MECE 3320 – Measurements & Instrumentation

Probability & Statistics

Dr. Isaac Choutapalli
Department of Mechanical Engineering
University of Texas – Pan American

Introduction

Suppose we have a large box containing thousands of similar bearings:

- We can choose a subset of bearings to determine the mean diameter of ALL the bearings
- But, another subset can yield different mean diameter
- So, how good is your measurement?

Sources that contribute to variation in measurements:

- Resolution
- Repeatability
- Spatial and temporal variation of the measured variables

So how do you ensure that a measurement represents a true average value?

- Quantify a representative value that characterizes the average value of the data set.
- Quantify a average value that provides a measure of the variation in the measured data set.
- Establish the bounds of the measured variable.

Statistical Measurement Theory

Sample: Data set obtained during repeated measurements of a variable, called the measurand, under fixed operating conditions.

Consider *x'* to be the ***true value*** based on repeated measurements of *x*.

N: Sample size

If *N* is small, the measurement error is high. As N -> ∞, measurement error tends to become smaller.

X' can be estimated as $x' = \bar{x} \pm u_x$

$\bar{x}$ represents the most probable estimate of *x'*

$\pm u_x$ represents the uncertainty interval at some probability level, *P%.*

Probability Density Functions

Regardless of the care taken during measurments, random scatter in the data will always occur.

Random variable: The measured variable

Central Tendency: The tendency of a data point to lie within some interval about one central value

Probability deals with the concept that certain values for a variable will be measured with some frequency of occurrence relative to others.

Probability density is used to measure this central value and the values scattered around it. It is the frequency with which a measured value assumes a particular value or interval of values.

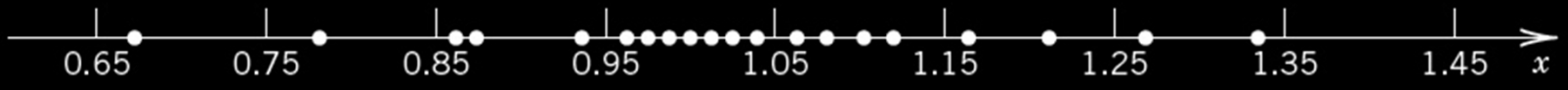


Figure 4.1 Concept of density in reference to a measured variable (from Example 4.1).

Probability Density Functions

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

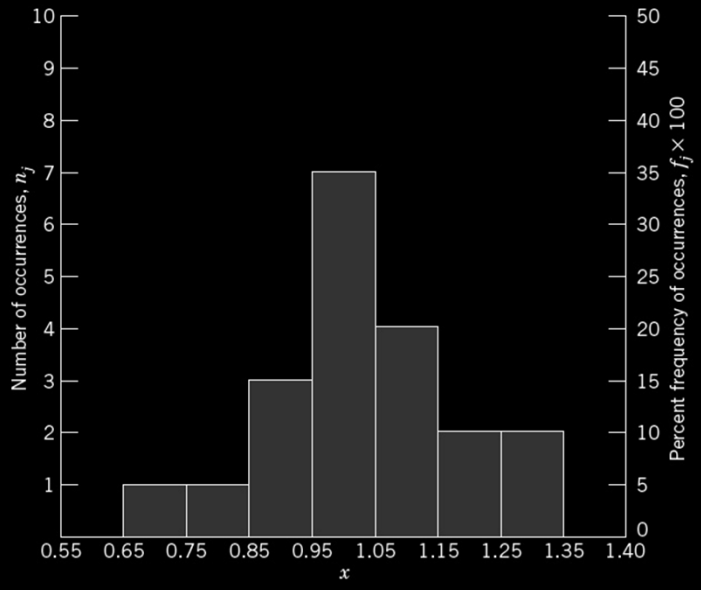


Figure 4.2 Histogram and frequency distribution for data in Table 4.1.

