Research Methods

Robert E Simpson

Singapore University of Technology & Design

robert_simpson@sutd.edu.sg

September 25, 2018

Objectives

You should be able to:

- Look up the standard normal probability values associated with z-statistics
- Calculate probabilities of a random normally distributed measurement existing between two values
- Creat probability plots for different distributions
- Identify distributions that accurately fit the measured data using probability plots
- Understand the origin of propagation of errors equation
- Analyse and calculate the propagation of errors by analytical differentiation
- Calculate the propagation of errors numerically using the perturbation approach

Normal distribution

A random variable x is normally distributed if its probability density function is described by:

$$p_G = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (1)$$

where $\langle x \rangle = \mu$ and $Var(x) = \sigma^2$.

Normal distribution

- Usually histograms of random variables exhibit a normal distribution shape, due to a fundamental effect known as the central limit theorem (CLT).
- When a measurement depends on many other independent random variables, the total error can be shown to have a normal distribution.
- The notation $y = N(\mu, \sigma)$ is often used to explain that y is normally distributed with mean, μ, and variance, σ.

Standard normal distribution

A normal random variable, x, with $\mu = 0$ and $\sigma^2 = 1$ is called a standard normal random vairable

If x is a standard normal random variable, the normal random variable is defined as:

$$Z = \frac{X - \mu}{\sigma} \tag{2}$$

Thus $\langle Z \rangle$=0 and $Var(Z)$=1.

And the normal distribution is standardised as:

$$p(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}} \tag{3}$$

Any normally distributed random variable, x, can be standardised to give a Z-score, which can then be used to look up the P(x) in the standard normal distribution probability table.

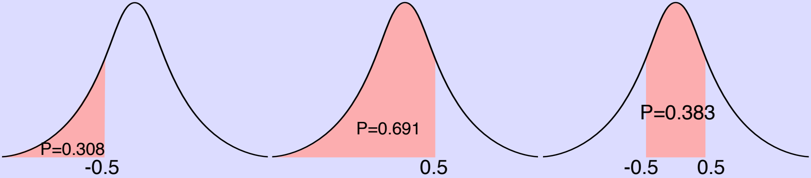
Class Problem 3.1

Normal Distributions

The electrical current flowing through a wire is assumed to follow a normal distribution with a mean of 10 mA and a variance of 4 mA^2. What is the probability that a measurement will be between 9 and 11 mA?

Class Problem 3.1

Normal Distributions



$$P(X < 9) = P(\frac{9-10}{2}) = P(Z < -0.5) = 0.30854$$

$$P(X < 11) = P(\frac{11-10}{2}) = P(Z < 0.5) = 0.69146$$

$$P(-0.5 < Z < 0.5) = 0.69146 - 0.30854 = 0.38292$$

Appropriate PDF model?

Knowing which PDF describes your data is useful because:

- It can verify assumptions
- It can provide insight into the physical mechanism that generates the data[1]

So far we have studied Bionomial, Poisson and Normal PDFs, but how do you know whether a particular PDF can reasonably model your data?

[1] E.g. when studying the reliability of a product, verifying an exponential distribution for the time-to-failure, identifies a failure rate is constant with respect to time

Probability Plots

Probability plotting is a graphical method used to check whether experimental data is accurately modelled by a particular distribution. The procedure is simple and quick. It is more reliable than checking whether the data fits a particular histogram for small to moderate sample sizes. The disadvantages are that determining the data is a good fit is rather subjective, and the result is rather qualitative and not quantitive.

Probability Plotting Procedure

1. Rank the observations in the sample from small to largest: x_j, x_1, x_2, x_3,... x_n. Where x_1 is the smallest observation and x_n is the largest.
2. Look up the standard normalised z-scores, z_j standard normal scores satisfy $\frac{j-0.5}{n}$ (Clearly, other statistics are available for other non-normal distributions).
3. Plot the standard normalised z-scores (z-statistic), z_j, against x_j
4. If the data can be modelled by a normal distribution, the relationship should be linear.

