

Numerical Methods Lecture 7 - Statistics, Probability and Reliability

Topics

- A summary of statistical analysis
- A summary of probability methods
- A summary of reliability analysis concepts

Statistical Analysis

The value of a measured quantity can often vary from one measurement to the next, and from one sample to the next (e.g. student grades on an exam, strength of concrete cylinders). We will refer to such a changing quantity as a 'random variable'. Statistical analysis allows us to view important characteristics of the random variable without having full information. That is, we won't know what the exact strength of the next concrete cylinder to be tested is, but we can take a good guess based on previous measurements and statistical analysis.

Mean and Standard Deviation of a Single Variable

The most fundamental statistics are the mean μ and standard deviation σ .

Given: A single random variable 'X' sampled 'N' times

The **mean** of X - denoted μ_x : average value of the measured quantity

$$\mu_x = E[x] = \frac{1}{N} \sum_{i=1}^N x_i$$

The **standard deviation** - denoted σ_x : the average distance from the mean, or the average spread

$$\sigma_x = \sqrt{\text{VAR}_x} = \sqrt{E[(x - \mu_x)(x - \mu_x)]} = \sqrt{\frac{1}{N} \sum_{i=1}^N ((x_i - \mu_x)^2)}$$

'var_x' is the variance of x. The standard deviation is the square root of the variance.
an equivalent expression is

$$\sigma_x = \sqrt{\left(\frac{1}{N} \sum_{i=1}^N (x_i)^2 \right) - \mu_x^2}$$

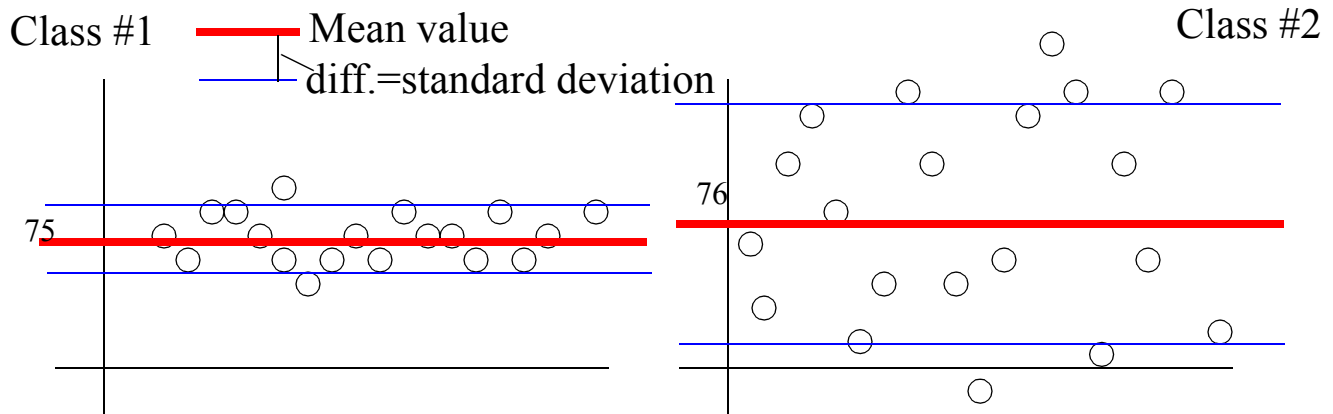
A higher standard deviation increases the odds of being far away from the mean.

Example: Two different sets of exam grades

Class #1 and class #2 have about the same mean value (red line)

Class #1 has a small standard deviation: most students are near the mean (blue line borders)

Class #2 has a larger standard deviation, so students have a higher probability of being well over or well under the class average grade.



We can use the mean and standard deviation to estimate the likelihood of deviating from the mean value. Higher σ = higher probability of being further from the mean. We will get into quantifying this probability in a few pages.

The mean and standard deviation are classified as first- and second-order statistics (involving the mean of X , and mean of X^2 , respectively). If we stick with using these two stats to describe data, we are making assumptions about the form of its probability. We assume the fluctuations about the mean are equally likely to be above or below the mean. That is, the probability behavior is **SYMMETRIC** about the mean. This will not always be realistic. For example, if I give an easy test, the class average may be 100, but the standard deviation may be 15. If we assume the distribution of grades is symmetric about the mean, that would result in scores above 100, which is out of bounds. So there are cases when just the mean and standard deviation are not enough.

