

Types of Data

	<i>summarized with</i>
Qualitative	
Non-numerical (Nominal/Categorical)	
Binary ("Quantal" or "All-or-None")	
Multicategory	<i>Proportions</i> (and derivatives thereof)
Ordered	
Unordered	
vs.	
Quantitative	
Numerical ("measured")	<i>Measures of</i>
	<i>Location</i>
Discrete	<i>and</i>
	<i>Spread</i>
Continuous	

References

Moore and McCabe -- Chapter 1 page 2

Colton: Ch 2

Armitage and Berry: Ch 1.4

Norman and Streiner: PDQ Stats Chs 1 and 2.

They also distinguish between 'dependent' and 'independent' variables.
Better terms would be 'outcome' and 'stimulus' variables.

Displaying Numerical Data

<u>Type of display</u>	<u>What it shows</u>
Frequency Distribution	Tabular Frequency Distribution of variable
Histogram	Graphical Frequency Distribution of variable
Dot Diagram	Individual values of variable
Stem and Leaf display	Frequency Distribution with detail
Box Plot	Values that split distribution into 1/4ths
Probability Density	Limiting case of Histogram as x 's $\rightarrow 0$
Cumulative Distribution	Graph of <i>Proportion of values $\leq x$</i> vs. x

See also

M&M Ch. 1.1 & 1.2 "Displaying / Describing Distributions".

Colton, Ch. 2, p 21. Histograms, Frequency Polygons & other graphs;

Freedman et al. Chapter 3 ("The Histogram")

Moses LE "Graphical Methods in Statistical Analysis " pp 309-353 in Annual Review of Public Health, Breslow L, Fielding JE, LaveLB (eds), Volume 8, 1987. This is a most readable and very helpful article.

Mosteller, F. Writing about numbers. Chapter 15, pp 305-321 in Medical Uses of Statistics, Bailar, JC and Mosteller, F (Eds) NEJM Books, Waltham, Mass 1986.

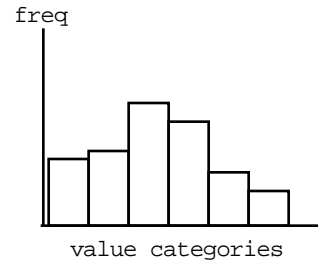
Tufte ER The Visual Display of Quantitative Information, Graphics Press, Cheshire Conn, 1983. A real treat, with lots of historical examples.

Data Displays (1 variable)

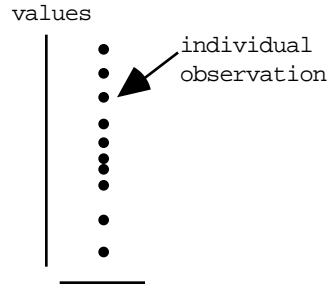
• **Frequency Table**

Interval	Freq	(%)
xx.x - yy.y	xx	xx.x
xx.x - yy.y	xx	xx.x
..
..
xx.x - yy.y	xx	xx.x
xx.x - yy.y	xx	xx.x

• **Histogram**



• **Dot Diagram**



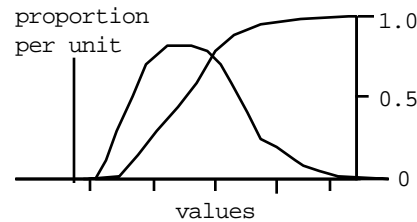
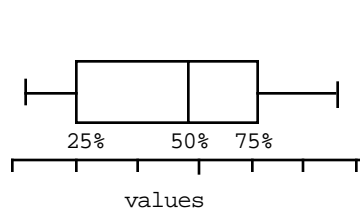
• **Stem and Leaf Plot**

```

4|69
5|36678
6|0003344567
7|01123478
8|0358
9|00
    
```

• **Probability Density & Cumulative Distrn.**

• **Box Plot †**



Notes: Dot Diagram: very visual; not easy if large # obsns.
Freq Table and Histogram: Watch boundaries; AREA proportional to frequency so if unequal intervals, be careful; Cumulative Distribution: useful for reading off percentiles. † See specific software packages for their conventions concerning 'outliers'

Shorthand for referring to Numerical Data

Variables and Subscripts

Variable X with n sample values denoted x_1, x_2, \dots, x_n in order of entry; The "1", "2", ... are called subscripts or indices. We use the letter i (or j) and the range "1 to n" to denote the n different x values and refer to the value of the i^{th} x as " x_i ".

