Statistics for Clinicians: (2) The Normal Distribution and the Intervals of Confidence.

Ahmed A. Hassouna, MD

Philosophically speaking, normal distribution does represent one of the empirically verified elementary "truths about the general nature of reality," and its status can be compared to the one of fundamental laws of natural sciences.

Table (1) shows the distribution of birth weights among 95 newborns at a maternity hospital (1). In order to facilitate data collection and clarify the presentation, birth weights are subdivided in 13 classes; each of 200 gm range. As shown, classes of a variable can be either presented by the range or the center of the class. The frequency in each class can figure either as absolute (number) or relative (percentage) value. We advise the reader to return to the previously published equations (2) and calculate the mean (m), variance (S2), standard deviation (SD) and standard error of mean (SEM); which equal 3196.8 gm, 210096.3 gm2, 458.4 gm and SEM 47.03 gm; respectively.

Table 1: The distribution of birth weights of 95 babies as recorded in a maternity hospital:

Birth weight classes (gm)		Birth weight frequency		Total weight (gm)
(a) Centers	(b) Range*	Absolute (number)	Relative (%)	
2100	2000-2200	2	2.1	4200
2300	2200-2400	4	4.2	9200
2500	2400-2600	6	6.3	15000
2700	2600-2800	4	4.2	10800
2900	2800-3000	10	10.5	29000
3100	3000-3200	18	18.9	55800
3300	3200-3400	21	22.1	69300
3500	3400-3600	17	17.9	59500
3700	3600-3800	5	5.3	18500
3900	3800-4000	4	4.2	15600
4100	4000-4200	3	3.2	12300
4300	4200-4400	0	0	0
4500	4400-4600	1	1.1	4500
Total		95	100	303700

**** = up to but not including and the upper limit of an interval is included in the next interval***

Accepted for publication Nov 10,2005

Address reprintrequest to Dr. A Hassouna

Department of Cardio-Thoracic Surgery

Ain Shams University 980 El Mokatam

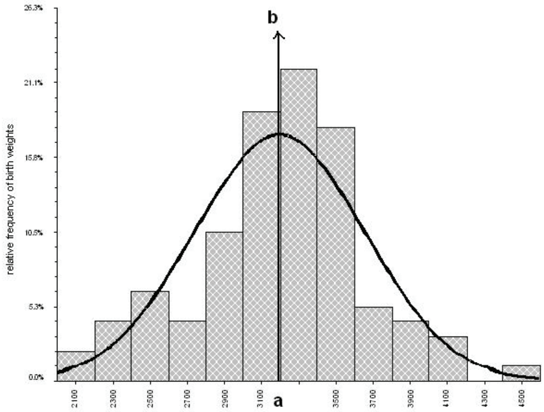
Cairo , Egypt

Email:ahmedhassouna@hotmail.com

Codex:04/02/edct/0512

Figure 1 is a relative frequency histogram of data presented in Table 1, with a vertical arrow (ab) passing through the mean birth weight (3196.8 gm).

Figure 1: Relative frequency histogram of data presented in Table1 with normal distribution curve



The vertical arrow (ab) passes through the mean birth weight (3196.8 gm).

Our histogram has several characters: firstly, birth weights are centered on their mean value and the numbers of births (and corresponding relative frequencies) decrease as we get further away of the mean. Secondly, there are specific relations between the mean and the SD. An interval sandwiching the mean by 1 SD on each side (the interval formed between 3196.8 - 458.4 = 2738.4 gm and 3196.8 + 458.4 = 3655.2 gm) comprises about 2/3 of values (66 births or 69.5% of the total 95 births).

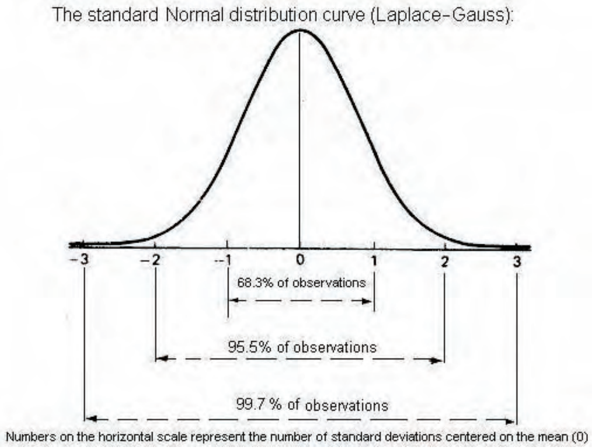
The interval formed by the mean + 2 SD (3196.8 + 916.8 = 2280 and 4113.6 gm) comprises about 95% of values (92 of the 95 births or 96.8% of the study sample) and nearly all births are comprised within the interval formed by the mean + 3 SD (3195.8 + 1375.2 gm).

Thirdly, a curved line joining the center of the classes creates an inverted bell-shaped curve which summit overlays the mean birth weight.

Such characteristics put our data within the limits of what is known as a "Normal distribution"; which is typically presented in figure 2.

As a reference, statisticians have created a perfect (standard) Normal distribution with a mean value of 0 and a SD of 1. Returning to our example, it appears that our data (vide-supra) are not far from the figures of the model, where "exactly" 68.3%, 95.5% and 99.7% of observations lie within a distance of 1, 2 and 3 SD from either sides of the mean; respectively (Figure 2).