We can look at **higher-order** statistics to help. That is, look beyond mean and standard deviation to explain non-symmetric data.

Higher-Order Statistics

Higher-order statistics simply extend the idea of mean and standard deviation to higher order.

skewness - third-order statistic: measures the asymmetry about the mean

$$skewness = \frac{E[(x - \mu_x)^3]}{\sigma_x^3} = \frac{\frac{1}{N} \sum_{i=1}^N ((x_i - \mu_x)^3)}{\sigma_x^3}$$

If we ignore the third-order statistic (skewness), we are assuming skewness = 0
i.e. fluctuations about the mean are symmetric.

Example: skewed class grades. Let's look at grades again, this time including the boundaries between 0 and 100.

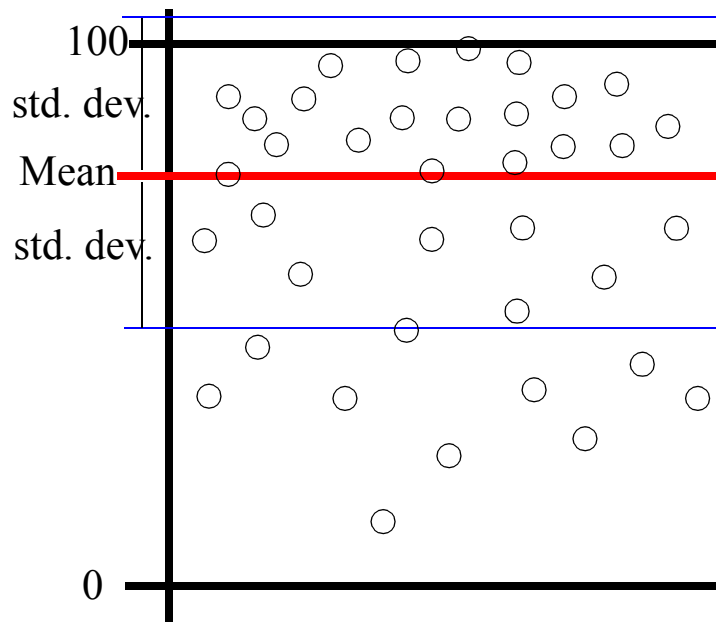
There is a large concentration of grades above and near the mean (red line)

There is a scattering of very low grades. This makes the fluctuations about the mean non-symmetric
====>>

One std. dev. above the mean is actually above the maximum value of 100 (blue line).

One std. dev. below the mean does not encompass the scatter of low grades very well.

Negative / positive skewness: The data is scattered further beneath / above the mean than above / beneath the mean.



The above example is negatively skewed. It would not be appropriate to use just the mean and standard deviation to assign letter grades. We need the skewness to give us a more complete view of grade distribution.

Later we'll see how to fit a curve to these statistics to quantitatively analyze the data.

Statistics of Two Variables

The relationship between two quantities is an often needed characteristic
e.g. relation between the ratio aggregate/water and concrete strength

Given: Two quantities X and Y sampled N times

The Covariance between two random variables X and Y is

$$COV_{xy} = E[(x - \mu_x)(y - \mu_y)] = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y), \quad \text{note no } \sqrt{\quad}$$

Note that if X and Y are the same process, the covariance becomes the variance, where $\sqrt{VAR_x} = \sigma_x$,
and the relationship between variance and standard deviation becomes $COV_{xx} = VAR_{xx} = \sigma_x^2$

Covariance can be used to measure the how much X and Y are related to each other by defining a **correlation coefficient**

Correlation Coefficient - a number that measures the linear relationship between X and Y

$$\rho_{xy} = \frac{COV_{xy}}{\sigma_x \sigma_y}, \quad \text{bounded by } -1 \leq \rho_{xy} \leq 1$$

The boundaries indicate the following property $|COV_{xy}| \leq \sigma_x \sigma_y$

