

Lecture 3 Gaussian Probability Distribution

1) Introduction

The Gaussian probability distribution is perhaps the most used distribution in all of science. Unlike the binomial and Poisson distribution the Gaussian is a continuous distribution. It is given by:

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

with μ = mean of distribution (also at the same place as mode and median)

σ^2 = variance of distribution

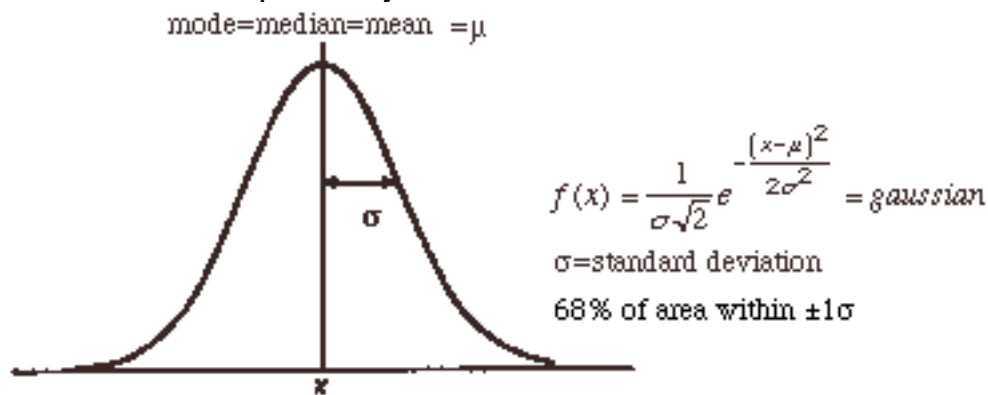
y is a continuous variable ($-\infty \leq y \leq \infty$)

The probability (p) of y being in the range $[a, b]$ is given by an integral:

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

This indefinite integral cannot be done analytically (at least no one's figured out how to do it in the last couple of hundred years) and the values of the integral have to be looked up in a table (e.g. Tables 3.2 and 3.3 of Barlow). The definite integrals of this form can be done analytically and some of the results are given on p.37 of Barlow.

A plot of the Gaussian probability distribution looks like this:



The total area under the curve is normalized to one. In terms of the probability integral we have:

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

Quite often we talk about a measurement being a certain number of standard deviations (σ) away from the mean (μ) 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

Thus it is very unlikely that a measurement taken at random from a Gaussian distribution will lie more than 3σ away from the mean of a Gaussian distribution.

2) 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.
- For more details see Section 3.4.5 of Barlow.

As an example of showing the small difference between the two distributions under the above conditions, consider tossing a coin 10,000 time. Here $p(\text{heads}) = 0.5$ and $N = 10,000$. For a binomial distribution we expect:

- mean number of heads = $\mu = Np = 5000$
- a variance $\sigma^2 = Np(1 - p) = 2500$ or standard deviation = 50.

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

$$P = \sum_{m=5000-50}^{5000+50} \frac{10^4!}{(10^4 - m)!m!} (1/2)^m (1/2)^{10^4 - m} = 0.69$$

How does this compare with a Gaussian distribution? Using the above table we find:

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

So, both distributions give about the same answer in this case!

3) Why is the Gaussian distribution so important?

“Things that are the result of the addition of lots of small effects tend to become Gaussian”

The above is a crude statement of the Central Limit Theorem:

A more exact statement is:

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 (μ) and variance (σ^2) of this distribution are both finite. Then for any numbers a and b :

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

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

Note: σ_m is sometimes called “the error in the mean” (more on that later).

Some things to note about the C.L.T. and the above statements:

a) A random *variable* is not the same as a random *number*! Devore in "Probability and Statistics for Engineering and the Sciences" defines a random variable as (page 81):

"A random variable is any rule that associates a number with each outcome in S".

Here S is the set of possible outcomes.

b) If y is described by a Gaussian distribution with mean (μ) = 0 and variance (σ^2) = 1 then the probability that $a < y < b$ is:

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

c) The C.L.T. is still true even if the Y_i 's are from different probability distributions! All that is required for the C.L.T. to hold is that the distribution(s) have a finite mean(s) and variance(s) and that no one term in the sum dominates the sum.

See Appendix 2 of Barlow for a proof of the Central Limit Theorem.

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 ± 30 minutes?

Let's assume that the daily errors are uniform in $[-1/2, 1/2]$. Then 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 distribution (the uniform distribution) with a well defined mean and variance the Central Limit Theorem is applicable.

Lets use the CLT to estimate the probability:

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

The upper limit corresponds to 30 minutes and hence:

$$a = \frac{Y_1 + Y_2 + \dots + Y_n - n\mu}{\sigma\sqrt{n}} = \frac{30 - 365 \times 0}{\frac{1}{\sqrt{12}} \sqrt{365}} = 5.4$$

The lower limit corresponds to -5.4 . Using the C.L.T. we have:

$$p = \frac{1}{\sqrt{2\pi}} \int_{-5.4}^{5.4} e^{-\frac{1}{2}y^2} dy \approx 1$$

The above integral is $\cong 1$ to about 1 part in 10^6 ! Thus there's less than a one in a million chance that the watch will be off by more than 30 minutes in a year!

Example 2: 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:

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

We need to calculate the average daily winning (μ) and its standard deviation (σ).

First calculated the average:

$$\mu = \frac{\int_{-40}^{50} yp(y)dy}{\int_{-40}^{50} p(y)dy}$$

Since we have a uniform distribution we can set $p(y) = 1$. Note $p(y) = 1$ is the un-normalized probability distribution function.

$$\mu = \frac{\int_{-40}^{50} yp(y)dy}{\int_{-40}^{50} p(y)dy} = 1/2(50^2 - 40^2)/90 = 5$$

To find the standard deviation (σ) we use the definition:

$$\sigma^2 = \frac{\int_{-40}^{50} y^2 p(y)dy}{\int_{-40}^{50} p(y)dy} - \mu^2 = 1/3[50^3 - (-40)^3]/90 - 25 = 675$$

We are now ready to use the CLT!

We are interested in the winnings over 60 days, so $n = 60$.

The lower limit of the winning is given to be \$500, so the lower limit of the integral is:

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

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

$$b = \frac{Y_1 + Y_2 + \dots + Y_n - n\mu}{\sigma\sqrt{n}} = \frac{3000 - 60 \times 5}{\sqrt{675}\sqrt{60}} = \frac{2700}{201} = 13.4$$

So the probability P is given by:

$$P = \frac{1}{\sqrt{2\pi}} \int_1^{13.4} e^{-(1/2)y^2} dy$$

Using a look up table for the integral we find:

$$P = 0.16 \text{ (i.e. a 16\% chance to win } > \$500 \text{ in 60 days)}$$

Note: we can approximate this integral by:

$$P = \frac{1}{\sqrt{2\pi}} \int_1^{\infty} e^{-(1/2)y^2} dy$$

which makes it easier to find in a table of Gaussian integrals.