

# Lecture 3

## Gaussian Probability Distribution

### Introduction

- Gaussian probability distribution is perhaps the most used distribution in all of science.
  - ◆ also called “bell shaped curve” or *normal* distribution
- Unlike the binomial and Poisson distribution, the Gaussian is a continuous distribution:

$$P(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

$\mu$  = mean of distribution (also at the same place as mode and median)

$\sigma^2$  = variance of distribution

$y$  is a continuous variable ( $-\infty \leq y \leq \infty$ )

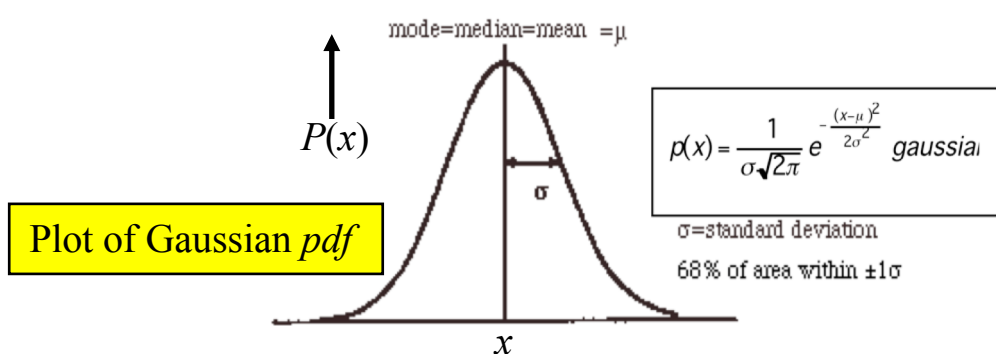
- Probability ( $P$ ) of  $y$  being in the range  $[a, b]$  is given by an integral:

$$P(a < y < b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy$$

- ◆ The integral for arbitrary  $a$  and  $b$  cannot be evaluated analytically
  - can be looked up in a table (e.g. Appendixes A and B of Taylor)
  - or use mathematical software.



Karl Friedrich Gauss 1777-1855



- The total area under the curve is normalized to one.

☞ the probability integral:

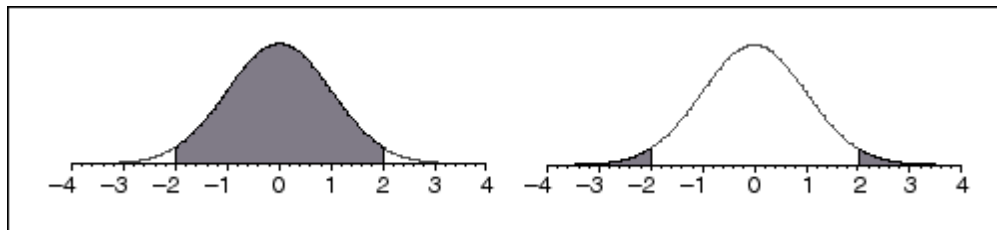
$$P(-\infty < y < \infty) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy = 1$$

- We often talk about a measurement being a certain number of standard deviations ( $\sigma$ ) away from the mean ( $\mu$ ) of the Gaussian.

☞ We can associate a probability for a measurement to be  $|\mu - n\sigma|$  from the mean just by calculating the area **outside** of this region.

$n\sigma$	Prob. of exceeding $\pm n\sigma$
0.67	0.5
1	0.32
2	0.05
3	0.003
4	0.00006

It is very unlikely ( $< 0.3\%$ ) that a measurement taken at random from a Gaussian *pdf* will be more than  $\pm 3\sigma$  from the true mean of the distribution.



95% of area within  $2\sigma$     Only 5% of area outside  $2\sigma$

## Relationship between Gaussian and Binomial distribution

- The Gaussian distribution can be derived from the binomial (or Poisson) assuming:
  - ◆  $p$  is finite
  - ◆  $N$  is very large
  - ◆ we have a continuous variable rather than a discrete variable

- An example illustrating the small difference between the two distributions under the above conditions:
  - ◆ Consider tossing a coin 10,000 time.

- $p(\text{heads}) = 0.5$

- $N = 10,000$

- For a binomial distribution:

- mean number of heads  $= \mu = Np = 5000$

- standard deviation  $\sigma = [Np(1 - p)]^{1/2} = 50$

- ☞ The probability to be within  $\pm 1\sigma$  for this binomial distribution is:

$$P = \sum_{5000-50}^{5000+50} \frac{10^4!}{(10^4-m)!m!} 0.5^m 0.5^{10^4-m} = 0.69$$

- For a Gaussian distribution:

$$P(\mu - \sigma < y < \mu + \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu-\sigma}^{\mu+\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy = 0.68$$

- ☞ Both distributions give about the same probability!

