1 Types of Data

Table 1: Types of Data

Type		***summarized with...***
Qualitative		
	Non-numerical (Nominal/Categorial)	
	Binary ("Quantal" or "All-or-None")	*Proportions (and derivatives thereof)*
	Multicategory (Ordered or Unordered)	
vs.		
Quantitative		
	Numerical ("Measured")	*Measures of Location and Spread*
	Discrete	
	Continuous	

References

Moore and McCabe ("M and M") – Chapter 1 page 2

Colton T, *Statistics in Medicine*, Chapter 2

Armitage P and Berry G : Chapter 1.4

Norman G and Streiner D: PDQ Stats Chapters 1 and 2.

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

Table 2: Displaying Numerical Data

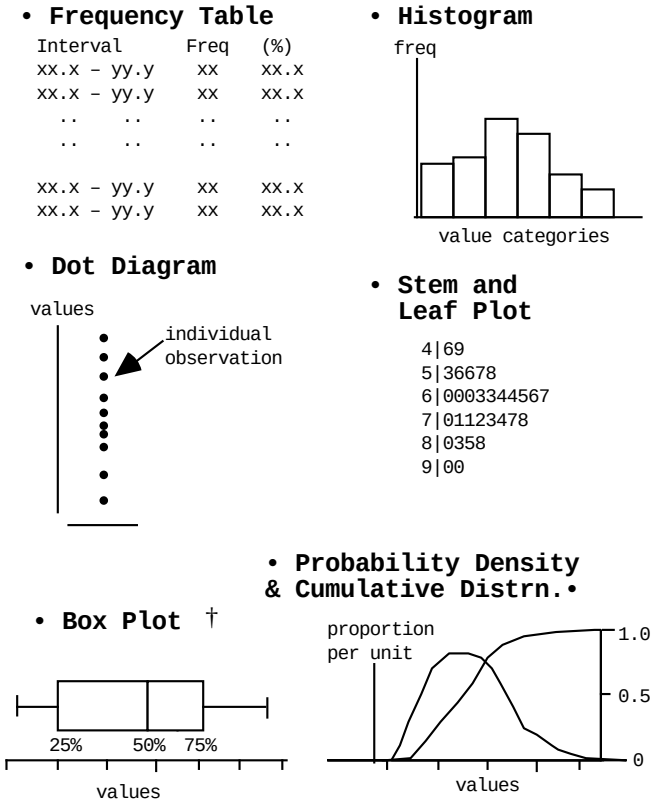
Type of display	***What it shows***
Frequency Distr'n.	Tabular Frequency Distribution of Y
Histogram	Graphical Frequency Distribution of Y
Dot Diagram	Individual values of Y
Stem & Leaf display	Frequency Distribution with detail
Box Plot	Q_2, Q_2, Q_3: distribution split into quarters
Probability Density	Limiting case of Histogram as $\Delta y \to 0$
Cumulative Distribution	Graph of *Proportion of values* $\leq y$ versus y

2 Displaying Numerical Data

See Also

1. M and M, Ch. 1.1 and 1.2: "Displaying/Describing Distributions".
2. Colton, Ch. 2, p. 21. "Histograms, Frequency Polygons and other graphs".
3. Freedman et al. Chapter 3: "The Histogram".
4. Moses LE. "Graphical Methods in Statistical Analysis" pp 309-353 in Annual Review of Public Health, Breslow L, Fielding JE, Lave LB (eds), Volume 8, 1987. *This is a very readable and helpful article.*
5. 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.
6. William S. Cleveland. The Elements of Graphing Data. [book]
7. Tufte ER. "The Visual Display of Quantitative Information". Graphics Press, Cheshire Conn, 1983. *A real treat, a classic with lots of historical examples. see "probably the best statistical graphic ever drawn, the map by Charles Joseph Minard which portrays the losses suffered by Napoleon's army in the Russian campaign of 1812.* `http://www.edwardtufte.com/tufte/posters`.
8. Links to R (and other) graphics:- under 'Resources.'

3 Data Displays (1 variable)



Notes:

Dot Diagram: *very visual: not easy if large number obsns.*

Freq Table and Histogram: Watch boundaries; AREA proportional to frequency so if unequal intervals, be careful;

Cumulative Distribution: useful for reading off percentiles.