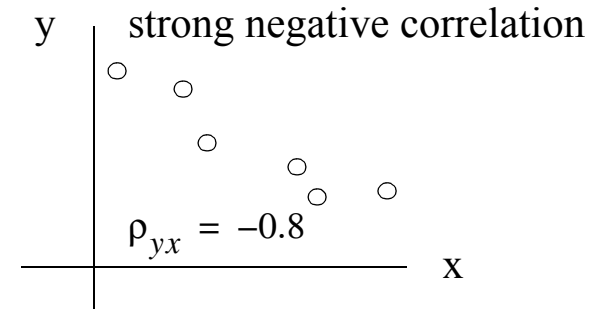
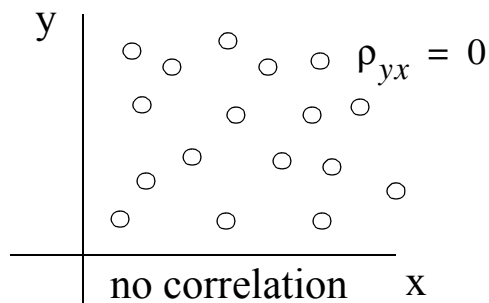
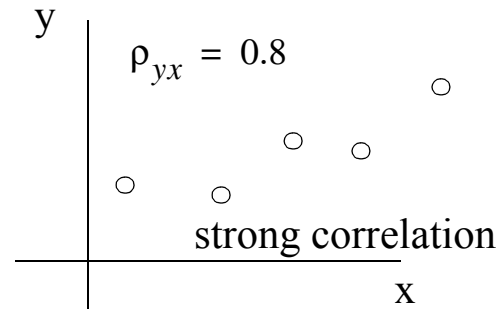
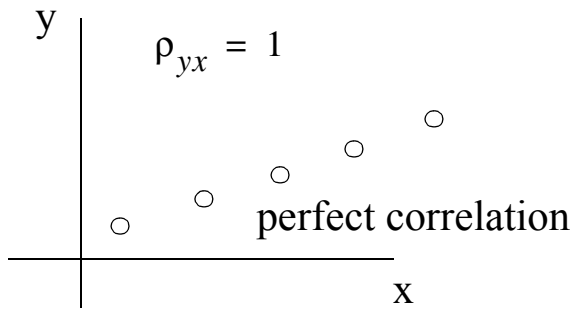
Meaning of the correlation coefficient

$\rho_{xy} = 1$: perfect linear correlation (identical processes)

$\rho_{xy} = 0$: no linear correlation between x and y

$\rho_{xy} = -0.8$: strong negative linear correlation (if x increases, y decreases)

Application - two different random variables x and y measured at the same time



Probability analysis - A formal framework for using statistical descriptions

Quantities provided in common engineering applications often are not exact. An **uncertainty** is often associated. Probability analysis incorporates this uncertainty when providing a solution.

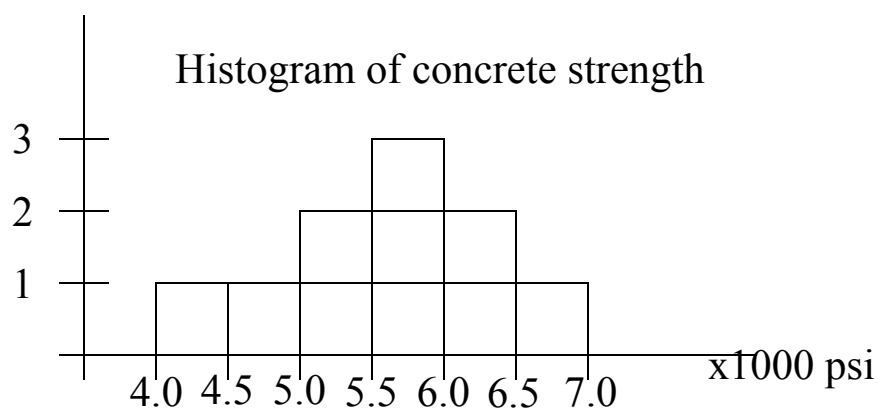
Key Issue: providing a statistical description of the uncertainty

- 1) take many samples
- 2) estimate statistics
- 3) fit a distribution using these statistics

Theme example: strength of concrete cylinders

10 concrete cylinders are tested for compressive strength rated in psi

The number of cylinders that break within a series of 500 psi ranges are plotted below



Now test cylinder # 11. If you had to bet on the range its strength will be in, where would you put your money? That's the basic idea. What is the likelihood of any particular strength?