## Central Limit Theorem

- Gaussian distribution is very applicable because of the Central Limit Theorem
- A crude statement of the Central Limit Theorem:
  - ◆ Things that are the result of the addition of lots of small effects tend to become Gaussian.
- A more exact statement:
  - ◆ Let  $Y_1, Y_2, \dots, Y_n$  be an infinite sequence of independent random variables each with the same probability distribution.
  - ◆ Suppose that the mean ( $\mu$ ) and variance ( $\sigma^2$ ) of this distribution are both finite.

Actually, the  $Y$ 's can be from different *pdf*'s!

- ☞ For any numbers  $a$  and  $b$ :

$$\lim_{n \rightarrow \infty} P \left[ a < \frac{Y_1 + Y_2 + \dots + Y_n - n\mu}{\sigma\sqrt{n}} < b \right] = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}y^2} dy$$

- ☞ C.L.T. tells us that under a wide range of circumstances the probability distribution that describes the sum of random variables tends towards a Gaussian distribution as the number of terms in the sum  $\rightarrow \infty$ .

- ☞ Alternatively:

$$\lim_{n \rightarrow \infty} P \left[ a < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < b \right] = \lim_{n \rightarrow \infty} P \left[ a < \frac{\bar{Y} - \mu}{\sigma_m} < b \right] = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}y^2} dy$$

- $\sigma_m$  is sometimes called “the error in the mean” (more on that later).

- For CLT to be valid:

- ◆  $\mu$  and  $\sigma$  of *pdf* must be finite.
- ◆ No one term in sum should dominate the sum.

- A random variable is not the same as a random number.

- ◆ Devore: *Probability and Statistics for Engineering and the Sciences*:

- ☞ A random variable is any rule that associates a number with each outcome in  $S$ 
  - $S$  is the set of possible outcomes.

- Recall if  $y$  is described by a Gaussian *pdf* with  $\mu = 0$  and  $\sigma = 1$  then the probability that  $a < y < b$  is given by:

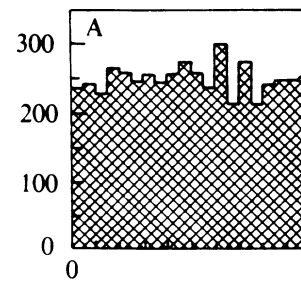
$$P(a < y < b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}y^2} dy$$

- The CLT is true even if the  $Y$ 's are from different *pdf*'s as long as the means and variances are defined for each *pdf*!
  - ◆ See Appendix of Barlow for a proof of the Central Limit Theorem.

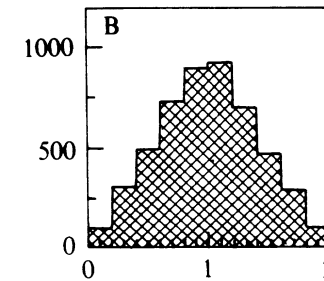
- Example: Generate a Gaussian distribution using random numbers.
  - ◆ Random number generator gives numbers distributed uniformly in the interval [0,1]
    - $\mu = 1/2$  and  $\sigma^2 = 1/12$
  - ◆ Procedure:
    - Take 12 numbers ( $r_i$ ) from your computer's random number generator
    - Add them together
    - Subtract 6
  - ☞ Get a number that looks as if it is from a Gaussian *pdf*!

$$\begin{aligned}
 P \left[ a < \frac{Y_1 + Y_2 + \dots + Y_n - n\mu}{\sigma\sqrt{n}} < b \right] \\
 &= P \left[ a < \frac{\sum_{i=1}^{12} r_i - 12 \cdot \frac{1}{2}}{\sqrt{\frac{1}{12}} \cdot \sqrt{12}} < b \right] \\
 &= P \left[ -6 < \sum_{i=1}^{12} r_i - 6 < 6 \right] \\
 &= \frac{1}{\sqrt{2\pi}} \int_{-6}^6 e^{-\frac{1}{2}y^2} dy
 \end{aligned}$$

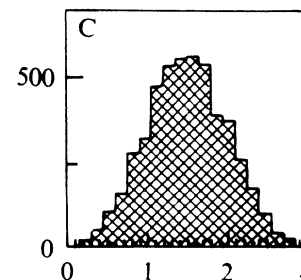
A) 5000 random numbers



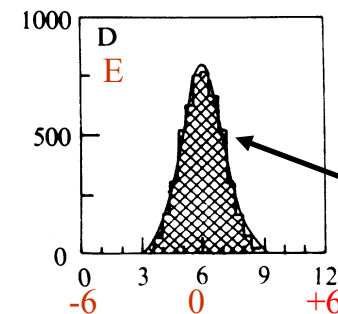
B) 5000 pairs ( $r_1 + r_2$ ) of random numbers



C) 5000 triplets ( $r_1 + r_2 + r_3$ ) of random numbers



D) 5000 12-plets ( $r_1 + r_2 + \dots + r_{12}$ ) of random numbers.



