2141-375
Measurement and Instrumentation

Probability and Statistics

Statistical Measurement Theory

A *sample* of data refers to a set of data obtained during repeated measurements of a variable under fixed operating conditions.

The estimation of true mean value, μ from the repeated measurements of the variable, x. So we have a sample of variable of x under controlled, fixed operating conditions from *a finite number of data points.*

$$\mu = \bar{x} \pm u_x \ (P\%)$$

$\bar{x}$ **is the sample mean**

u_x **is the confidence interval or uncertainty in the estimation at some probability level, $P\%$. The confidence interval is based both on estimates of the precision error and on bias error in the measurement of x.** ***(in this chapter, we will estimation μ and the precision error in x caused only by the variation in the data set)***

Infinite Statistics: Normal Distribution

A common distribution found in measurements e.g. the measurement of length, temperature, pressure etc.

The probability density function for a random variable, x having normal distribution is defined as

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

μ is the true mean and σ^2 is the true variance of x

Notation for random variable X has a normal distribution with mean μ and σ variance

$$X \sim N(\mu, \sigma^2)$$

The probability, $P(x)$ within the interval a and b is given by the area under $p(x)$

$$P(a \le x \le b) = \int_a^b p(x)dx$$

The Standard Normal Distribution

The standard normal distribution:

$$X \sim N(0,1)$$

The probability density function has the notation, $p(x)$

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

for $-\infty \le x \le \infty$

The cumulative distribution function of a standard normal distribution

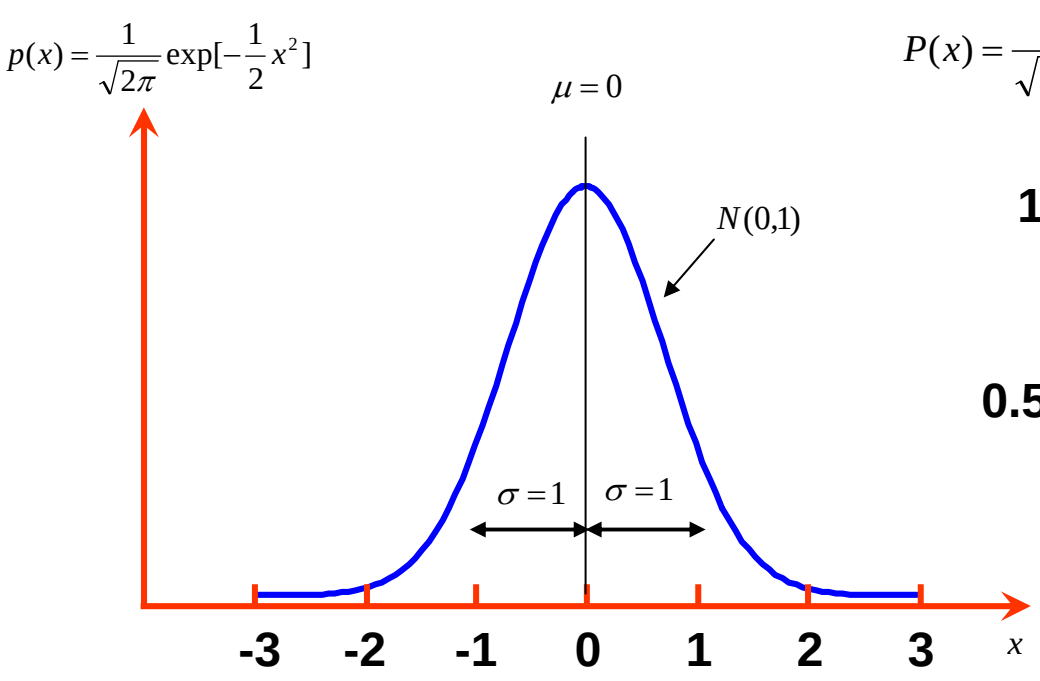
$$P(x) = \int_{-\infty}^{x} p(y)\, dy$$

$$P(-\infty \le x \le a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} \exp[-\frac{x^2}{2}]dx$$

