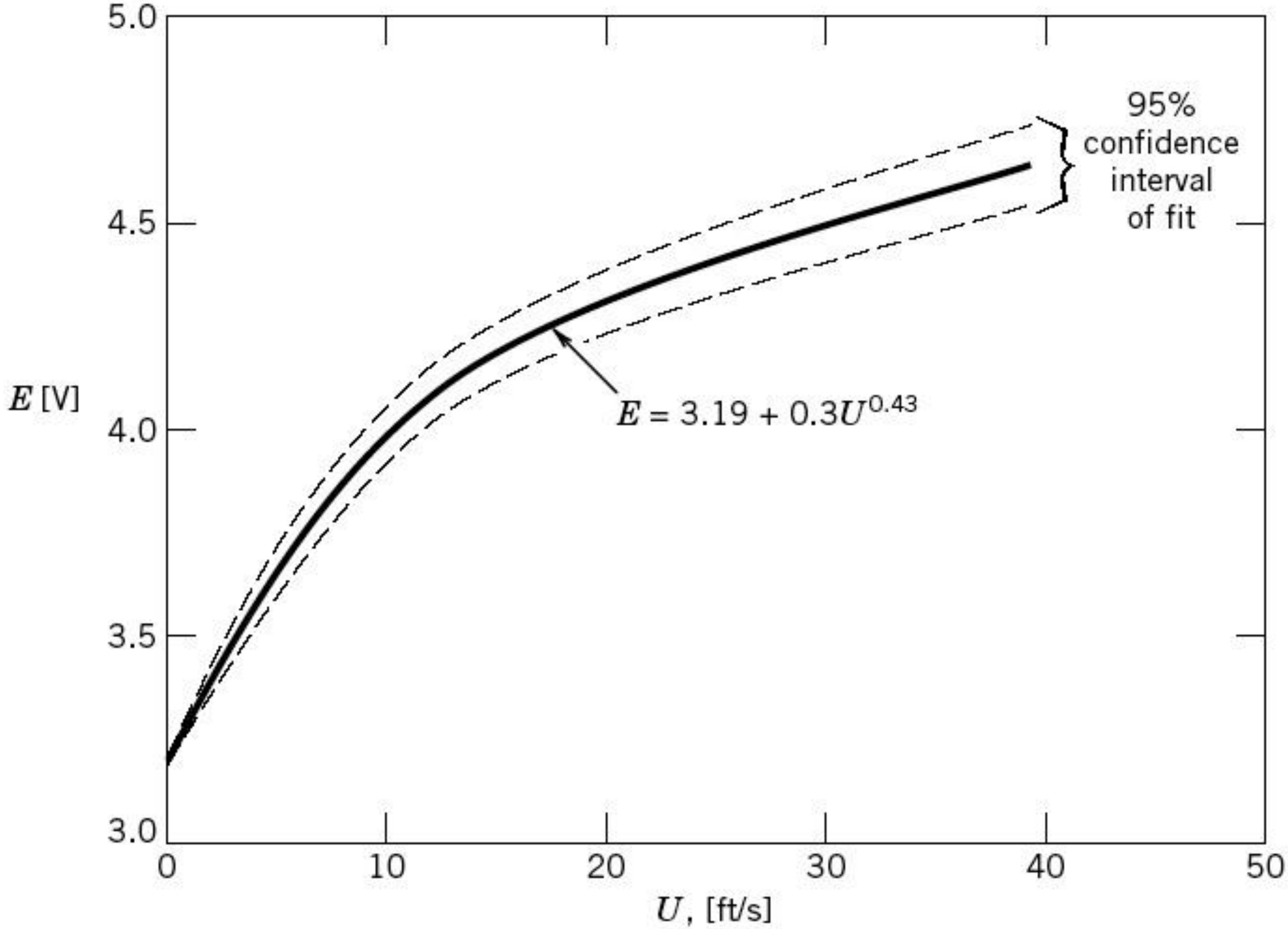


Figure 4.11 A curve fit for Example 4.10.

Data Outlier Detection

Wrong data causes

- offset the mean
- inflate the random error
- influence the least square correlation

How to detect data that is outside the normal variation?

Once the outlier data is removed, the statistics are re-calculated

Data Outlier Detection

Chauvenet's criterion

Outlier data point having less than 1/2N probability of occurrence

Test criterion

Calculate sample statistics i.e. $\bar{x}$ and s_x

Calculate $$z_0 = \frac{x - \bar{x}}{s_x}$$

if $$[1 - 2P(z_0)] < \frac{1}{2N}$$ Data point could be rejected.

Example 4.11

i	1	2	3	4	5	6	7	8	9	10
x_i	28	31	27	28	29	24	29	28	**18**	27

Required: Statistics and outliers

From table 4.3

$$\bar{x} = 27, \quad s_x = 3.8$$

For data point x=18

$$z_0 = \left|\frac{18 - 27}{3.8}\right| = 2.368, \quad P(z_0) = 0.4910$$

Chauvenet's criterion $[1 - 2P(z_0)] < \frac{1}{2N}$ 1/(20)=0.05

[1-2P(z0)]=[1-2*0.4910]=0.018 ≤ 0.05

Therefore this data point can be rejected

For the remaining 19 data points $\bar{x} = 28, \quad s_x = 2.0$

Number of measurements required

Range of values of x with certain probability

$$x' = \bar{x} \pm t_{\nu,P} s_{\bar{x}} \qquad (P\%)$$

Confidence interval CI

$$CI = \pm t_{\nu,P} s_{\bar{x}} = \pm t_{\nu,P} \frac{s_x}{\sqrt{N}}$$

One sided precision d=CI/2= $\dfrac{t_{\nu,P} s_x}{\sqrt{N}}$

$$N = \left(\frac{t_{\nu,95} s_x}{d} \right)^2 \qquad (95\%)$$

This is equation has two unknowns N and s_x

$$N = \left(\frac{t_{v,95} s_x}{d} \right)^2 \qquad (95\%)$$

A trail and error procedure is utilized to find N

Or If for N_1 measurements one has calculate s_1 then

$$N_T = \left(\frac{t_{N-1,95} s_1}{d} \right)^2 \qquad (95\%)$$

Additional N_T-N_1 measurements will be required

Example 4.13 Given: 21 measurements, S_1=160, CI=30 units, P=95%

Required: Total number of measurements required

$$d = \frac{CI}{2} = 15$$

$t_{v,P}=t_{20,95}=2.093$

Use $$N_T = \left(\frac{t_{N-1,95} s_1}{d}\right)^2 \qquad (95\%)$$

$$N_T = \left(\frac{2.093*160}{15}\right)^2 = 125 \qquad (95\%)$$

Therefore additional (125-21)=104 measurements will be required to achieved the required confidence interval