Statistics & Probability

What is error?

Two kinds: Random & Systematic

Random error is the scatter you get from repeated measurements

Systematic error is systematic: it happens to the same degree each time you make a measurement.

Multiple measurements will average out (reduce) random error, but won't affect systematic error!

If you do many measurements, the error in the avg. will be completely dominated by systematic error, but this is very difficult to estimate!

Be wary of error estimates – always look to see how it is calculated.

As an engineer, a detailed understanding of statistics and error is <u>critical</u>; You <u>must</u> master this material!

OK, now for some elementary statistics:

The most important <u>probability distribution</u> is the Gaussian or normal distribution:

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

$p(x) \equiv$ probability density

$$= \frac{\text{probability of meas. being in interval } [x, x+dx]}{dx}$$

We can define the cumulative probability:

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

This is the probability that a meas. is less than some value x

The normal distribution is characterized by 2 quantities:

$\mu \equiv$ mean of distribution

$\sigma \equiv$ population standard deviation

About 69% of observations lie within 1σ of μ, 95% within 2σ, and 99% within 3σ.

In working w/ statistics, we define an expectation value

$E(x) \equiv$ what you expect to get if you do a meas. many times and average it together!

Mathematically:

$$E(x) \equiv \int_{-\infty}^{\infty} x \; p(x)\, dx$$

prob. density

meas. quantity

$$= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx$$

$$= \int_{-\infty}^{\infty} \frac{x-\mu}{\sqrt{2\pi}\,\sigma^2}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx + \mu \int_{-\infty}^{\infty} p(x)\, dx$$

Let $z \equiv x - \mu$

$$\therefore E(x) = \int_{-\infty}^{\infty} \underbrace{\frac{z}{\sqrt{2\pi}\,\sigma^2}}_{\text{odd}}\, \underbrace{e^{-\frac{z^2}{2\sigma^2}}}_{\text{even}}\, dz + \mu \underbrace{\int_{-\infty}^{\infty} p(x)\, dx}_{=1}$$

The integral of an <u>odd</u> function over an <u>even</u> domain is <u>zero</u>, so first integral vanishes!

Thus $E(x) = \mu$

What about the mean of N observations?

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

$$\therefore E(\bar{x}) = E\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N} E(x_i) = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu!$$

This may have been self-evident, but we can use the same approach for the variance:

What is $E((x-\mu)^2)$?

$$E((x-\mu)^2) \equiv \int_{-\infty}^{\infty} (x-\mu)^2 \, p(x)\, dx$$

$$= \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Let $z = (x-\mu)$

$$\therefore \quad E\left((x-\mu)^2\right) = \int_{-\infty}^{\infty} \frac{z^2}{\sqrt{2\pi}\,\sigma} e^{-\frac{z^2}{2\sigma^2}} dz$$

Integrate by parts:

$$= \underbrace{-\frac{\sigma^2 z}{\sqrt{2\pi}\,\sigma} e^{-\frac{z^2}{2\sigma^2}} \Bigg|_{-\infty}^{\infty}}_{\text{vanishes at } \pm\infty} + \sigma^2 \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{z^2}{2\sigma^2}} dz}_{=1}$$

$$\therefore \quad E\left((x-\mu)^2\right) \equiv \sigma^2$$

What about the variance of the mean of N measurements?

$$E\left((\bar{x}-\mu)^2\right) \equiv \sigma_{\bar{x}}^2$$

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

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

$$= \frac{1}{N^2} E\left(\sum_{i=1}^{N} (x_i-\mu)^2 + \sum_{i=1}^{N} \sum_{j \neq i} (x_i-\mu)(x_j-\mu)\right)$$

$$= \frac{1}{N} E\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2\right)$$

$$+ E\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}(x_i-\mu)(x_j-\mu)\right)$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}E\left((x_i-\mu)^2\right)$$

$$+ \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}E\left((x_i-\mu)(x_j-\mu)\right)$$

First expectation is just σ^2

Second term is zero provided x_i & x_j are independent

$$E\left((x_i-\mu)(x_j-\mu)\right) \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i-\mu)(x_j-\mu) \cdot p(x_i)\,p(x_j)\,dx_i\,dx_j$$

$$\equiv \int_{-\infty}^{\infty}(x_i-\mu)\,p(x_i)\,dx_i \int_{-\infty}^{\infty}(x_j-\mu)\,p(x_j) = 0 \quad i\neq j$$

Thus:

$$\sigma_{\bar{x}}^2 \equiv \frac{\sigma^2}{N}$$

That is why multiple measurements reduce the random error!