Box Plot: See specific software packages for their conventions concerning 'outliers'.

In `R`: `hist/truehist boxplot plot`

4 Statistical Shorthand

4.1 Variables and Subscripts

Variable Y with n sample values denoted $y_1, y_2, \ldots, y_n$ 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 y values, and refer to the value of the i^{th} y *as* "y_i".

4.2 Order Statistics

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

4.3 Summation $\sum$ and Product $\prod$

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

$$y_1 + y_2 + \cdots + y_n.$$

The Greek capital letter $\prod$ (Pi) is used as a shorthand for "*product of*". Thus, $\prod y$ means

$$y_1 \times y_2 \times \cdots \times y_n.$$

4.4 Powers, Logarithms and Anti-logarithms

The term $y^{1/2}$ is shorthand for the square root of y or $\sqrt{y}$. Likewise, $y^{1/n}$ denotes the n^{th} root of y or $\sqrt[n]{y}$.

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

Note: y must be positive; $\ln y$ ranges from $-\infty$ to $+\infty$.

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

$\exp(y)$ is shorthand for e^y or "exponential of y" or the natural anti-log of y. y ranges from $-\infty$ to $+\infty$ and so $\exp(y)$ yields a positive value. eg. $\exp(-1) = 0.36$...; $\exp(0) = 1$; $\exp(0.5) = 1.64$...; $\exp(1) = 2.71$...

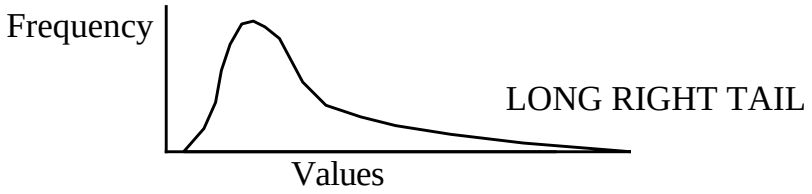
5 Summarizing Numerical Data

See Table 3.

Notes:

**Armitage and Berry*, Chapter 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.



6 Arithmetic/Geometric/Harmonic Mean (for positive numbers)

Example: $y_1 = 1 \quad y_2 = 2 \quad y_3 = 3 \quad y_4 = 4 \quad y_5 = 5$

$ArithmeticMean = (1+2+4+8+16)/5 = 31/5 = 6.2.$

$GeometricMean = (1 \times 2 \times 4 \times 8 \times 16)^{1/5} = 4.$

or, since it is difficult to keep a large product in the calculator,

$$\exp\left[(\ln 1 + \ln 2 + \ln 4 + \ln 8 + \ln 16)/5\right]$$

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

HERE....

$$2^{\{\frac{\log 2[1] \; + \; \log 2[2] \; + \; \log 2[4] \; + \; \log 2[8] \; + \; \log 2[16]}{5}\}}$$

Table 3: Summarizing Numerical Data

Location		
(Central Tendency)		
	Individual	Grouped
Arithmetic Mean $\bar{x}$	$\frac{(\Sigma x)}{n}$	$\frac{\sum f \times x_{mid}}{\Sigma f}$
Geometric Mean	$(\prod x)^{\frac{1}{n}}$	$\frac{\sum f \times x_{mid}}{\Sigma f}$
or if $y = \ln(x)$	$\exp[\frac{(\sum y)}{n}]$	$\exp[\frac{(\sum f \times y_{mid})}{\sum f}]$
Median $(50^{th}\ \% - ile)$	$\frac{(n+1)}{2^{th}x}$	see A and B p30*
Mode	(most popular) the	x value with largest f
Spread		
(Dispersion/Scatter)		
Range	lowest to highest	
IQR	$25\% - ile$ to $75\% - ile$	(Inter-Quartile Range)
Deviations from mean:	$d = x - \bar{x}$	
Mean $\lvert d \rvert$	$\frac{\sum \lvert d \rvert}{(n-1)}$	$\frac{\sum f \lvert x_{mid} - \bar{x} \rvert}{(\sum f - 1)}$
Mean d^2	$\frac{\sum (x-\bar{x})^2}{(n-1)}$	$\frac{\sum f (x_{mid} - \bar{x})^2}{(\sum f - 1)}$
Root Mean d^2	$\sqrt{Mean\ d^2}$	$\sqrt{Mean\ d^2}$
or $s_x = SD(x)$		
Relative Spread	$CV = 100 \times (\frac{SD}{Mean})\%$	
	"***Coefficient of Variation***"	

$$\underline{\textit{Harmonic Mean}} = \frac{1}{\frac{(\frac{1}{1}+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16})}{5}}$$

$$= \frac{5}{(\frac{1}{1}+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16})}$$

$$= \frac{5}{\frac{31}{16}}$$

$$= 2.6$$

Notes:

Harmonic mean = reciprocal of mean of reciprocals

Geometric mean = inverse log of mean of logs

Harmonic mean $\leq$ *Geometric mean* $\leq$ *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.

7 "'Standard" Deviation (écart-type; typical deviation)

SS	x	$x-\bar{x}$	$\bar{x}$	$x-\bar{x}$	$(x-\bar{x})^2$	$\bar{x}$	$(x-\bar{x})^2$
01	4.2	-1.75	5.95		3.0765	5.95	
02	3.0	-2.95	5.95		8.7261	5.95	
03	7.6		5.95	1.65		5.95	2.7093
04	3.8	-2.15	5.95		4.6397	5.95	
05	6.4		5.95	0.45		5.95	0.1989
06	7.7		5.95	1.75		5.95	3.0485
07	9.4		5.95	3.45		5.95	11.8749
08	6.7		5.95	0.75		5.95	0.5565
09	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	
$\sum$	77.4	-10.92	+	10.92	24.6023	+	22.7500
				=0			= 47.3523

$\bar{x} = \frac{77.4}{13} = 15$	Variance: $s^2 = \frac{47.3523}{12} = 3.9460$
$ave \times (x-\bar{x}) = 0$ $ave \times (x-\bar{x}) = \frac{\lvert -10.92 - 10.92 \rvert}{12}$ $= \frac{21.84}{12}$ $= 1.92$	Std Deviation: $= \sqrt{Variance}$ $= \sqrt{3.946}$ $= 1.99$

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

Standard Deviation is Close to Average Absolute Deviation

i.e. deviations without regard to sign .. 1.75, 2.95, ... 0.05, 2.25. Average of these absolute deviations is very close to $\sqrt{}$ of ave of squared deviations.

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

see M and M, middle of page 53

8.1 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.

8.2 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 σ^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 $\sum(x-\bar{x})^2$ tends to be too small.

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

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

Thus,

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

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 $\frac{\sigma^2}{n}$, so that the average value of the calculable estimate on the left hand side is

$$\sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

i.e., on average, if we used a divisor of n, we would <u>*underestimate*</u> σ^2 by a factor of $\frac{1}{n}$.

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

8.2.1 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 $\frac{1}{3}$, $\frac{1}{3}$ & $\frac{1}{3}$, then $\mu \;=\; 3$

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

- Samples $[x_1,\ x_2]$ of size $n = 2$ ...

  9 equiprobable samples

		x_2		
		1	3	5
x_1	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:

		x_2		
		1	3	5
x_1	1	1	2	3
	3	2	3	4
	5	3	4	5

  e.g. $xbar \;=\; \frac{(1\;+\;5)}{2} \;=\; 3$

  i.e., 9 equiprobable variance estimates

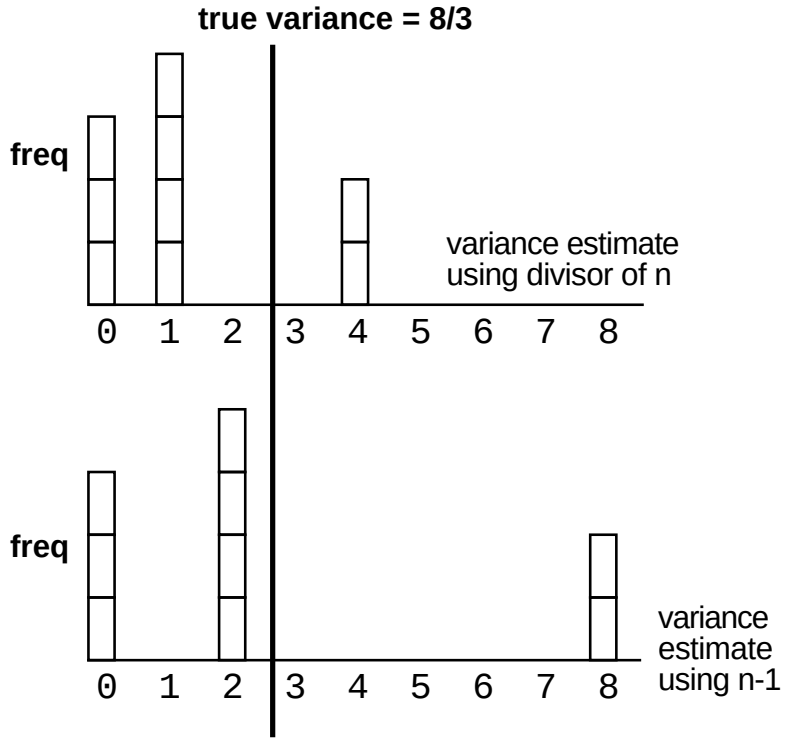
  (estimate in [ ] uses divisor of $n \;=\; 2$; estimate outside [ ] uses $(n \;-\; 1) \;=\; 1$)

		x_2			
		1	3	5	Var estimates with $(n \;-\; 1) \;=\; 1$
x_1	1	0[0]	2[1]	4[8]	$\frac{[(1\;-\;3)^2]}{1} \;=\; 8$
	3	2[1]	0[0]	2[1]	
	5	4[8]	2[1]	0[0]	$\frac{[(1\;-\;3)^2]}{2} \;=\; 8$

(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	$\frac{4}{3}$	$\frac{8}{3}$

In Pictures...



8.3 Note of the usage of "±" 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±SD in the description of a set of observations. **The use of the ± 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 "±" 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).

8.4 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

9 Re-Locating and Re-Scaling Numerical Data

see M&M page 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 the Std. Deviation σ_Y of Y?

Change	What it does to		
	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. $Y = bX + a$	$b(\mu_X + a)$	$b\sigma_X$	$b^2\sigma_X^2$
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 $b = \frac{1}{sigma_X}$

i.e. $Y = b(X + a)$	$b(\mu_X + a)$	$b\sigma_X$	
	$(\frac{1}{\sigma_X})(\mu_X + -\mu_X)$	$(\frac{1}{\sigma_X})\sigma_X$	
	0	1	1
Application —	$C^\circ = \frac{5}{9}(F^\circ - 32)$;	$F^\circ = 32^\circ + \frac{9}{5}C^\circ$	

Application — computational: Use of "working units"
... cf A& B p27

10 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 $N(\mu\ ,\ \sigma)$

How it arises

- "Naturally"

  Biological measurements...
  e.g. height

- "Manmade"

  Sampling distribution
  - Binomial and Poisson as $\mu = n\pi$
  - 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

10.1 Using the Gaussian Tables

What % of observations would be

(What is prob that single observation would be)

$X \rightarrow \%$

$>?$

Take advantage of the 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$(x, \mu, \sigma, TRUE)$ gives the

10.2 How to use one Gaussian distribution table for all N(μ,σ) 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 or 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

$$z \ = \ \frac{x \ - \ \mu}{\sigma} \ = \ \frac{130 \ - \ 100}{13} \ = \ 2.31 \qquad (1)$$

The place o 130 on the (100,130) distrn. is the same as the place of $z = 2.31$ on the "Standardized" $N(0, 1)$ or "N" 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

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

$$8.8\ UQ\ points\ above\ \mu(= 100)\ is\ 100 \ + \ 8.8 \ = \ 108.8$$

In this algebraic notation, what we have done is calculate

1. $z \ \times \ SD$
2. $X \ = \ \mu \ + \ z \ \times \ SD$

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

10.3 Using Excel functions instead of Table A

The function **NORMDIST**$(x, \ \mu, \ \sigma, \ TRUE)$ 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 **NORMMINV**(Prob,μ, σ) gives the reverse (INVerse) i.e. that value of X below which lites $100 \ \times \ Prob\%$ of the Normal distribution with mean μ and standard deviation σ.

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