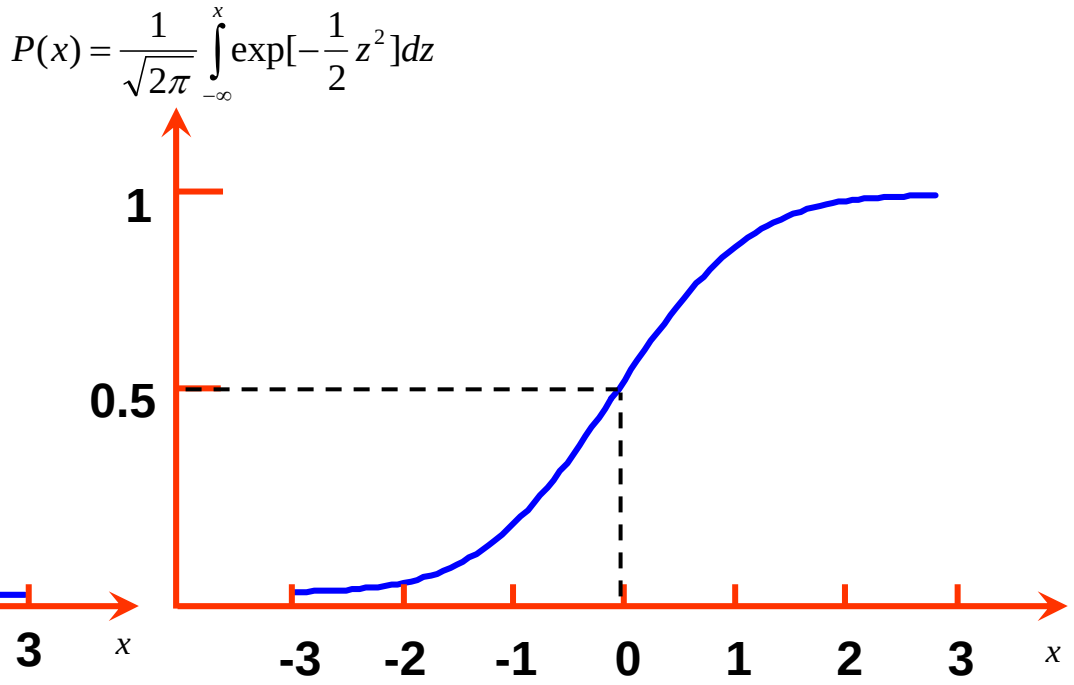
The Normal error function

$$P(0 \le x \le a) = \frac{1}{\sqrt{2\pi}} \int_{0}^{a} \exp[-\frac{x^2}{2}]dx$$

The Standard Normal Distribution

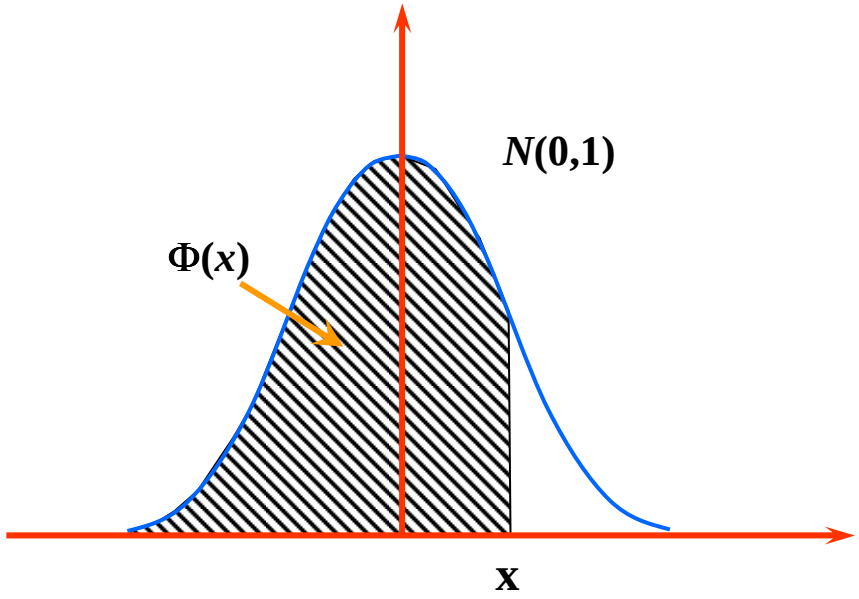


The standard normal distribution



The cumulative distribution of the standard normal distribution

The Standard Normal Distribution



$$\Phi_{(x)} = P(Z \leq x)$$

The symmetry of the standard normal distribution about 0 implies that if the random variable Z has a standard variable normal distribution, then

$$1-\Phi(x) = P(Z \geq x) = P(Z \leq -x) = \Phi(-x)$$

$$\Phi(x) + \Phi(-x) = 1$$

Probability Calculations for Normal Distribution

If $X \sim N(\mu,\sigma^2)$, then

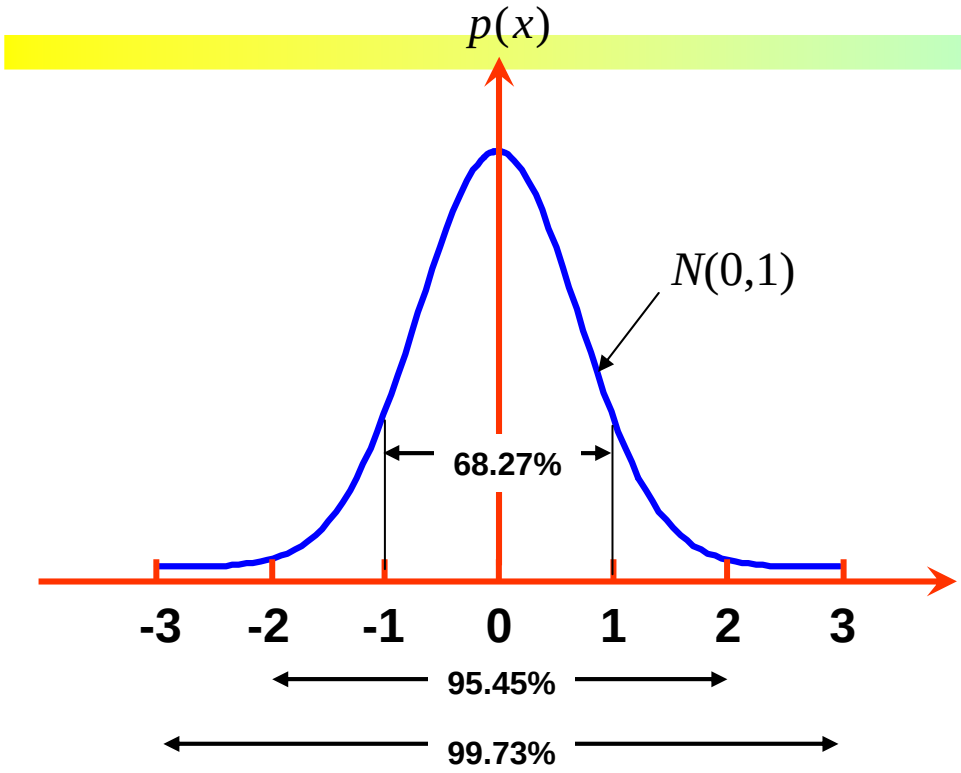
$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

The random variable Z is known as the "standardized" version of the random variable, X. This results implies that the probability values of a general normal distribution can be related to the cumulative distribution of the standard normal distribution $\Phi(x)$ through the relation

$$P(a \le x \le b) = P\left(Z \le \frac{b-\mu}{\sigma}\right) - P\left(Z \le \frac{a-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

The Normal Distribution



If $X \sim N(\mu,\sigma^2)$, notice that

$$P(\mu - c\sigma \le x \le \mu + c\sigma) = P(-c \le Z \le c)$$

Normal random variables

There is a probability about 68% that a normal random variable takes a value within one SD of its mean