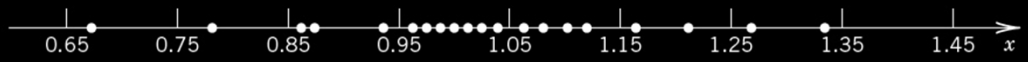


Figure 4.1 Concept of density in reference to a measured variable (from Example 4.1).

The number of intervals *K* required for a viable statistical analysis is given by

$$K = 1.87(N-1)^{0.40} + 1$$

Dividing n_j (no. of times data falls within an interval) by *N* gives the frequency distribution for the data. The area under the curve gives the total frequency of occurrence of 100%.

Probability Density Functions

Probability Density function, *p(x)* results from frequency distribution when N ->∞ and δx -> 0

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

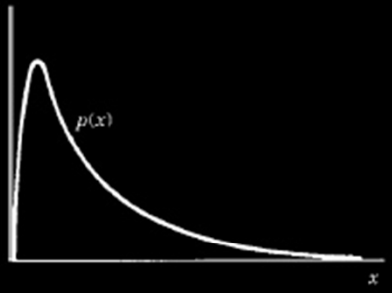
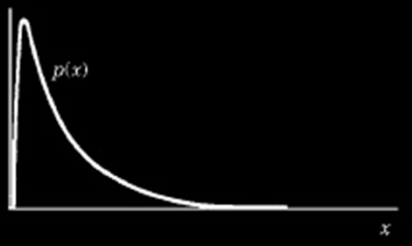
p(x) defines the probability that a measured value might assume a particular value during a measurement.

Shape of p(x) depends on the variable being measured and the measurement conditions affecting the measurement.

The specific values of the variable and the width of the distribution depends on the actual process but the overall shape will likely fit some standard distribution.

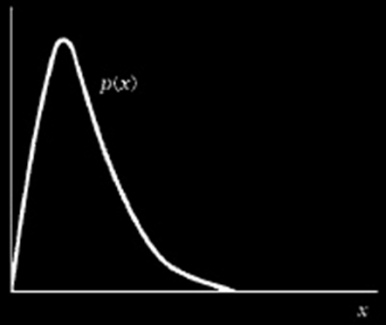
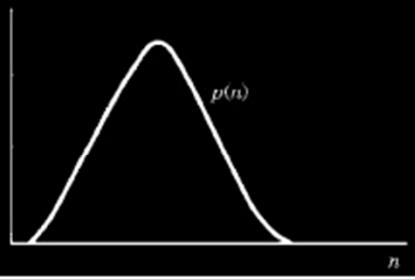
Probability Density Functions

Table 4.2 Standard Statistical Distributions and Relations to Measurements

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

Probability Density Functions

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.	$p(n) = \left[\frac{N!}{(N-n)!n!}\right] P^n (1-P)^{N-n}$	

Probability Density Functions

Regardless of the type of distribution, a variable can be quantified through its mean and variance.

$$True\ mean\ value\ x' = \lim_{T\to\infty}\frac{1}{T}\int_0^T x(t)dt = \int_{-\infty}^{\infty} xp(x)dx$$

$$True\ \text{var}iance\ \sigma^2 = \lim_{T\to\infty}\frac{1}{T}\int_0^T \left[x(t)-x'\right]^2 dt = \int_{-\infty}^{\infty}\left(x-x'\right)^2 p(x)dx = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left(x_i-x'\right)^2$$

The square root of the variance is defined as the *standard deviation.*

The underlying assumption in the above equations is that they assume an infinite number of measurements.

Infinite Statistics

Most common distribution in measurements is the normal or Gaussian distribution (bell curve), i.e. data is scattered around a central value.

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

$p(x)$ is maximum when $x = x'$, i.e. in the absence of systematic error, a variable having a normal distribution has a central tendency towards its true value.

The probability that a random variable x falls between $x' \pm \delta x$ is given by the area under $p(x)$

$$p(x' - \delta x \le x \le x' + \delta x) = \int_{x'-\delta x}^{x'+\delta x} p(x)dx = 2\left[\frac{1}{\sqrt{2\pi}} \int_0^{z_1} e^{-\frac{\beta^2}{2}} d\beta \right] = 2(error\ function)$$

The area under $p(x)$ defined by $x' - z_1\sigma \le x \le x' + z_1\sigma$ is the probability that a measurement will lie within this interval.