Order Statistics

If the x's are sorted in ascending numerical order, then, $x_{[1]}$ is shorthand for "the smallest x" (the 1st "order statistic"), $x_{[i]}$ is known as the i^{th} "order statistic", and so on.

Summation Σ and Product \prod

The term Σx (spoken: "sigma x" or "sum of x's") is used as a shorthand for the sum

$$x_1 + x_2 + \dots + x_n$$

The Greek capital letter \prod is used as a shorthand for "product of". Thus, $\prod x$ means

$$x_1 \cdot x_2 \cdot \dots \cdot x_n$$

Powers, Logarithms and Anti-logarithms

The term $x^{1/2}$ is shorthand for the square root of x or \sqrt{x} .

Likewise, $x^{1/n}$ denotes the nth root of x or $\sqrt[n]{x}$.

ln (x) denotes the "natural log of x". or "log of x to the base e" i.e. $\log_e(x)$, where e is 2.718.

Note: x must be positive; $\ln(x)$ ranges from - to

$$\ln(0.1) = -2.30; \ln(1) = 0; \ln(2) = 0.69; \ln(10) = 2.30$$

exp(x) is shorthand for e^x or "exponential of x" or the natural anti-log of x. x ranges from - to and $\exp(x)$ yields a positive value. eg. $\exp(-1) = 0.36$; $\exp(0) = 1$; $\exp(.5) = 1.64$; $\exp(1) = 2.71\dots$

Summarizing Numerical Data

Location (Central Tendency)	Individual	Grouped
Arithmetic Mean \bar{x}	$(\sum x) / n$	$\sum f \cdot x_{mid} / \sum f$
Geometric Mean	$(\prod x)^{1/n}$ or if $y = \ln(x)$ $\exp[(\sum y) / n]$	$\sum f \cdot x_{mid} / \sum f$ $\exp[\sum f \cdot y_{mid} / \sum f]$
Median (50 th %-ile)	$(n+1) / 2$ th x	see A&B p30*
Mode	(most popular) the x value with largest f	

Spread (Dispersion / Scatter)

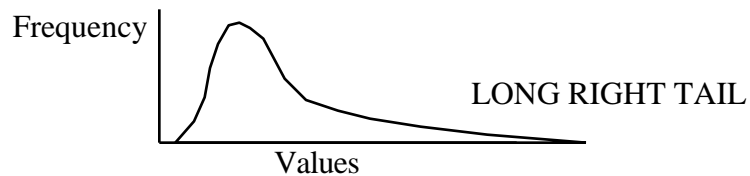
Range	lowest to highest value	
IQR	25%-ile to 75%-ile (Inter- <u>Q</u> uartile <u>R</u> ange)	
Deviations from mean: $d = x - \bar{x}$		
Mean $ d $	$ d / (n - 1)$	$\sum x_{mid} - \bar{x} / (\sum f - 1)$
Mean d^2	$(\sum (x - \bar{x})^2) / (n - 1)$	$\sum f (x_{mid} - \bar{x})^2 / (\sum f - 1)$
Root Mean d^2 or $s_x = SD(x)$	$\sqrt{\text{Mean } d^2}$	$\sqrt{\text{Mean } d^2}$

Relative Spread **CV = 100 • (SD / Mean) %**
(Coefficient of Variation)

Notes

* Armitage and Berry §1.5/1.6. Page 30 shows calculation of median from grouped data.