Normalized Histograms

Let's do some normalizing of the values on the y-axis. Right now, it looks like the probability of concrete strength being between 5.5 and 6.0 ksi is:

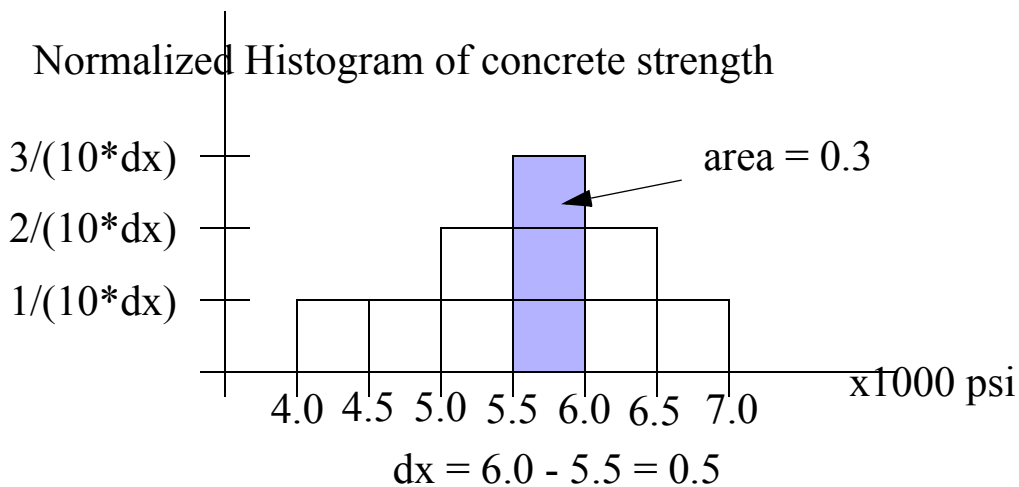
$$\begin{aligned} \text{prob}(5.5 < x < 6.0) &= \# \text{ of samples in range} / \text{total } \# \text{ of samples} \\ &= 3/10 = 0.3 = 30\% \end{aligned}$$

Let's manipulate the y-axis so that the area under the histogram between 5.5 and 6.0 is 0.3 (30%)

i.e. we want to alter the y-axis so that

$$\int_{5.5}^{6.0} \text{y-axis } dx = 0.3$$

To get there: $\text{y-axis} = \# \text{ of samples in bin} / (\text{total } \# \text{ of samples} * \text{distances between bins})$



Now the **blue area** is equal to the probability of being within the range [5.5 6.0] ksi

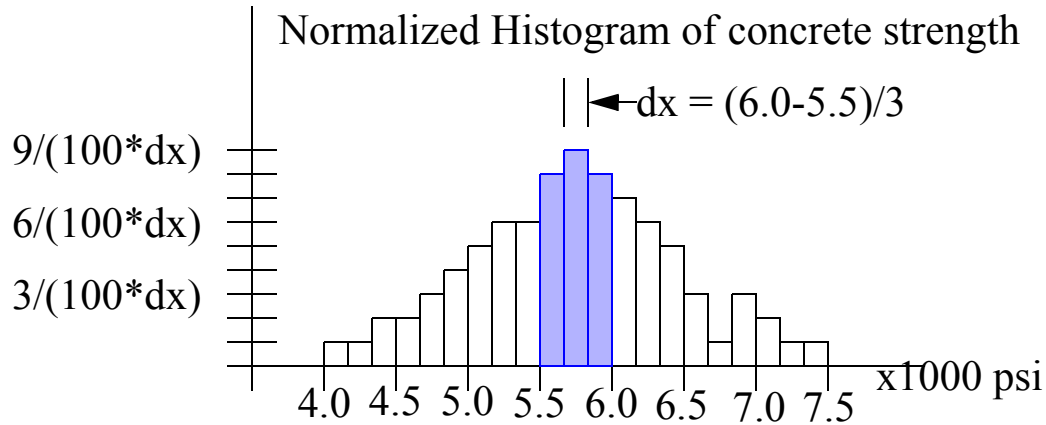
This display is called the normalized histogram. The *total area* under the normalized histogram is = 1.0

Next we'll extend the concept of discrete bins of probability to a continuous display.