Infinite Statistics

$$p(x'-\delta x \le x \le x'+\delta x) = \int_{x'-\delta x}^{x'+\delta x} p(x)dx = 2\left[\frac{1}{\sqrt{2\pi}}\int_{0}^{z_1} e^{-\frac{\beta^2}{2}}\,d\beta\right] = 2(error\ function)$$

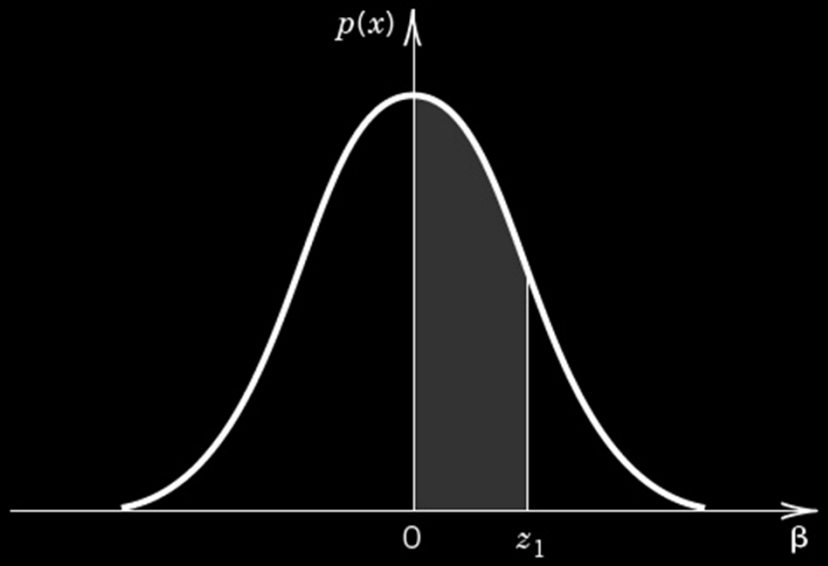


Figure 4.3 Integration terminology for the normal error function.

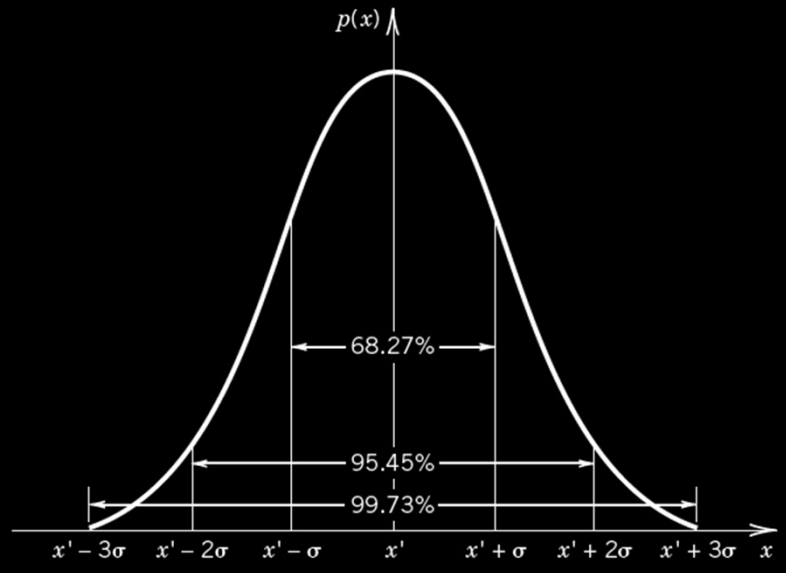


Figure 4.4 Relationship between the probability density function and its statistical parameters, x' and σ, for a normal (Gaussian) distribution.

z_1 = 1.0: 68.26% of the area under *p(x)* lies within $\pm z_1\,\sigma$ of *x'*.
z_2 = 2.0: 95.45% of the area under *p(x)* lies within $\pm z1\,\sigma$ of *x'*.
z_3 = 3.0: 99.73% of the area under *p(x)* lies within $\pm z1\,\sigma$ of *x'*.