The term "skewed": Texts don't agree on what is "skewed to right". To avoid confusion, use the terms "long left tail" and "long right tail". Use the latter for the following histogram "distribution"



Arithmetic/Geometric/Harmonic Mean (for +ve numbers)

Example: $x_1 = 1 \quad x_2 = 2 \quad x_3 = 4 \quad x_4 = 8 \quad x_5 = 16$

$$\text{Arithmetic Mean} = \frac{1 + 2 + 4 + 8 + 16}{5} = \frac{31}{5} = \mathbf{6.2}$$

$$\text{Geometric Mean} = (1 \cdot 2 \cdot 4 \cdot 8 \cdot 16)^{1/5} = \mathbf{4}$$

or, since difficult to keep a large product in calculator,

$$e^{\left\{ \frac{\ln[1] + \ln[2] + \ln[4] + \ln[8] + \ln[16]}{5} \right\}}$$

or, if all data are in powers of 2, as here,

$$2^{\left\{ \frac{\log_2[1] + \log_2[2] + \log_2[4] + \log_2[8] + \log_2[16]}{5} \right\}}$$

$$\begin{aligned} \text{Harmonic Mean} &= \frac{1}{\frac{1}{1} + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}} \\ &= \frac{5}{\frac{1}{1} + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}} \\ &= \frac{5}{\frac{31}{16}} \\ &= \mathbf{2.6} \end{aligned}$$

Notes: Harmonic mean = reciprocal of mean of reciprocals
 Geometric mean = inverse log of mean of logs

Harmonic mean Geometric mean Arithmetic mean

The harmonic mean will be useful later when dealing with the variability of the odds ratio from a 2x2 table in relation to cell sizes, and also for the variability of a weighed average

"standard" deviation (écart-type ; typical deviation)

x = Half-life of Caffeine in n=13 healthy non-smokers

SS.	x	x - \bar{x}	\bar{x}	x - \bar{x}	$(x - \bar{x})^2$	\bar{x}	$(x - \bar{x})^2$
1	4.2	-1.75	5.95		3.0765	5.95	
2	3.0	-2.95	5.95		8.7261	5.95	
3	7.6		5.95	1.65		5.95	2.7093
4	3.8	-2.15	5.95		4.6397	5.95	
5	6.4		5.95	0.45		5.95	0.1989
6	7.7		5.95	1.75		5.95	3.0485
7	9.4		5.95	3.45		5.95	11.8749
8	6.7		5.95	0.75		5.95	0.5565
9	4.2	-1.75	5.95		3.0765	5.95	
10	7.7		5.95	1.75		5.95	3.0485
11	7.1		5.95	1.15		5.95	1.3133
12	5.9	-0.05	5.95		0.0029	5.95	
13	3.7	-2.25	5.95		5.0805	5.95	
Σ	77.4	-10.92	+	10.92	24.6023	+	22.7500
				=0			= 47.3523

$\bar{x} = 77.4 / 13 = 5.95$

ave. $(x - \bar{x}) = 0$

ave. $|x - \bar{x}| = \frac{|-10.92| + |10.92|}{12} = 21.84 / 12 = 1.92$

Variance:

$s^2 = 47.3523 / 12 = 3.9460$

Std Deviation

$= \sqrt{\text{Variance}} = \sqrt{3.946} = 1.99$

STANDARD DEVIATION IS CLOSE TO AVERAGE ABSOLUTE DEVIATION ie. deviations without regard to sign.. 1.75, 2.95, ...0.05, 2.25. Average of these absolute deviations is very close to of ave of squared deviations

Why divide by n-1 rather than n to obtain SD?

(See M&M, middle of page 53)

The "wave your hands" explanation: Because with n independent observations, we have n-1 independent evaluations of variation. We have to use one 'degree of freedom' to calculate the sample mean, from which to measure the variation. The n deviations from the sample mean are linked by 1 constraint, namely that they add to zero. If we had only n=1 observation, we would have no opportunity to assess variation, if n=2, then 1 piece of information, etc.