If there is a systematic deviation from the straight line, then the hypothesised model is not appropriate. The same procedure can be used for different distributions to see which has the best fit.

Example 1

Probability plots

10 observations for the life time (minutes) of a well known computer battery are as follows: 176, 191, 214, 220, 205, 192, 201, 190, 183, 185.

Does a normal distribution accurately model the observations?

Probability Plots

Order the observations and look up the z-scores

j	x	(j-0.5)/10	Z
1	176	0.05	-1.645
2	183	0.15	-1.036
3	185	0.25	-0.674
4	190	0.35	-0.385
5	191	0.45	-0.126
6	192	0.55	0.126
7	201	0.65	0.385
8	205	0.75	0.674
9	214	0.85	1.036
10	220	0.95	1.645

Probability Plots

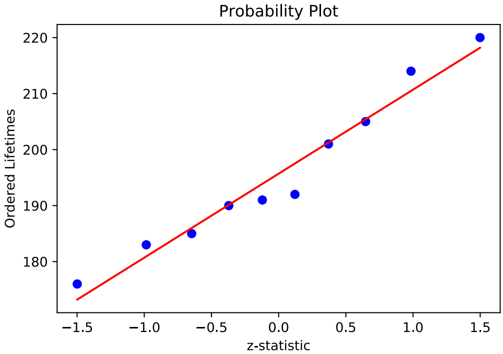
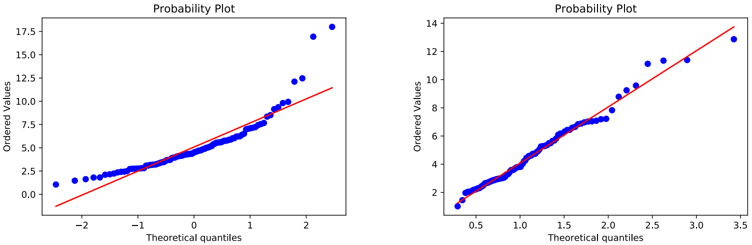


Figure: Normal probability plot

If you are unsure which model is most suited to your data, try different types of continuous distributions, and see which probability plot is most linear.

Different PDF Probability Plots

100 points were generated with a lognormal random variable. The plots below show the line fits where the theoretical statistic is calculated using a (a) normal distribution and (b) a lognormal distribution.



(a) Normal

(b) LogNormal

Clearly the lognormal distribution is preferred to represent this data.

Propagation of Errors (PoE)

- Errors typically occur due to human imprecision or imperfections in equipment
- Usually we quote errors with $\pm x$, where is x is the is the $0.5\times$ the highest decimal place of the measurement
- This number can represent the standard deviation in the precision.
- We expect 68% of the measurements to be $\pm\frac{x}{2}$ of the mean.

Systematic or a statistical errors

Manufactures often state the tolerance of their system. The tolerance can be interpreted in different ways– so be careful! Is it a systematic error or a statistical error?

Digital instruments contain many components each with a tolerance of $\sim 1\%$. Therefore each instrument will have a different systematic error, which can be estimated by measuring a large sample of similar instruments.
Or did the manufacturer calculate the statistical error of a single machine using PoE from the tolerances of each part? These two uncertainties should be the same– otherwise there is probably a problem with the instrument or the measurement procedure.

How can we calculate the statistical error of a single machine from the tolerances of each part?

Propagation of errors (PoE)

繁殖 传播

We cannot always directly measure the variable of interest but we can measure several associated variables from which the variable of interest can be deduced.

If each measurable variable carries an uncertainty, we need to estimate the buildup of uncertainty in the variable of interest.

Class Problem 3.2

Volume of a box error propagation

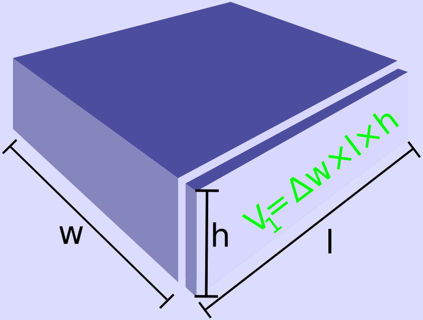
Estimate the error in the volume of a box from measured values of the sides: $w_0 \pm \Delta w$, $h_0 \pm \Delta h$, and $l_0 \pm \Delta l$. Where:

$$\Delta w = w - w_0$$
$$\Delta l = l - l_0$$
$$\Delta h = h - h_0$$

The true value of the sides are: w_0, h_0, l_0.