Infinite Statistics

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

Infinite Statistics

A very large data set (N > 10,000) has a mean value of 9.2 units and a standard deviation of 1.1 units. Determine the range of values in which 50 % of the data set should be found assuming a normal probability density.

Finite Statistics

Is it possible to obtain the true mean and variance if the sample size is finite? *Yes we can.*

For finite sized data sets:

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

$$Variance\ S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$\text{Standard deviation } S_x = \sqrt{S_x^2}$$

The sample mean provides the *most probable estimate* of the true value *x'.*

$(x_i - \bar{x})$ is called the deviation of x_i *and (N-1)* are the degrees of freedom in the statistical estimate.

The degrees of freedom is equal to the number of data points minus the number of previously determined statistical parameters used in estimating that value.

Finite Statistics

For a normal distribution of *x* about a sample mean value,

$$x_i = \bar{x} \pm t_{v,P} S_x$$

Here $t_{v,P}$ replaces the z variable. The numerical values of $t_{v,P}$ are given by the *Student's t-distribution*.

The Student's t-distribution was developed by William S. Gosset.

The interval $\pm t_{v,P} S_x$ represents the precision interval at some probability (confidence level) *P%* within which the measured value would fall.

As the value of *N* increases, the value of *t* approaches the values given by the *z* variable.

Student's t-distribution

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

Suppose we were to measure the sizes of bearing from different number of samples, the *mean from each sample would be different* due to finite sample size and random variation in the bearing size from manufacturer's tolerances.

So what is a good estimate of the true mean based on sample mean?

Let's say we measure a variable N times and duplicate this procedure M times, the sample mean and sample variance for each of the M data sets would be different.

However, the mean values obtained from M replications will themselves follow a normal distribution

Standard Deviation of the Means

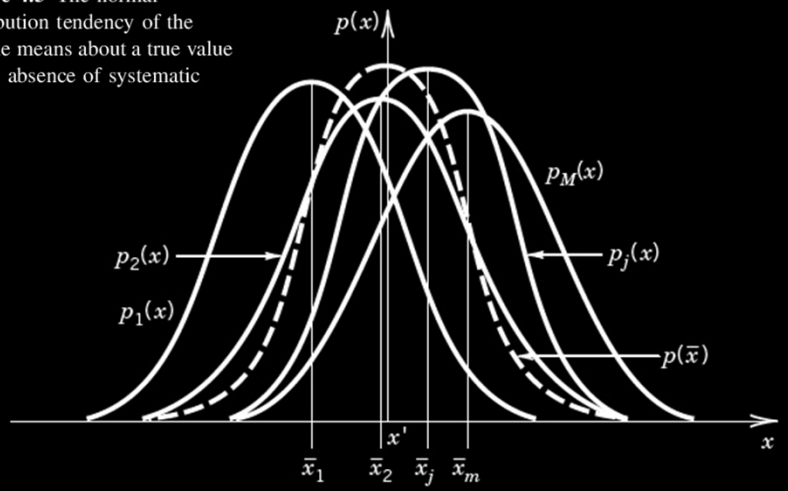


Figure 4.5 The normal distribution tendency of the sample means about a true value in the absence of systematic error.