Probability Density Function

How can we get a more accurate picture of the probability?

If more examples are available, say 100, then the histogram can use smaller bins, providing:



What if we could take 1000 samples, 10,000 samples? As the number of samples go to infinity, the size of the bins approach zero, and the histogram becomes the probability density function (PDF)

probability density function (PDF) - p(x)

The area under p(x) represents the probability of the variable x occurring between the integration limits.

$$prob(a < x < b) = \int_a^b p(x) dx$$

again, the total area under the entire curve is unity (= 1)

$$prob(-\infty < x < \infty) = \int_{-\infty}^{\infty} p(x) dx = 1.0 = 100\%$$

Usually we'll have an equation to describe p(x)

Any equation for p(x) **must** satisfy the above restriction, total area = 1

Q: Usually we don't have an infinite # of samples to numerically generate a PDF p(x).

So how do we get an equation for p(x) ???

A: To get this curve $p(x)$, we'll assume a certain form for the probability of the variable x.

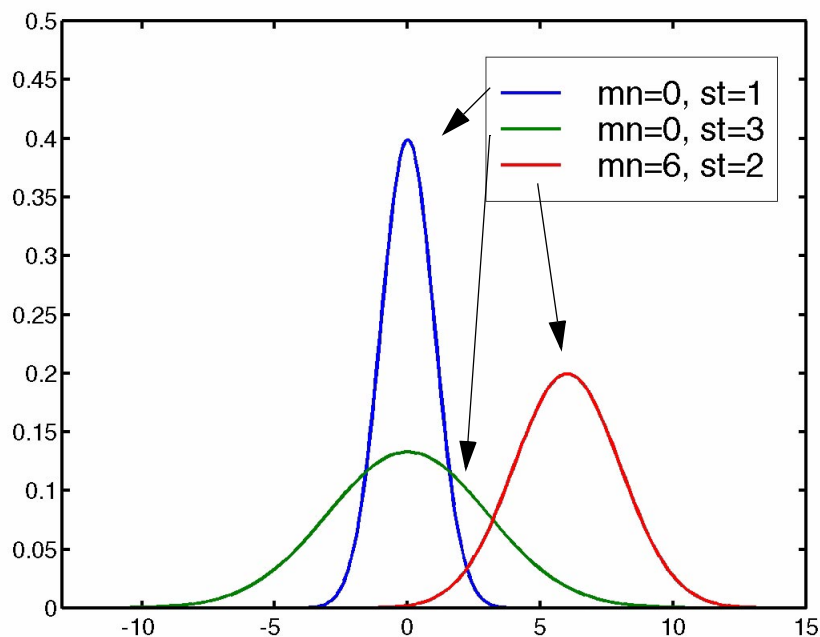
That is, we will curve fit a probability distribution model.

Gaussian distribution

The most common assumed form is the ‘bell curve’, also called the Gaussian distribution. The Gaussian $p(x)$ is completely defined by the mean μ_x and standard deviation σ_x .

Gaussian distribution:
$$p(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x - \mu_x}{\sigma_x} \right)^2 \right]$$

This distribution is symmetric about the mean, and tapers in the tails following an exponential shape. Several examples are given below as mean (μ_x) and standard deviation (σ_x) vary.



Properties of Gaussian distribution (Normal distribution)

- Fully described by μ and σ . Referred to as $N(\mu, \sigma)$
- Third-order statistic skewness = 0, i.e. Symmetry about the mean
- Most commonly used distribution. Many natural random events follow the Gaussian form (e.g. wind speed, rain fall, financial markets...)
- Mathematical limits $-\infty, \infty$
- Practical Limits are plus/minus 5 standard deviations from the mean:
 $[\mu - 5\sigma, \mu + 5\sigma]$. Beyond this range, area under $p(x) \sim 0$
- Linear operations on a Gaussian variable yields a Gaussian result.