Class Problem 3.2

Volume of a box error propagation

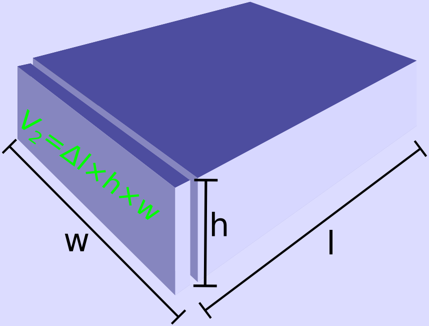


$$V_1 = V_0 + (l \times h)\Delta w$$

$$V_1 = V_0 + \left(\frac{\partial V}{\partial w}\right)_{l_0 h_0} \Delta w$$

Class Problem 3.2

Volume of a box error propagation

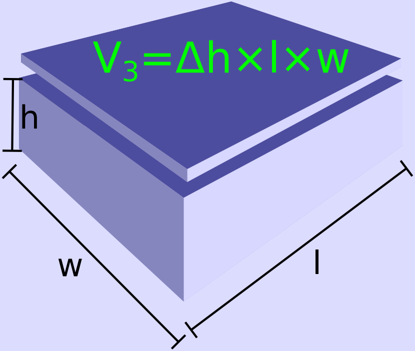


$$V_2 = V_0 + (w \times h)\Delta l$$

$$V_2 = V_0 + \left(\frac{\partial V}{\partial l}\right)_{w_0 h_0} \Delta l$$

Class Problem 3.2

Volume of a box error propagation



$$V_3 = V_0 + (l \times w)\Delta h$$

$$V_3 = V_0 + \left(\frac{\partial V}{\partial h}\right)_{l_0 w_0} \Delta h$$

Class Problem 3.2

Volume of a box error propagation

$$\Delta V = V - V_0 \quad = \Delta w \left(\frac{\partial V}{\partial w}\right)_{l_0 h_0} + \Delta l \left(\frac{\partial V}{\partial l}\right)_{w_0 h_0} + \Delta h \left(\frac{\partial V}{\partial h}\right)_{l_0 w_0}$$

Where δl, δh, δw is the error in each of the box lengths.

$$\Delta V = \sigma_w \left(\frac{\partial V}{\partial w}\right)_{l_0 h_0} + \sigma_l \left(\frac{\partial V}{\partial l}\right)_{w_0 h_0} + \sigma_h \left(\frac{\partial V}{\partial h}\right)_{l_0 w_0}$$

Derivation of the Error Propagation Equation

$$x = f(u, v)$$ x depends on two variables

$$x_i = f(u_i, v_i)$$

$$x_i = x_0 + u_i \frac{\partial x}{\partial u} + v_i \frac{\partial x}{\partial v}$$ 1st order Taylor Expansion

$$\bar{x} = f(\bar{u}, \bar{v})$$ $\bar{x}$ depends on $\bar{u}$ & $\bar{v}$

$$\bar{x} = x_0 + \bar{u} \frac{\partial x}{\partial u} + \bar{v} \frac{\partial x}{\partial v}$$ 1st order Taylor Expansion

$$x_i - \bar{x} = (u_i - \bar{u}) \frac{\partial x}{\partial u} + (v_i - \bar{v}) \frac{\partial x}{\partial v}$$ Deviation of x_i

$x_i - \bar{x}$ is the deviation in x due to single measurements of u & v. This is similar to the previous box example, where the error in volume can be found if we know the errors in the side lengths.