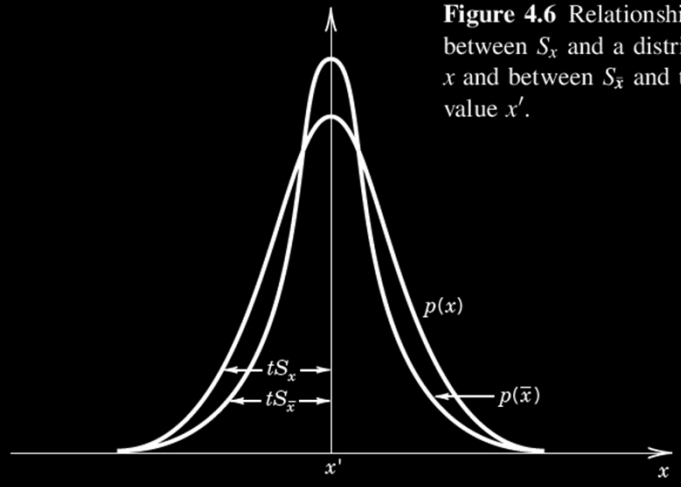


Figure 4.6 Relationships between S_x and a distribution of x and between $S_{\bar{x}}$ and the true value x'.

The variation in the sample statistics are characterized by a normal distribution of the sample mean values about a true value. *The variance of the distribution of mean values from different data sets can be estimated from a single finite data set through the standard deviation of means,* $S_{\bar{x}}$

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

$S_{\bar{x}}$ represents a measure of how well the sample mean represents the true mean. The range over which the true mean might lie is given by: $\bar{x} - t_{v,P} S_{\bar{x}} \le x' \le \bar{x} + t_{v,P} S_{\bar{x}}$

$t_{v,P} S_{\bar{x}}$ is the random uncertainty in the mean value due to variation in the measured data set.

Pooled Statistics

A good test plan uses both repetition and replication. Since replication is an independent estimate of the same measured value, its data represents separate data samples that can be combined to provide a better statistical estimate of the measured variable. *When samples are grouped together in this manner, they are said to be pooled*.

If the number of replications are *M* and the number of repetitions are *N*, the pooled mean is given by

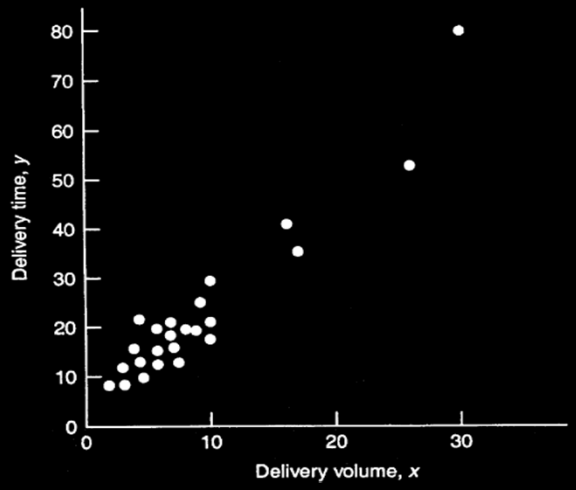
$$\left\langle \bar{x} \right\rangle = \frac{1}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N} x_{ij}$$

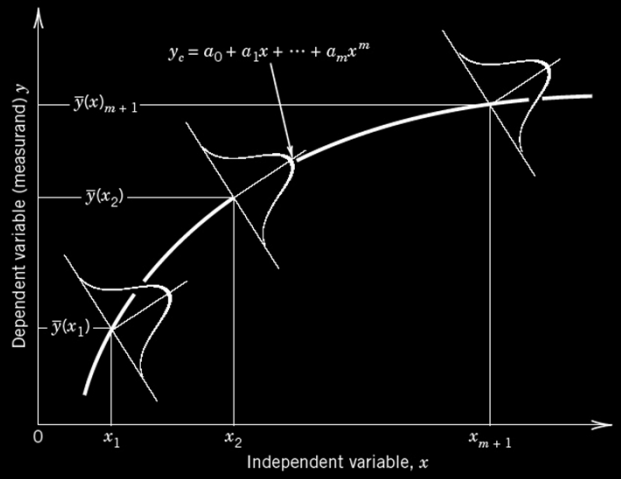
The pooled standard deviation is given by $\left\langle S_x \right\rangle = \sqrt{\frac{1}{M(N-1)} \sum_{j=1}^{M} \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2} = \sqrt{\frac{1}{M} \sum_{j=1}^{M} S_{x_j}^2}$