$z = 5 + 0.6 * x$ If x is a Gaussian random variable, so is z

Application of PDFs in engineering

Given: 1000 concrete cylinders are tested for ultimate compressive strength.

The 1000 measured sample strengths become samples of a random variable denoted x .

Mean: $\mu_x = 6000 \text{ psi}$

Std. Dev.: $\sigma_x = 500 \text{ psi}$

skewness: $skew = 0.02$

Example #1)

Find: What is the probability that the strength of the next cylinder tested is < 5500 psi?

- skewness close to zero, so we'll assume Gaussian PDF

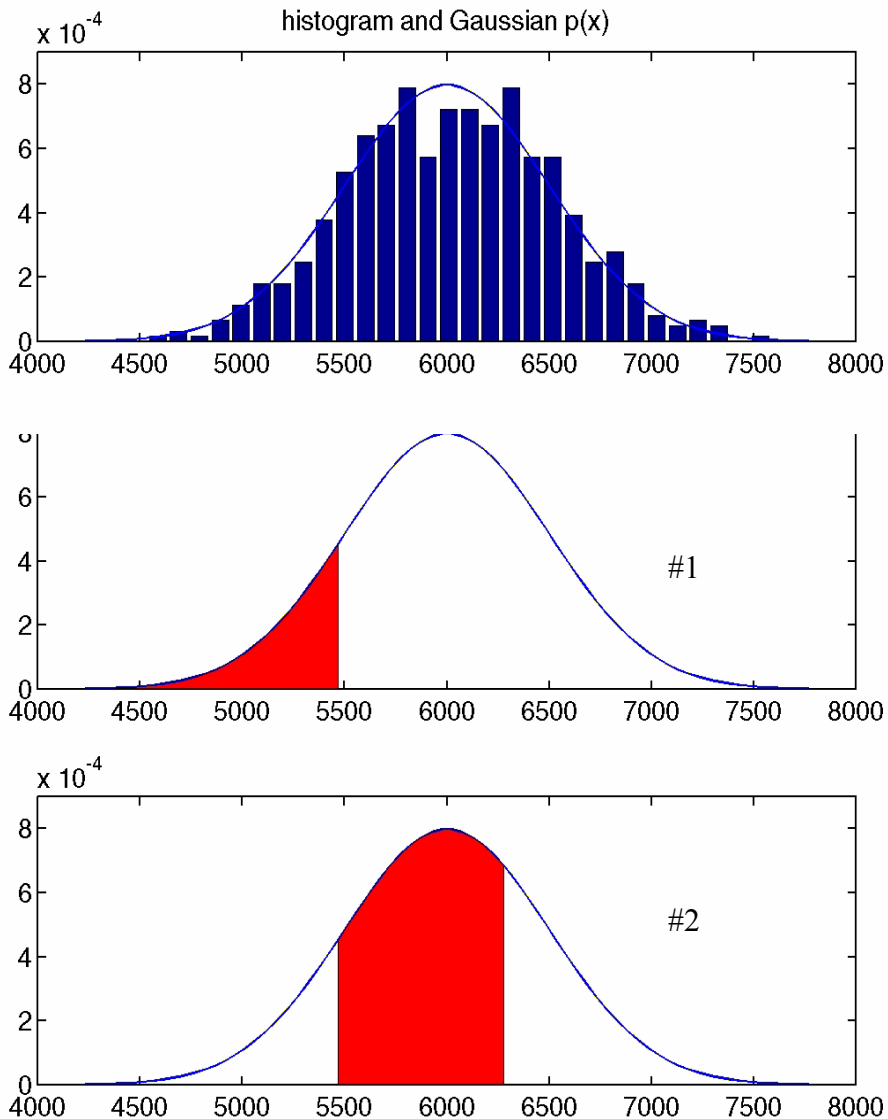
$$prob(x < 5500) = \int_{\mu_x - 5\sigma_x}^{5500} \frac{1}{\sigma_x \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x - \mu_x}{\sigma_x}\right)^2\right) dx$$