There is a probability about 95% that a normal random variable takes a value within two SD of its mean

There is a probability about 99.7% that a normal random variable takes a value within three SD of its mean

The Normal Distribution

Example: It is known that the statistics of a well-defined voltage signal are given by μ = 8.5 V and σ^2 = 2.25 V^2. If a single measurement of the voltage signal is made, determine the probability that the measured value will be between 10.0 and 11.5 V

Known: μ = 8.5 V and σ^2 = 2.25 V^2

Assume: Signal has a normal distribution

Solution:

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

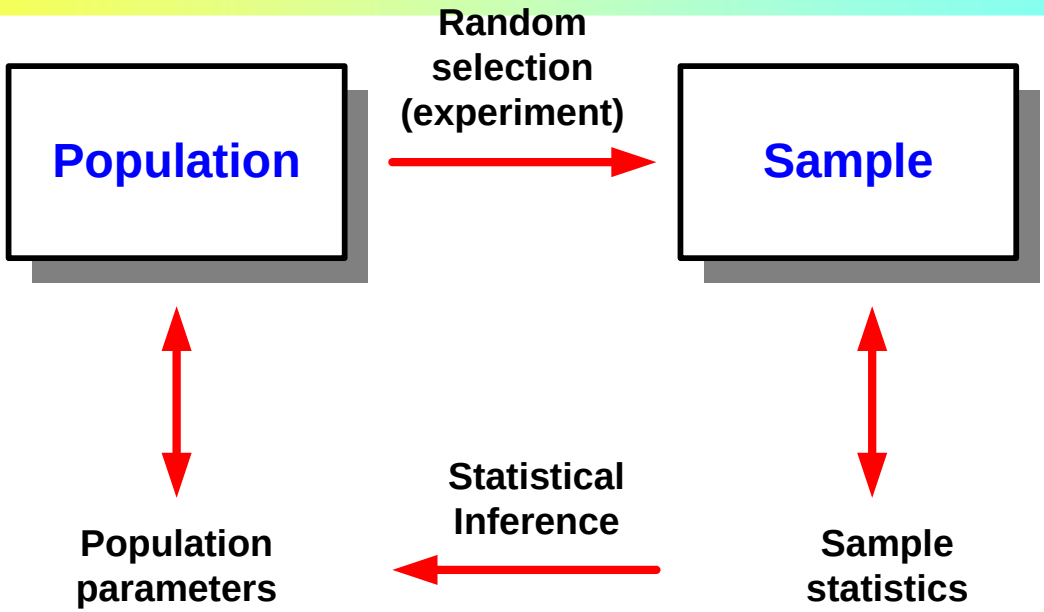
$$= P[Z \le (11.5\text{-}8.5)/1.5] - P[Z \le (10.0\text{-}8.5)/1.5]$$

$$= P[Z \le 2] - P[Z \le 1]$$

$$= 0.9772 - 0.8413$$

$$P(10.0 \le x \le 11.5) = 0.1359$$

Finite Statistics: Sample Versus Population



The finite sample versus the infinite population

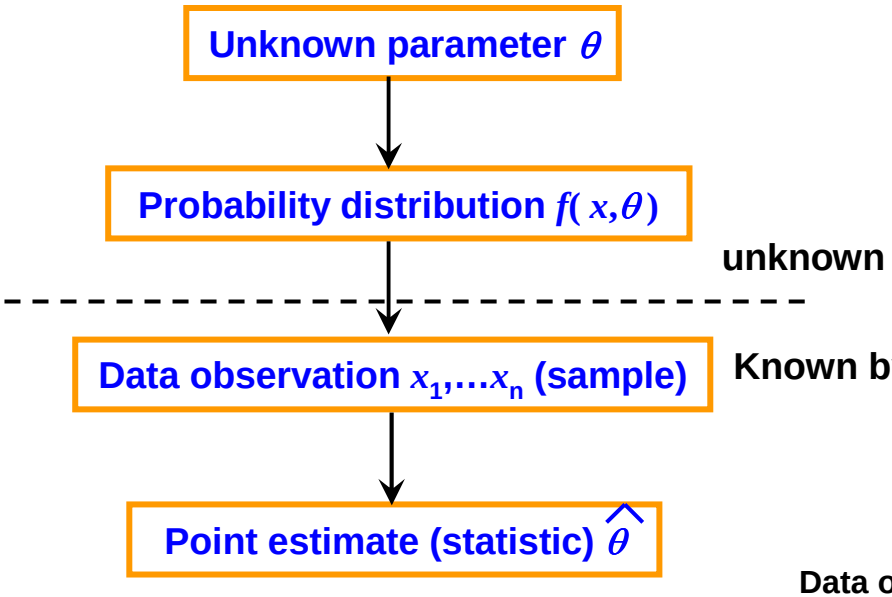
Parameter is a quantity that is a property of unknown probability distribution. This may be the mean, variance or a particular quantity.

