The Gaussian ("Normal") Distribution

What it is

- For Continuous-type data
  (or data discrete enough to be "continuous")
- (technically) Infinite range $-\infty$ to ∞
- Symmetric "Bell-shaped" distribution
- Described fully by two parameters μ and σ (tabulated)

Shorthand X is N(μ , σ)

How it arises

- "Naturally"

  Biological measurements ...
  e.g. height

- "Manmade"

  Sampling distribution

  - Binomial and Poisson as $\mu = n\pi \rightarrow \infty$
  - Sums (or Means) of Non-Gaussian random variables

  (Central Limit Theorem)

la loi des erreurs

« Tout le monde y [la loi des erreurs] croit cependant, me disait un jour M. Lippmann, car les expérimentateurs s'imaginent que c'est un théorème de mathématiques, et les mathématiciens que c'est un fait expérimental »

“ Everyone believes in it [the law of errors] however, said Monsieur Lippmann to me one day, for the experimenters fancy that it is a theorem in mathematics and the mathematicians that it is an experimental fact. ”

H. Poincaré, Calcul des Probabilités, 2nd Ed. (Paris: Gauthier–Villars, 1912), p. 171. quoted in text "Distribution–Free Statistical Tests" by James V Bradley, Prentice-Hall, 1968

Using the Gaussian Tables

What % of observations would be
(What is prob that single observation would be)

$>$ ___ ? **X -> %**

$<$ ___ ?

$>$ ___ and $<$ ___ ?

What X value(s) will

% -> X

Exclude the lower _____ % of population ?

Exclude the upper _____ % of population ?

Include the middle _____ % of population ?

Take advantage of fact that no matter what the values of μ and σ are, the % of the Gaussian distribution falling between the two values

$\mu + m_1\sigma$ and $\mu + m_2\sigma$

where m_1 and m_2 are any multiples,

will remain the same. Z is a generic or context-free measure of deviation

Use of Standardization

Notes on Diagrams on previous page

Standardization:

The diagram illustrates how to use one Gaussian distribution table for all N(μ,σ) calculations, no matter what the value of μ and σ. Illustrate via e.g. of an IQ score of 130 in relation to a N(100,13) distribution of scores. The two steps are:

change of location from μ = 100 to μ' = 0

change of scale from σ=13 to σ'=1

Combined, they become

$$z = \frac{x - \mu}{\sigma} = \frac{130 - 100}{13} = 2.31 \qquad \text{eqn[1]}$$

The place of 130 on the (100,130) distrn. is the same as the place of z=2.31 on the "Standardized" N(0,1) or "Z" distribution. Percent above X=130 = Percent above Z=2.31 =1.1% [obtained by looking up area of 0.9896 corresponding to z=2.3 in Table A and subtracting from 1 to get 0.0104, and turning it into 1.04%. 130 is at the 98.96th percentile.

Suppose we are asked the reverse question: what is the 75th %-ile of the IQ distribution? In this case, we reverse the sequence of calculations:

Start at a probability of 0.75 in the body of table: it corresponds to a z value of +0.675. Since this z value refers to a N(0,1), distribution, we need to convert it to a score on the IQ scale. So, reversing our steps

0.675 SD's is 0.675 x 13 = 8.8 IQ points

8.8 IQ points above μ(=100) is 100 + 8.8 = 108.8

In algebraic notation, what we have done is calculate

(i) z • SD
(ii) X = μ + z • SD eqn[2]

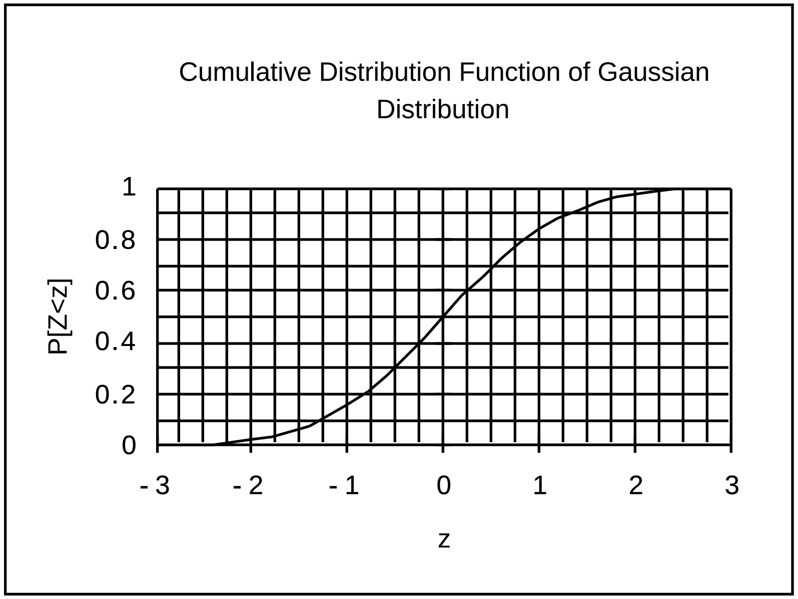
which is the reverse of eqn [1] above, i.e.