A more theoretical explanation: The above explanation is a bit loose. It doesn't explain why -- for example -- we just don't drop the redundant deviation and work with the average of the squares of the n-1 remaining ones. A fuller understanding comes from recalling that we define s^2 as the average of all the possible $(x - \mu)^2$'s in our universe of x's. We could try to calculate $(x - \mu)^2$ for each of the x's we observe in our sample -- IF we KNEW μ ! Because we don't, we are forced to measure the variation of each x, not from μ , but from the "next best thing" \bar{x} . If our \bar{x} happens to be smaller than μ , the individual x's from which this \bar{x} is calculated are themselves also likely on the (same) lower side of μ . Thus $(x - \bar{x})^2$ tends to be too small.

This can be seen by expanding the numerator of the formula for the variance, and using some algebra to rearrange it:

$$(x - \bar{x})^2 = (x - \mu + \mu - \bar{x})^2 = (x - \mu)^2 - n(\bar{x} - \mu)^2.$$

Thus

$$\frac{(x - \bar{x})^2}{n} = \frac{(x - \mu)^2}{n} - (\bar{x} - \mu)^2.$$

Thus, over all possible samples, i.e. over all possible estimates), the average [in statistical terms, the EXPECTATION] of the first term on the right is indeed σ^2 -- and the average of the second term is $\frac{2}{n}$, so that the average value of the calculable estimate on the left hand side is $\sigma^2 - \frac{2}{n} = \frac{n-1}{n} \sigma^2$, i.e., on average, if we used a divisor of n , we would underestimate σ^2 by a factor of $1/n$.

Dividing by $n-1$ corrects this so that the average of all the possible s^2 's is σ^2 i.e. s^2 's is an unbiased estimator of σ^2

An example with $n=2$ is given next --in a separate exercise, try estimating the variance σ^2 "both ways" in a spreadsheet with samples of $n=2, 3, ..$ observations from a distribution with a known variance σ^2

Why divide by $n-1$ rather than n when using the sample variance as an estimate of variance in a population ???

**Empirical Demonstration of Theoretical reason:
Because doing so makes s^2 an unbiased estimator**

Example

- IF X takes on values 1, 3, 5 with probabilities 1/3, 1/3 & 1/3, then $\mu = 3$

X:	1	3	5	$\mu = \text{average}(X) = 3$
Prob	1/3	1/3	1/3	
X-μ	-2	0	2	
(X-μ)²	4	0	4	Var(X) = ave [(X-μ)²] = 8/3

- Samples {x1,x2} of size $n = 2$...

9 equiprobable samples

		x2		
		1	3	5
x1	1	{1,1}	{1,3}	{1,5}
	3	{3,1}	{3,3}	{3,5}
	5	{5,1}	{5,3}	{5,5}

i.e., 9 equiprobable sample means:

		x2		
		1	3	5
x1	1	1	2	3
	3	2	3	4
	5	3	4	5

e.g. $\bar{x} = (1+5)/2 = 3$

i.e., 9 equiprobable variance estimates

(estimate in [] uses divisor of $n = 2$; estimate outside [] uses $n-1=1$)