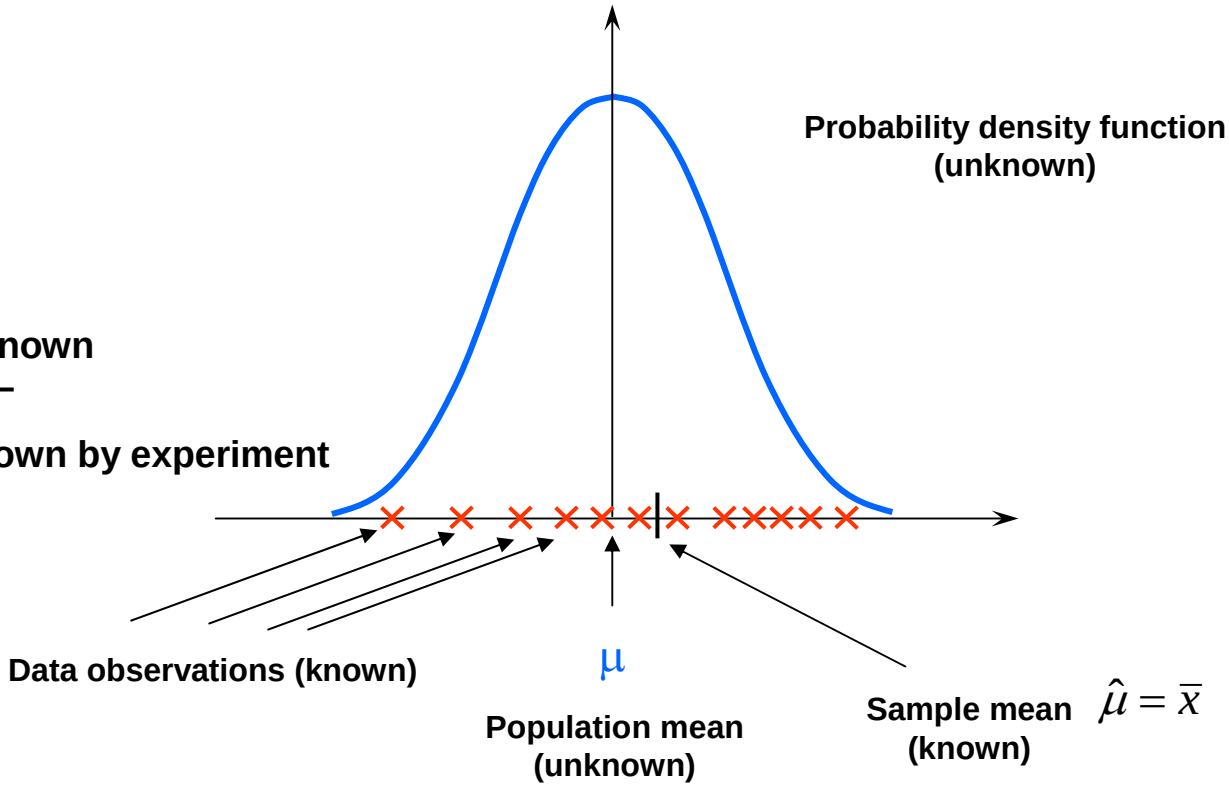
Statistic is a quantity that is a property of sample. This may be the mean, variance or a particular quantity. Statistics can be calculated from a set of data observations.