Ok, how do we estimate μ and σ?

In general we have a sample of N observations drawn from an underlying population of possible observations!

we define the sample mean

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

we have already shown that

$$E(\bar{x}) = \mu$$

That is, $\bar{x}$ is an unbiased estimate of μ.

Now for the variance:

We <u>define</u> the sample variance:

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

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

$$= \frac{1}{N-1} \left\{ \sum_{i=1}^{N} x_i^2 - 2\bar{x} \underbrace{\sum_{i=1}^{N} x_i}_{\equiv N\bar{x}} + N\bar{x}^2 \right\}$$

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

Note that to calculate this we <u>don't</u> have to store all the x_i's!

We just need to keep N, $\sum_{i=1}^{N} x_i$ and $\sum_{i=1}^{N} x_i^2$ $\Rightarrow$ that's how a calculator with only a few memories can calc. S_x^2.

85

Ok, how is s_x^2 related to σ^2?

We need the following identities:

$$\sum_{i=1}^{N} (x_i-\mu)^2 = \sum_{i=1}^{N} (x_i-\bar{x})^2 + N(\bar{x}-\mu)^2$$

(this is because $\sum_{i=1}^{N} (x_i-\bar{x}) = 0$ by the definition of $\bar{x}$)

Also, we had:

$$E\left((\bar{x}-\mu)^2\right) \equiv \sigma_{\bar{x}}^2 = \frac{\sigma^2}{N}$$

and $E\left((x_i-\mu)^2\right) = \sigma^2$

Thus:

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

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

86

$$= \frac{1}{N-1}\left\{N\sigma^2 - \frac{N}{N}\sigma^2\right\} = \sigma^2!$$

So S_x^2 is an unbiased estimate for σ^2!

The $N-1$ part comes about because we are calculating the mean from the same sample that we are calculating the variance

We have reduced the number of degrees of freedom by one.

Propagation of Errors:

Usually when doing experimental calculations you must combine several measurements to get the final result.

Each individual meas. has some error associated with it. How do these combine?

Let x_i, y_i be random variables w/ mean $\bar{x}$, $\bar{y}$ and st. dev. s_x & s_y

Let

$$z_i = c_1 x_i + c_2 y_i$$

where c_1 & c_2 are csts.

$$\therefore \bar{z} = \frac{1}{N} \sum_{i=1}^{N} z_i = c_1 \bar{x} + c_2 \bar{y}$$

What about the variance?

$$s_z^2 \equiv \frac{1}{N-1} \sum_{i=1}^{N} (z_i - \bar{z})^2$$

$$= \frac{1}{N-1} \left\{ \sum_{i=1}^{N} (z_i^2 - \bar{z}^2) \right\}$$

Substituting in:

$$z_i^2 = c_1^2 x_i^2 + c_2^2 y_i^2 + 2 c_1 c_2 x_i y_i$$

$$\bar{z}^2 = c_1^2 \bar{x}^2 + c_2^2 \bar{y}^2 + 2 c_1 c_2 \bar{x} \bar{y}$$

$$\therefore S_z^2 \equiv \frac{1}{N-1} \sum_{i=1}^{N} \Big(c_1^2 (x_i^2 - \bar{x}^2) + c_2^2 (y_i^2 - \bar{y}^2)$$
$$+ 2 c_1 c_2 (x_i y_i - \bar{x}\bar{y}) \Big)$$

$$= c_1^2 S_x^2 + c_2^2 S_y^2 + \frac{2 c_1 c_2}{N-1} \sum_{i=1}^{N} (x_i y_i - \bar{x}\bar{y})$$

Let's look at this last term. We define:

$$S_{xy}^2 \equiv \frac{1}{N-1} \sum_{i=1}^{N} (x_i y_i - \bar{x}\bar{y})$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

This is the covariance of x and y. If they are independent, then:

$$E(S_{xy}^2) = 0$$

So for addition:

$$S_z^2 = c_1^2 S_x^2 + c_2^2 S_y^2 + 2 c_1 c_2 S_{xy}^2$$