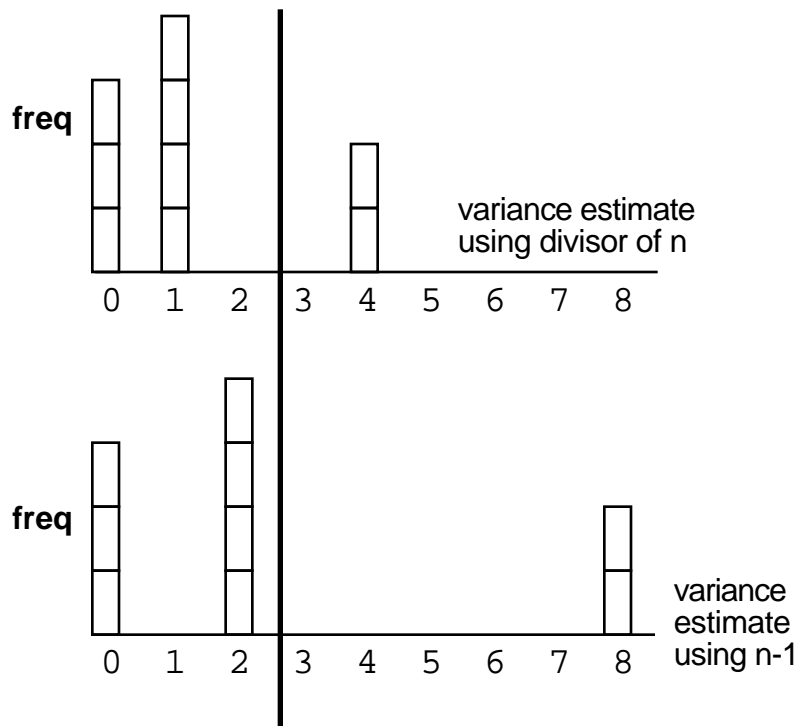
		x2			
		1	3	5	var estimates with n-1 and n
x1	1	0[0]	2[1]	4[8]	[(1-3) ² +(5-3) ²] / 1 = 8
	3	2[1]	0[0]	2[1]	[(1-3) ² +(5-3) ²] / 2 = 4
	5	4[8]	2[1]	0[0]	

(Sampling) distribution of variance estimates:

frequency (out of 9)	variance estimate using divisor of n	variance estimate using divisor of n-1
3	0	0
4	1	2
2	4	8
Average	4/3	8/3

In Pictures...

true variance = 8/3



jh 5/91

Note on usage of " \pm " SD

Dear Editor: ... First, all of the current calcium antagonists have peak and trough effects. It is therefore vital that in studies where potentially additive combinations are being looked at, the timing of the blood pressure measurements following the last dose is carefully controlled and stated in the article. The efficacy of the calcium antagonists either alone or in combination will depend on when the blood pressure was measured. In the article the authors state that the blood pressure was measured 12 ± 2 hours after a dose. Assuming that this is an SD, blood pressure was measured in 99% of patients between 6 and 18 hours after taking a dose. Furthermore... (excerpt from a letter to Editor) *In Reply* ... The measurement of blood pressure at 12 ± 2 hours after receiving a dose was not an SD but a commonly used protocol requirement that BP measurements be made within 12 ± 2 hours after dosing and in fact 100% of the patients had these measurements at 10 to 14 hours after taking a dose, just before the next dose.

Commentary by JH: Although the above objection may be pedantic, it does warn the user (and the reader) to be careful as to the presentation of the standard deviation in reports. It is common to see Mean \pm SD in the description of a set of observations. **The use of the \pm in such situations is incorrect and misleading.**

First, the SD is by definition positive (or at least, as the mathematicians say, "non-negative"). Second, using it this way may tend to give the impression that the data have a symmetric, or possibly even a Gaussian, distribution, and that multiples of the SD can be used to calculate the full pattern of the data. Most data are not symmetric, let alone Gaussian, in their distribution. The use of " \pm " has a lot more justification when we are dealing with Confidence Intervals for population parameters. The reason for this is that CI's are calculated from statistics (aggregates of the observations) the variation of which are more likely, by virtue of the Central Limit Theorem, to be Gaussian in their variation. Even then, one needs to be careful as to whether the margin or error is 1 or 2 or some other multiple. So better to write the mean and SD as mean(SD).

The Average and the Standard Deviation

It is difficult to understand why statisticians commonly limit their enquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull as to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that "IF ITS MOUNTAINS COULD BE THROWN INTO ITS LAKES, TWO NUISANCES WOULD BE GOT RID OF AT ONCE"

Sir Francis Galton 1833-1911 quoted in Freedman's STATISTICS text

ReLocating and ReScaling Numerical Data

see M&M pages 49-51 "Changing Units: Linear Transformation"

Variable X with mean μ_X and Std. Deviation σ_X Change X to Y.
 What will be the mean μ_Y and Std. Deviation σ_Y of Y?