Estimation is a procedure by which the information contained within a sample is used to investigate properties of the population from which the sample is drawn.

Point Estimates of Parameters

A point estimate of unknown parameter θ is a statistic $\hat{\theta}$ that represent of a "best guess" at the value of θ.





Estimation of the population mean by the sample mean

Point Estimation of Mean and Variance

Sample mean: Point Estimate of a Population Mean

If $X_1, \ldots, X_n$ is a sample of observation from a probability distribution which a mean μ, then the sample mean

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

is the best guess of the point estimate of the population mean μ

Sample variance: Point Estimate of a Population Variance:

If $X_1, \ldots, X_n$ is a sample of observation from a probability distribution which a variance σ^2, then the sample variance

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

is the best guess of the point estimate of the population variance σ^2

Inference of Population Mean

For a normal distribution of x about some sample mean value, $\bar{x}$ one can state that

$$x_i \in \bar{x} \pm t_{v,P} S_x \ (P\%)$$

Where the variable $t_{v,p}$ is a function of the probability P, and the degree of freedom $v = n - 1$ of the Student-t distribution.

The standard deviation of the mean

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

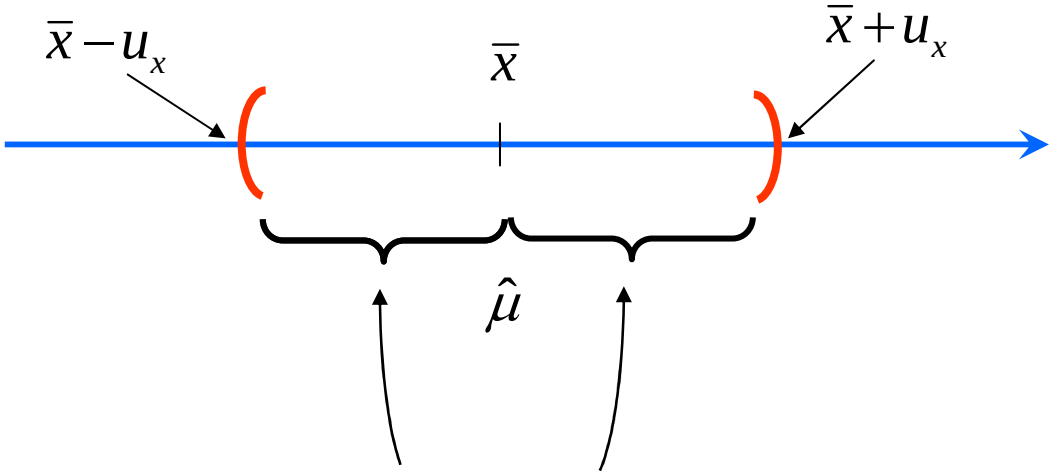
The estimate of the true mean value based on a finite data set is

$$\mu \in \left(\bar{x} - t_{v,P} S_{\bar{x}}, \bar{x} + t_{v,P} S_{\bar{x}}\right) \text{or } \mu \in \left(\bar{x} - t_{v,P} S_x / \sqrt{n}, \bar{x} + t_{v,P} S_{\bar{x}}\right)$$

Inference of Population Mean

Inference methods on a population mean based on the *t*-procedure for large sample size $n \geq 30$ and also for small sample sizes as long as the data can reasonably be taken to be approximately normally distributed. Nonparametric techniques can be employed for small sample sizes with data that are clearly not normally distributed

$$\mu \in (\bar{x} - u_x, \bar{x} + u_x)(P\%)$$



Precision error u_x

with confident level P or $1 - \alpha$

Inference of Population Mean

Example: Consider the data in the table below. (a) Compute the sample statistics for this data set. (b) Estimate the interval of values over which 95% of the measurements of the measurand should be expected to lie. (c) Estimate the true mean value of the measurand at 95% probability based on this finite data set