Derivation of the Error Propagation Equation

Rather than looking at the deviation of a single measurement we should find how the variance of the independent variables influences the variance in the resultant distribution of x.

$$\sigma^2 = \lim_{N\to\infty} \frac{1}{N} \sum (x_i - \bar{x})^2 \qquad \text{Variance}$$

$$\sigma^2 = \lim_{N\to\infty} \frac{1}{N} \sum \left[(u_i - \bar{u}) \frac{\partial x}{\partial u} + (v_i - \bar{v}) \frac{\partial x}{\partial v} \right]^2$$

Derivation of the Error Propagation Equation

$$\sigma^2 = \lim_{N\to\infty} \frac{1}{N} \sum \left[(u_i - \bar{u})^2 \left(\frac{\partial x}{\partial u}\right)^2 + (v_i - \bar{v})^2 \left(\frac{\partial x}{\partial v}\right)^2 + 2(u_i - \bar{u})(v_i - \bar{v}) \frac{\partial x}{\partial u}\frac{\partial x}{\partial v} \right] \tag{4}$$

Note that:

$$\sigma_u^2 = \lim_{N\to\infty} \frac{1}{N} \sum (u_i - \bar{u})^2$$

$$\sigma_v^2 = \lim_{N\to\infty} \frac{1}{N} \sum (v_i - \bar{v})^2$$

and if there is no correlation between u & v we get:

$$\boxed{\sigma_x^2 = \sigma_u^2 \left(\frac{\partial x}{\partial u}\right)^2 + \sigma_v^2 \left(\frac{\partial x}{\partial v}\right)^2} \tag{5}$$

Derivation of the Error Propagation Equation

If u & v are correlated then we define the co-variance term:

$$\sigma_{uv}^2 = \lim_{N\to\infty}\left[\frac{1}{N}\sum[(u_i-\bar{u})(v_i-\bar{v})]\right]$$

and the error propagation equation becomes:

$$\boxed{\sigma_x^2 = \sigma_u^2(\frac{\partial x}{\partial u})^2 + \sigma_v^2(\frac{\partial x}{\partial v})^2 + 2\sigma_{uv}^2\frac{\partial x}{\partial u}\frac{\partial x}{\partial v}} \tag{6}$$

The uncertainty in the result depends on the summed square of the uncertainties in the independent variables.

Case Problem 3.3

Angles

Calculate the error in the time for a boat to move 30° around a curve of radius 10 m, moving at a speed of 0.6 ms^{-1}. Assume that the captain is able to estimate angle with an error of 5°, radius with an error of 0.1 m, and velocity with an error of 0.1 ms^{-1}.

Case problem 3 solution

$$\begin{aligned} t = \frac{r\theta}{v} =& \frac{10 \times 30\pi/180}{0.6} \\ =& 8.73\ s \end{aligned} \tag{7}$$

$$\begin{aligned} \sigma_t^2 =& \sigma_r^2 \left(\frac{\theta}{v}\right)^2 + \sigma_\theta^2 \left(\frac{r}{v}\right)^2 + \sigma_v^2 \left(-\frac{r\theta}{v^2}\right)^2 \\ \sigma_t^2 =& 0.1^2 \left(\frac{30\pi/180}{0.6}\right)^2 + \left(\frac{5\pi}{180}\right)^2 \left(\frac{10}{0.6}\right)^2 + (0.1)^2 \left(-\frac{10 \times 30\pi/180}{0.1^2}\right)^2 \\ \sigma_t^2 =& 7.61 \times 10^{-3} + 2.115 + 2.115 \\ \sigma_t =& 2.06\ s \end{aligned} \tag{8}$$

The boat will take 8.7 ± 2.1 s.

Case problem 3.4

Reynold's Number

A liquid is flowing in a pipe of diameter, d. The flow rate (V), viscosity (μ), and density (ρ) of the liquid is measured within some tolerance.

(a) Determine the *relative error* in the Reynolds number (Re) under low and high flow conditions.

(b) How can the measurement be improved?