<u>Change</u>	<u>What it does to</u>		
	μ_Y	σ_Y	σ_Y^2
	-----	-----	-----
Add a constant "a"			
i.e. $Y = X + a$	$\mu_X + a$	σ_X	σ_X^2
Multiply by constant "b"			
i.e. $Y = bX$	$b\mu_X$	$b\sigma_X$	$b^2\sigma_X^2$
Add a constant "a" then Multiply by constant "b"			
i.e. $Y = b(X + a)$	$b(\mu_X + a)$	$b\sigma_X$	$b^2\sigma_X^2$
Multiply by constant "b" then Add a constant "a"			
i.e. $Y = bX + a$	$b\mu_X + a$	$b\sigma_X$	$b^2\sigma_X^2$

APPLICATION --- standardized variable
Subtract a constant μ i.e. add $a = -\mu_X$. . . then
Divide by a constant σ_X i.e. multiply by $b=1/\sigma_X$
i.e. $Y = b(X + a)$
$\frac{b(\mu_X + a)}{(1/\sigma_X)(\mu_X + -\mu_X)}$
$\frac{b\sigma_X}{(1/\sigma_X)\sigma_X}$
$\frac{0}{1} \quad \frac{1}{1} \quad \frac{1}{1}$
APPLICATION --- $^{\circ}\text{C} = \frac{5}{9} (^{\circ}\text{F} - 32)$; $^{\circ}\text{F} = 32 + \frac{9}{5} ^{\circ}\text{C}$
APPLICATION --- computational: Use of "working units"... cf A & B p27

The Gaussian ("Normal") Distribution

What it is

- For Continuous-type data (or data discrete enough to be "continuous")
- (technically) Infinite range - 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 $\mu = n$ ->
 - 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

Using Excel instead of Table A

The function NORMDIST(x, μ , σ , TRUE) gives the

How to use one Gaussian distribution table for all $N(\mu, \sigma)$ calculations, no matter what the value of μ and σ .

Standardization

Illustration via e.g. of an IQ score of 130 in relation to a $N(100,13)$ distribution of scores.

Q1: What percent of scores are above 130?

The two steps are:

1. change of location from $\mu = 100$ to $\mu' = 0$
2. change of scale from $\sigma = 13$ to $\sigma' = 1$

Combined, they become

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

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 entering Table A at $z=2.3$, finding lower-tail area of 0.9896, and subtracting it from 1 to get upper-tail area of 0.0104, or 1.04%]

130 is the 98.96th percentile -- 98.96% are below 130.

Q2: 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

$$0.675 \text{ SD's} = 0.675 \times 13 = 8.8 \text{ IQ points}$$

$$8.8 \text{ IQ points above } \mu (=100) \text{ is } 100 + 8.8 = \underline{108.8}$$

In algebraic notation, what we have done is calculate

$$(i) z \cdot \text{SD}$$

$$(ii) X = \mu + z \cdot \text{SD} \quad \text{eqn[2]}$$

which is the reverse of eqn [1] above, i.e. $+0.675 = \frac{108.8 - 100}{13}$

Using Excel functions instead of Table A

The function **NORMDIST(x, μ , σ , TRUE)** gives the cumulative or lower-tail area corresponding to a value of x on a Normal distribution with mean μ and standard deviation σ .

The function **NORMSDIST(z)** gives the cumulative or lower-tail area corresponding to a value of z on a ("Standard") Normal distribution with mean 0 and standard deviation 1.

[Careful: NORMDIST(x, μ , σ , FALSE) gives height of the density curve]

The function **NORMINV(Prob, μ , σ)** gives the reverse (INVerse) i.e. that value of X below which lies 100xProb % of the Normal distribution with mean μ and standard deviation σ .

The function **NORMSINV(Prob)** gives that value of Z below which lies 100xProb % of the ("Standard") Normal distribution with mean 0 and standard deviation 1.