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

Known: the given table, N = 20

Assume: data set follows a normal distribution

Solution:

Inference of Population Variance

For a normal distribution of x, sample variance S^2 has a probability density function of chi-square χ^2

$$\chi^2 \sim \nu S_x^2 / \sigma^2$$

with the degree of freedom $v = n - 1$

Precision Interval in a Sample Variance

$$P(\chi^2_{1-\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}) = 1 - \alpha$$

with a probability of $P(\chi^2) = 1 - \alpha$

$$P(\nu S_x^2/\chi^2_{\alpha/2} \le \sigma^2 \le \nu S_x^2/\chi^2_{1-\alpha/2}) = 1 - \alpha$$

$$\sigma^2 \in (\nu S_x^2/\chi^2_{\alpha/2}, \nu S_x^2/\chi^2_{1-\alpha/2}) \quad (P = 1 - \alpha)$$

Inference of Population Variance

Example: Ten steel tension specimens are tested from a large batch, and a sample variance of $(200\ \text{kN/m}^2)^2$ is found. State the true variance expected at 95% confidence.

Known: $S^2 = 40000\ (\text{kN/m}^2)^2$, $N = 10$

Solution: assume data set follows a normal distribution, with $v = n - 1 = 9$

From chi-square table $\chi^2 = 19$ at $\alpha = 0.025$ and $\chi^2 = 2.7$ at $\alpha = 0.975$

$$9 \cdot 40000 / 19 \le \sigma^2 \le 9 \cdot 40000 / 2.7$$

$$138^2 \le \sigma^2 \le 365^2 \quad (95\%)$$

Pooled Statistics

Consider M replicates of a measurement of a variable, x, each of N repeated readings so as to yield that data set x_{ij}, where i = 1,2,…, N and j = 1, 2, …, M.

The pooled mean of x

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

The pooled standard deviation of x

$$\langle S_x \rangle = \sqrt{\frac{1}{M(N-1)} \sum_{j=1}^{M} \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2} = \sqrt{\frac{1}{M} \sum_{j=1}^{M} S_{x_j}^2}$$

With degree of freedom $v = M(N\text{-}1)$

The pooled standard deviation of the means of x

$$\langle S_{\bar{x}} \rangle = \frac{\langle S_x \rangle}{(MN)^{1/2}}$$

Data outlier detection

Detect data points that fall outside the normal range of variation expected in a data set base on the variance of the data set. This range is defined by some multiple of the standard deviation.

Ex. 3-sigma method: consider all data points that lie outside the range of 99.8% probability, $\bar{x} \pm t_{v,99.8} S_x$ as outlier.

Modified 3-sigma method: calculated the z variable of each data point by

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

The probability that x lies outside range defined by -∞ and z is 1 - $P(z)$. For N data points is $N[1 - P(z)] \leq 0.1$, the data points can be considered as outlier.

Data outlier detection

Example: Consider the data given here for 10 measurements of tire pressure taken with an inexpensive handheld gauge. Compute the statistics of the data set; then test for outlier by using the modified three-sigma test

Known: N = 10

Assume: Each measurement obtained under fixed conditions

Solution:

n_i	x [psi]	$\|z\|$	$P(\|z\|)$	$N(1\text{-}P(\|z\|))$
1	28	0.3	0.6199	3.8010
2	31	1.1	0.8724	1.2764
3	27	0.0	0.5111	4.8893
4	28	0.3	0.6199	3.8010
5	29	0.6	0.7199	2.8005
6	24	0.8	0.7895	2.1051
7	29	0.6	0.7199	2.8005
8	28	0.3	0.6199	3.8010
9	18	2.5	0.9932	0.0677
10	27	0.0	0.5111	4.8893

Statistics at start

N = 10: $\bar{x}$ = 27, S_x = 3.604

After removing the spurious data

N = 9: $\bar{x}$ = 28, S_x = 2.0, $t_{8,95}$ = 2.306

$$\mu = \bar{x} \pm \frac{t_{\nu,P} S_x}{\sqrt{N}} = 28 \pm 1.6 \text{ psi} \quad (95\%)$$