Example #2)

Find: What is the probability that the strength of the next cylinder is between 5500 and 6300 psi?

$$prob(5500 < x < 6300) = \int_{5500}^{6300} \frac{1}{\sigma_x \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x - \mu_x}{\sigma_x}\right)^2\right) dx$$

Answer: In this case, we assume a Gaussian distribution fits the data well. The histogram from the measured data and the Gaussian PDF fit are shown in the top figure on the next page. The resemblance is good, so we proceed to use the two equations above to estimate the probability. Graphically, the answer to these two questions is now shown as the red area in the two plots on the next page.



Problem: There is no analytical solution to the integration of the Gaussian function...

Option #1: Use Numerical Integration to get the solution!! We've done this already.

Option #2: Use tables for probability of standard normal Gaussian PDF

standard normal = Gaussian PDF with $\mu = 0$, $\sigma = 1$, or $N(0,1)$.

We can 'normalize' any quantity associated with the Gaussian PDF by removing the mean value and dividing by the standard deviation:

Example: If x is a Gaussian random variable with mean of 5 and standard deviation of 15

$$x = N(5, 15)$$

then we can produce a standard normal variable y with the following operation

$$y = \frac{(x - \mu_x)}{\sigma_x} = N(0, 1) \text{ where } y \text{ is just a manipulation of each of the } x \text{ values}$$

For problem #1,

find: $prob(x < 5500)$ for $N(6000, 500)$

In standard normal space this is equal to:

$$prob\left(x < \frac{(5500 - 6000)}{500}\right) = prob(x < -1) \text{ for } N(0, 1)$$

Tables for $N(0,1)$ say the answer is $prob(x < -1) = 0.1446 = 14.46\%$

There is a built in mathcad function with this table information in there (nice!!)

For problem #2

find: $prob(5500 < x < 6300)$ for $N(6000, 500)$

Look at the red area on the graph on the previous page, this is calculated by:

$prob(x < 6300) - prob(x < 5500)$ for $N(6000, 500)$

In standard normal space this is equal to:

$$prob\left(x < \frac{6300 - 6000}{500}\right) - prob\left(x < \frac{5500 - 6000}{500}\right) \quad \text{for } N(0, 1)$$

Tables for $N(0,1)$ say the answer is $.7106 - .1446 = .5440 = 54.5\%$

Example #3) How do we use these probability models?

Once put into service, the maximum weight placed on a single concrete column is expected to be 200 kips. The concrete column has a surface area of 50 in^2 . The μ and σ of the concrete strength are the same as the previous example. What is the probability that the column will fail under the maximum expected load?

Answer:

$$\mu = 6000 \text{ psi} \times 50 \text{ in}^2 = 300,000 \text{ lbs.}$$

$$\sigma = 500 \text{ psi} \times 50 \text{ in}^2 = 25,000 \text{ lbs.}$$

$$\text{load} = 200,000 \text{ lbs.}$$

$$\text{Find: } prob(\text{strength} < \text{load}) = \int_{(\mu - 5\sigma)}^{200000} \frac{1}{\sigma_x \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x - \mu_x}{\sigma_x}\right)^2\right) dx$$

from $N(300000, 25000)$ space to $N(0,1)$ space, the equivalent is

$$\text{prob}\left(x < \frac{200000 - 300000}{25000}\right) = \int_{-5}^{\left(\frac{200000 - 300000}{25000}\right)} \frac{1}{\sigma_x \sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx$$

We either evaluate the above integral using Simpson's rule, or do a table look-up as in:

$$\text{prob}(x < -4) = 3.13 \text{ E-5} = \mathbf{0.00313 \% \text{ chance of failure}}$$

Reliability Analysis

The final lecture topic is the application of the basic probability concepts to the evaluation of system reliability. Reliability is a formal numerical way of expressing the likelihood of a system failure. This system may be a window pain in high winds, a tall building during an earthquake, the soil above a water table subjected to pollution, etc. 'System' then is a loose term to describe whatever it is we are analyzing and possibly controlling/changing. The definition of failure changes depending on the system, and on what aspect of the system is being investigated. For example, a tall building may fail if its main support columns crack during an earthquake. But failure may also be defined as too much swaying motion during a windy day. The motion will not make the structure unsafe, but the building is not as functional as it should be (perhaps the rent for the top floor apartments has to be reduced).