$$0.675 = \frac{108.8 - 100}{13}$$

"Spinning Top" that yields N(0,1) observations

Imagine a disk with 2 concentric circles, and a spindle through the centre. Suppose that when spun it is equally likely to come to rest at any point on the circumference. This unbiased-ness is reflected in markings of 0% to 50% uniformly on the circumference of the inner circle. **Q:** How should we mark the circumference of the outer circle so that repeated spins produce values with a Gaussian distribution? [see "spinner" in fig 4.8 page 311 of M&M]
A: Use the z values corresponding to the percentiles!
The spinner shown will produce z values from 0 to infinity.. For values from minus to plus infinity, one could use a separate random determination as to the sign of the z, or else use half the circumference for negative, and half for positive z values.

The **nomograms** on the rightmost column illustrate the same idea, the function that links the shaded area under the Gaussian curve with the corresponding z value. I have shown it two ways, first showing areas (as %'s) as a function of z, and then vice-versa (as is done in Table A of M&M). Actually, Table A tabulates Prob[Z<z] as a function of z., but the idea is the same... because of a shortage of space, and because one can use the symmetry of the z distribution, I used only half the full z scale. The graph below is another way of looking at Table A!



5 Distributions with same Mean and Standard Deviation

Exercises on M&M Ch. 1.3

1 [from Armitage] The iodine level of a tin of salt is stated to be between 433µg and 753 µg. Assuming that the iodine content is a normally distributed random variable and that it lies within the given limits with probability 0.94 and below the lower limit with probability 0.01, find the probability that the iodine content exceeds:

(a) 500 µg (b) 700 µg

2 The following table was derived from random samples of males and females in the "Alberta Study" by Spitzer et al. in mid 1980-s

Variable	n	Mean	SD	Min	Max
males					
Height (cm)	102	176.9	7.1	150	197
Weight (Kg)	101	83.0	15.1	35	125
females					
Height (cm)	107	164.3	6.1	142	182
Weight (Kg)	107	72.3	14.8	49	115

(a) Is height more variable than weight (among males)?

(b) Are women's heights more variable than men's?

(c) Fit a Normal distribution to the observed distribution of weights in males

```
 3|5
 4|6
 5|9
 6|1223455678
 7|01122333444556666777777778999
 8|00000000000122333344444555556 7779
 9|0011111445789
10|11333579
11|278
12|115
```

by calculating how many you would "expect" in each weight category if the weights followed a Normal distribution with mean 83 Kg and standard deviation 15.1 Kg .

ANSWERS 1. Iodine contents

```
0.01                0.94                0.05
-----|--------------------------|-----------
Z
   -2.326                      1.645
     |<-------- 3.971SD’s ---->|
    433                         753
```

So 1 SD = (753 – 433)/3.971 = 80.58µg

µ ?? (753 – µ) / sd = 1.645
==> µ = 753 – 1.645SD = 620.44

500 µg is 120.44 µg below mean or 120.44/80.58 SD’s below mean, i.e., it corresponds to z = –1.49; the tables say that 0.136 is beyond this z (2 directions) so 0.068 is below this z. Then 1 – 0.068 = 0.932 is above –1.49.

Likewise for 700: (700 – 620.4)/80.58 = 0.99 on a 1 tail table is 0.161 (or by 2 tail = 0.322/2).

ANSWERS 2. Data from the Alberta Study

a CV(ht) = 100% x SD/Mean = 100 x 7.1 / 176.9 = 4.0%;
CV(wt) = 100% x SD/Mean = 100 x 15.1 / 83.0 = 18.2%
So height is less variable.

b Likewise CV(women’s height) = 3.7%; less variable
CV(men’s height) = 4.0%; more variable

c Eg of first category (assuming 30-39 means 30 to <40)

30 corresponds to z = (30 – 83)/15.1 = –3.5 (SD’s below)

40 corresponds to z = (40 – 83)/15.1 = –2.65 etc…

so the task reduces to finding what % of the normal distribution lies between –3.5 and –2.65, … then converting the percents to numbers of people out of 101.
The other (faster) way is to start with z boundaries and convert to weight cutoffs: i.e., z = –2, –1.5, –1, –0.5, 0, 0.5, 1, 1.5, 2 …