E) 5000 12-plets ( $r_1 + r_2 + \dots + r_{12} - 6$ ) of random numbers.  
 Gaussian  $\mu = 0$  and  $\sigma = 1$

Thus the sum of 12 uniform random numbers minus 6 is distributed as if it came from a Gaussian *pdf* with  $\mu = 0$  and  $\sigma = 1$ .

- Example: A watch makes an error of at most  $\pm 1/2$  minute per day.  
After one year, what's the probability that the watch is accurate to within  $\pm 25$  minutes?
  - ◆ Assume that the daily errors are uniform in  $[-1/2, 1/2]$ .
    - For each day, the average error is zero and the standard deviation  $1/\sqrt{12}$  minutes.
    - The error over the course of a year is just the addition of the daily error.
    - Since the daily errors come from a uniform distribution with a well defined mean and variance
      - ☞ Central Limit Theorem is applicable:
- $$\lim_{n \rightarrow \infty} P \left[ a < \frac{Y_1 + Y_2 + \dots + Y_n - n\mu}{\sigma\sqrt{n}} < b \right] = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}y^2} dy$$
- ☞ The upper limit corresponds to +25 minutes:
 
$$b = \frac{Y_1 + Y_2 + \dots + Y_n - n\mu}{\sigma\sqrt{n}} = \frac{25 - 365 \times 0}{\sqrt{\frac{1}{12}} \cdot \sqrt{365}} = 4.5$$
  - ☞ The lower limit corresponds to -25 minutes:
 
$$a = \frac{Y_1 + Y_2 + \dots + Y_n - n\mu}{\sigma\sqrt{n}} = \frac{-25 - 365 \times 0}{\sqrt{\frac{1}{12}} \cdot \sqrt{365}} = -4.5$$
  - ☞ The probability to be within  $\pm 25$  minutes:
 
$$P = \frac{1}{\sqrt{2\pi}} \int_{-4.5}^{4.5} e^{-\frac{1}{2}y^2} dy = 0.999997 = 1 - 3 \times 10^{-6}$$
  - ☞ less than three in a million chance that the watch will be off by more than 25 minutes in a year!

- Example: The daily income of a "card shark" has a uniform distribution in the interval  $[-\$40, \$50]$ .  
What is the probability that s/he wins more than \$500 in 60 days?

- ◆ Lets use the CLT to estimate this probability:

$$\lim_{n \rightarrow \infty} P \left[ a < \frac{Y_1 + Y_2 + \dots + Y_n - n\mu}{\sigma\sqrt{n}} < b \right] = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}y^2} dy$$

- ◆ The probability distribution of daily income is uniform,  $p(y) = 1$ .

☞ need to be normalized in computing the average daily winning ( $\mu$ ) and its standard deviation ( $\sigma$ ).

$$\mu = \frac{\int_{-40}^{50} yp(y)dy}{\int_{-40}^{50} p(y)dy} = \frac{\frac{1}{2}[50^2 - (-40)^2]}{50 - (-40)} = 5$$

$$\sigma^2 = \frac{\int_{-40}^{50} y^2 p(y)dy}{\int_{-40}^{50} p(y)dy} - \mu^2 = \frac{\frac{1}{3}[50^3 - (-40)^3]}{50 - (-40)} - 25 = 675$$

- ◆ The lower limit of the winning is \$500:

$$a = \frac{Y_1 + Y_2 + \dots + Y_n - n\mu}{\sigma\sqrt{n}} = \frac{500 - 60 \times 5}{\sqrt{675} \cdot \sqrt{60}} = \frac{200}{201} = 1$$

- ◆ The upper limit is the maximum that the shark could win (50\$/day for 60 days):

$$b = \frac{Y_1 + Y_2 + \dots + Y_n - n\mu}{\sigma\sqrt{n}} = \frac{3000 - 60 \times 5}{\sqrt{675} \cdot \sqrt{60}} = \frac{2700}{201} = 134$$

$$P = \frac{1}{\sqrt{2\pi}} \int_1^{134} e^{-\frac{1}{2}y^2} dy = \frac{1}{\sqrt{2\pi}} \int_1^{\infty} e^{-\frac{1}{2}y^2} dy = 0.16$$

☞ 16% chance to win > \$500 in 60 days