Failure is defined by the Limit State Equation:

$$G = R - L$$

where R is the system's resistance, and L is the load on the system.

Failure is the state where load exceeds resistance $L > R$ which gives $G < 0$

The variables R and L may be deterministic, random, or a combination. Random variables require a description of their most likely values, called a Probability Density Function. The task of defining the system reliability is then to find the probability of $G < 0$ given the descriptions of R and L .

We will examine two methods of evaluating system reliability: 1) Analytical, 2) Simulation

Analytical

Under certain circumstances, the probability of $G < 0$ can be determined without numerical methods. For example, if our system variables R and L are both Gaussian, the description for G is also Gaussian. Combining Gaussian variables is as done as follows:

Say we want to combine four Gaussian random variables a, c, d, f like so:

$$z = a + c - f + d$$

The resultant z is also Gaussian with the following properties:

mean:
$$\mu_z = \mu_a + \mu_c - \mu_f + \mu_d$$

standard deviation :
$$\sigma_z = SRSS = \sqrt{\sigma_a^2 + \sigma_c^2 + \sigma_f^2 + \sigma_d^2}$$

note that all σ are added, regardless of the minus sign in the equation, while the mean combination includes the minus sign in front of the f term.

Applying this concept to the limit state equation $G = R - L$, the probability of failure can be determined as the probability that the resultant Gaussian variable G is less than zero:

$$\text{prob(failure)} = \int_{-\infty}^0 p(G) dg .$$

This can be evaluated as we did in the probability section. We are assuming all gaussian variables are uncorrelated for simplicity.

We will do an example or two in class on the board. I've left some room below to take notes

Simulation

Sometimes an analytical solution is not possible (or difficult) because we can't easily find the resultant distribution of G . The nice formulas on the previous page to find μ_g and σ_G are not easily gotten. Further, if the distribution for G is not Gaussian (like we've been using so far), then we need more information beyond μ_g and σ_G to get a suitable description of G .

case 1

We can combine as many different Gaussian variables as we like so long as we are adding and subtracting, and still use the analytical approach. However, often there is a need to combine variables in forms other than adding and subtracting. For example, recall that the tip deflection of a cantilevered beam is

$$\text{deflection} = \frac{PL^3}{3EI}.$$

Suppose our system is such a beam with a random load on its tip. Suppose also that the length, E , and I are random with their own Gaussian distributions. If we define failure as deflection exceeding a limit, the limit state would look like this:

$$G = \text{threshold} - \frac{PL^3}{3EI}, \text{ where threshold is the limit to deflection.}$$

If the deflection exceeds the threshold, then $G < 0$, indicating system failure as we define it. Now we are not combining Gaussian variables through addition / subtraction. The nice formulas on the previous page to find μ_g and σ_G do not apply here. We'll look at how to solve this after case 2.

case 2

The probability of load may differ from Gaussian. This can also be true for the resistance R . If the variables in the limit state vary from a Gaussian distribution, we can count on G also differing from Gaussian. For the simple case of $G = R - L$ where either R or L or both aren't Gaussian, again our nice formulas don't apply.

Some special cases can still be solved analytically, but in general it takes great effort to do this.

Solution - Simulation

In both cases above it can be important to account for this change in the distribution of G from Gaussian. To handle these cases, we employ simulation methods. This is a pure numerical brute force method of simulating many random numbers. Each set of random numbers fits the distribution of one of the variables in the limit state equation. We then run many experiments on the computer by seeing how many times out of, say 1,000,000, our limit state is < 0 . Probability of failure is then

$$\text{prob}(fail) = \frac{\# \text{ failures}}{\text{total} \# \text{ simulations}}.$$

The more experiments we perform, the more reliable our final answer becomes. This is because it takes many random numbers to accurately describe a given distribution. Examples given on the board.