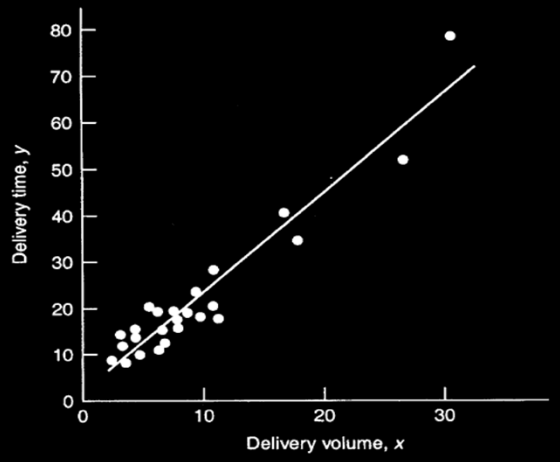
The pooled standard deviation of means is given by $\left\langle S_{\bar{x}} \right\rangle = \frac{\left\langle S_x \right\rangle}{\sqrt{MN}}$

Regression Analysis

A regression analysis is a statistical technique for modeling the relationship between a dependent and an independent variable.







Regression analysis is useful for parameter estimation and prediction, data description and control.

The regression is only an approximation of the true functional relationship between the variables.

Linear-Squares Regression Analysis

Linear squares regression analysis for *y=f(x)* provides an m[th] order polynomial fit of the form

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

The values of the coefficients are determined by the method of least squares. *This method minimizes the sum of the squares of deviations between the actual data and the polynomial fit by adjusting the values of the coefficients.*

For linear polynomials, the correlation coefficient represents the quantitative measure of the linear association between *x and y.*

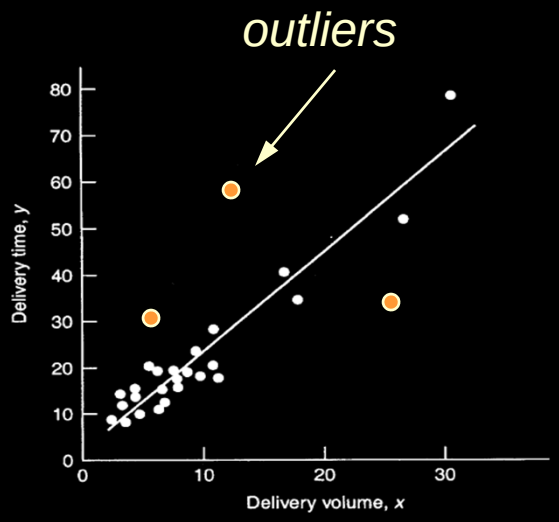
$$r = \sqrt{1 - \frac{S_{yx}^2}{S_y^2}}$$

$$\text{where } S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2; \qquad S_{yx} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}{\upsilon}} \text{ (standard error of fit)}$$

Data Outlier Detection

Outliers are the data that lie outside the probability of normal variation.

- incorrectly offsets the sample mean value estimate
- inflates the random error estimates
- influences the least squares correlation



Each data point can be checked if it is an outlier by using the *three-sigma test*.

$$z_0 = \left| \frac{x_i - \bar{x}}{S_x} \right|$$

Probability that *x* lies outside the one-sided range defined by *0 and* Z_0 *is* $[0.5\text{-}P(z_0)]$.

For N points, if N $[0.5\text{-}P(z0)] \leq 0.1$, *the data point can be considered as an outlier.*

Number of Measurements Required

How many number of measurements do we need to make to get a good estimate of a the variable being measured?

There is *no fixed formula* for estimating the exact number of measurements. However the *following techniques will help in giving an idea about the approximate number of measurements required*:

a) The degree of statistical convergence
b) By assuming a value for the sample standard deviation (a very crude technique)

$$N = \left(\frac{t_{\upsilon,95} S_x^2}{d} \right)^2 \text{ at } 95\% \text{ confidence level}$$

$$d = \frac{95\% \text{ confidence interval}}{2}$$