The Reynold's number is given by:

$$Re = \frac{\rho V d}{\mu}$$

Quantity	Min Flow	Max Flow	Random Error
V (ms^{-1})	1	20	0.1
d (m)	0.2	0.2	0
ρ ($kg\ m^{-2}$)	1E3	1E3	1
μ ($kg\ m^{-1}s^{-1}$)	1E-3	1E-3	5E-6

Table: Error table of four quantities that define the Reynold's Number

Class Problem 3.4

Reynold's Number

$$\partial R/\partial V = \rho d/\mu \qquad \partial R/\partial \rho = dV/\mu \qquad \partial R/\partial \mu = -\rho V d/\mu^2$$

Calculate the **relative** POE:

$$\frac{\sigma^2}{Re^2} = \frac{(\frac{\rho d}{\mu})^2\sigma_V^2 + (\frac{dV}{\mu})^2\sigma_\rho^2 + (\frac{-\rho V d}{\mu^2})^2\sigma_\mu^2}{(\frac{\rho V d}{\mu})^2} = \frac{\sigma_V^2}{V^2} + \frac{\sigma_\rho^2}{\rho^2} + \frac{\sigma_\mu^2}{\mu^2}$$

$$\frac{\sigma}{Re} = \sqrt{\frac{\sigma_V^2}{V^2} + \frac{\sigma_\rho^2}{\rho^2} + \frac{\sigma_\mu^2}{\mu^2}}$$

High Flow

$$\frac{\sigma}{Re} = \sqrt{2.5^{-5} + 1 \times 10^{-6} + 2.5 \times 10^{-5}} = 7.14 \times 10^{-3}$$

Low Flow

$$\frac{\sigma}{Re} = \sqrt{1 \times 10^{-2} + 1 \times 10^{-6} + 2.5 \times 10^{-5}} = 0.1$$

We see that at low flow rate the velocity measurement error is 1000 times more prominent than errors in viscosity and density.

Class Problem 3.5

Exponential growth uncertainty

The consumption of resources is modelled as:

$$Q(t) = \int_0^t P_0 e^{rt} dt \tag{9}$$

Where P_0 is the initial consumption rate, and r is the exponential rate of growth. The world coal consumption in 1986 was equal to 5.0 billion tonnes per a year and the estimated recoverable reserves of coal were estimated at 1000 billion tonnes.

(a) If the growth rate is 2.7 % per a year, how many years before the coal reserves are depleted?

(b) Assume that the growth rate, r, and the recoverable reserves, Q, are subject to random uncertainty with $\sigma_r = 0.2$ % (absolute) and $\sigma_Q = 10\%$ (relative) respectively. Compute the standard error in the estimated time before the coal reserves are depleted, σ_t.

Class Problem 3.5

Exponential growth uncertainty

This PoE problem can be solved analytically,

$$t = \ln\left(\frac{Qr}{P_0} + 1\right) r^{-1}$$

P_0= 5 billion tonnes/year
$r = 0.027$ and $\sigma_r = 0.002$
$Q = 1000$ billion tonnes and $\sigma_Q = 100$ billion tonnes

Perturbation Approach

This numerical method is based on approximating the partial derivatives by a central finite-difference approach. If $y = f(x_i, x_2, ..., x_n)$, then:

$$\begin{aligned} \frac{\partial y}{\partial x_1} &= \frac{y(x_1 + \Delta x_1, x_2, ...) - y(x_1 - \Delta x_1, x_2, ...)}{2\Delta X_1} \\ \frac{\partial y}{\partial x_2} &= \frac{y(x_1, x_2 + \Delta x_2, ...) - y(x_1, x_2 - \Delta x_2, ...)}{2\Delta X_2} \end{aligned} \tag{10}$$

Typically a computer is used to watch how $\frac{\partial y}{\partial x_1}$ and $\frac{\partial y}{\partial x_2}$ converge with Δx_1 and Δx_2. The converged values are then used in the PoE equation to numerically estimate to combined error of the system. This is useful for computing PoE for equations that cannot easily be differentiated. We will practice this approach in a future problems class.

Summary

- Learnt how to graphically/visually analyse how well a particular distribution fits the data
- Learnt how to analytically analyse how errors combine to give an overall error for a system
- Learnt how to numerically analyse how errors combine to give an overall error for a system