Number of Measurement

The precision interval is two sided about sample mean that true mean to be within. $-t_{v,P}S_x/\sqrt{N}$ to $+t_{v,P}S_x/\sqrt{N}$. Here, we define the one-side precision value d as

$$d = \frac{t_{v,P}S_x}{\sqrt{N}}$$

The required number of measurement is estimated by

$$N \approx \left(\frac{t_{v,P}S_x}{d}\right)^2 \qquad P\%$$

The estimated of sample variance is needed. If we do a preliminary small number of measurements, N_1 for estimate sample variance, S_1. The total number of measurements, N_T will be

$$N_T \approx \left(\frac{t_{N_1-1,P}S_1}{d}\right)^2 \qquad P\%$$

$N_T - N_1$ additional measurements will be required.

Number of Measurement

Example: Determine the number of measurements required to reduce the confidential interval of the mean value of a variable to within 1 unit if the variance of the variable is estimated to be ~64 units.

Known: $P = 95\%$ $d = ½$ $\sigma^2 = 64$ **units**

Assume: $\sigma^2 \approx S_x^2$

Solution: $N = 983$

$$N \approx \left(\frac{t_{v,P} S_x}{d} \right)^2 \qquad 95\%$$

Least-Square Regression Analysis

The regression analysis for a single variable of the form $y = f(x)$ provides an m^{th} order polynomial fit of the data in the form. variable

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

Where y_c is the value of the dependent variable obtained directly from the polynomial equation for a given value of x. For N different values of independent and dependent value included in the analysis, the highest order, m, of the polynomial that can be determined is restricted to $m \leq N - 1$.

The values of the $m + 1$ coefficients $a_0, a_1, \ldots, a_m$ are determined by the least square method. The least-squares technique attempts to minimize the sum of the square of the deviations between the actual data and the polynomial fit

Minimize $\longrightarrow D = \sum_{i=1}^{N} (y_i - y_c)^2$

Least-Square Regression Analysis

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

Now the total differential of *D* is dependent on the *m* + 1 coefficients

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

To minimize the sum of square error, one wants *dD* to be zero. This is accomplished by setting each of the partial derivatives equal to zero

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

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

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

Least-Square Regression Analysis

This yields $m + 1$ equations, which are solved simultaneously to yield the unknown regression coefficients, $a_0, a_1, \ldots, a_m$

Standard deviation based on the deviation of the each data point and the fit by

$$S_{yx} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - y_c)^2}{v}}$$

A measure of the precision

We can state that the curve fit with its precision interval as

$$y_c \pm t_{v,P} S_{yx} \ (P\%)$$

Least-Square Regression Analysis

Linear Polynomials ($y_c = a_0 + a_1 x$)

For linear polynomials a correlation coefficient, *r*

$$r = \sqrt{1 - \frac{S_{yx}^2}{S_y^2}} \qquad \text{where} \qquad S_y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

The correlation coefficient is the measure of the linear association between *x* and *y*

$$S_{a1} = S_{yx}\sqrt{\frac{N}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}}$$

The precision estimate of the slope

$$S_{a0} = S_{yx}\sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}}$$

The precision estimate of the zero

Least-Square Regression Analysis

Example: The following data are suspected to follow a linear relationship. Find an appropriate equation of the first-order form

x [cm]	y [V]
1	1.2
2	1.9
3	3.2
4	4.1
5	5.3

Known: Independent variable, x
Dependent variable, y
N = 5

Assume: Linear relation $y_c = a_0 + a_1 x$

ν = N - (m+1) = 5 - (1+1) = 3

Solution:

x[cm]	y[V]	x^2	y^2	xy
1	1.2	1	1.44	1.2
2	1.9	4	3.61	3.8
3	3.2	9	10.24	9.6
4	4.1	16	16.81	16.4
5	5.3	25	28.09	26.5
Σ	**15**	**15.7**	**55**	**60.19**	**57.5**

$$y_c = 0.02 + 1.04x$$