Figure 2: The standard normal distribution curve: zLaplace-Gauss



Hence, our role is to try to fit (not to force) our data in this model, in order to ensure Normality (which is another term that can be used) however, some rules have to be drawn here (1, 3):

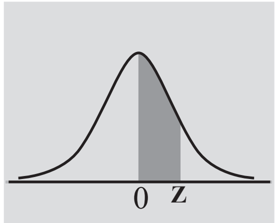
1- Normal is just a name and does not mean that other distributions are abnormal.

2- By Normality we mean: Normal distribution of the studied variable in "the population of concern", and not necessarily in the studied "sample". As an example, if we are studying serum albumin in a group of patients - and even if serum albumin can not be demonstrated to be normally distributed in this particular sample or group- normality can be assumed because we already know that serum albumin is normally distributed among the "population" from which our sample was drawn.

3- On the other hand, if the distribution of the studied variable in the population is either unknown or known to be other than normal, the inclusion of > 30 patients per studied group is sufficient to consider a practical near normal distribution. This is based upon the central limit theorem, where the means of random samples from any distribution (Normal or other) will themselves have a Normal distribution. In consequence, the more we include patients, the more the variability is diluted and the more we approach a Normal distribution.

4- There are tests for checking normality and the simpler of which, even if not totally reliable, is to plot a histogram of the data as shown. If the distribution of the recorded values (x) is far from Normality, a simple change of value like (1/x), log, and log-10 may be all what is needed. In fact, a "perfect" Normal distribution is rare and a "near" Normal distribution is usually sufficient.

5- Lastly, what is the big deal about the variable being Normally distributed? Normality is a plus but not a necessity: Most of the commonly used statistical tests (Student's test, ANOVA, correlation, regression, etc…) do necessitate the presence of certain parameters for being applied (that is why those tests are called parametric statistical tests); the most important of which is the Normal distribution of the studied variable. Even though other tests that are known as distribution-free tests are as effective and do not necessitate parameters for application; normality included.



A confidence interval can be thought of as the set of true

Figure 3: The Standard (Z) Normal distribution and Table.

	0.00	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

The values inside the given table represent the areas under the standard normal curve for values between 0 and the relative z-score. For example, to determine the area under the curve between 0 and 2.36, look in the intersecting cell for the row labeled 2.30 and the column labeled 0.06. The area under the curve is .4909. To determine the area between 0 and a negative value, look in the intersecting cell of the row and column which sums to the absolute value of the number in question. For example, the area under the curve between -1.3 and 0 is equal to the area under the curve between 1.3 and 0, so look at the cell on the 1.3 row and the 0.00 column (the area is 0.4032).

(but unknown) differences that are statistically compatible with the observed difference (4).

Returning to the Standard Normal distribution (Figure 2), 95.5% of observations are included in the interval formed by the mean + 2 SD. Looking the other way round; we can state that the probability for an observation to be included in the interval formed by the mean + 2SD is 95.5%. Statisticians have calculated this probability by measuring the surface area under the curve that is included between the mean and the number of SD in question (2 in our case), on either side of the mean. Then, this area was referred to the total surface area under the curve; which in our case was found to be 0.955 of that area. In fact, statisticians had the courtesy to calculate all expected probabilities (relative surface areas under the curve) not for 1, 2 and 3 SD; but for all possible positions as one moves on the horizontal scale, further away from the mean. They did so, for whole as well as fractions of SD (0.01 to 3 SD) and presented the "Z table" (table 4). However, the units on the horizontal scale are no more the units and fractions of SD of the standard Normal curve, but are those of any other Normal distribution, that are first standardized so as to fit the perfect model before reading from the table the expected probability.

Those units are called the "relative Z values" and are calculated by the following simple equation: for any given value (x) of any studied Normal distribution with a mean value (m) and a SD (σ);

Z = (x-m)/ σ.

The probability of a Z value of 2 is the probability for a value (x) to be included in an interval formed by 2 σ (2SD), on either of the mean (m). As shown in the figure above Table (4), probabilities are given for Z values on 1 side of the mean or the absolute Z values; the other side being a mirror image of the one figuring in the table. In other words, the probability for an observation to be included in the interval formed by the mean + 2SD = that figuring in the table for a Z value of 2 multiplied by 2 = 0.4772x2= 0.95.4 = 95.4%. Checking Table 4 at the intersection of line 1.9 and column 0.06 which sums a Z value of 1.96 shows that a variable chosen at random from the population of which our studied sample was drawn has a 95% chance (0.475x2) to lie in the interval formed by 1.96 SD, on either sides of the mean value of our sample.

Applying this to our example: there is a 95% chance that the birth weight of (a coming baby) -from the population from which our sample was drawn- will be included in the interval formed by the mean of our sample (3196.8 gm) + 1.96 SD (1.96x 458.4=898.5 gm) or between 2298.3 and 4095.3 gm. This is the interval of confidence of a subject at 95%:

IC of a subject at 95% = mean + 1.96 SD

By checking Table 4, one can calculate other intervals of confidence with more or less precisions as indicated; e.g. the interval of confidence at 99% = mean + 2.6 SD, and the interval of confidence at 75% = mean + 0.385. The more we include patients in our study, the more SD will decrease and the more the interval of confidence narrows. Those upper and lower limits of the interval of confidence figure for normal values on our laboratory sheets (e.g. blood sugar levels, renal functions, hepatic functions, red blood cell count, white cell count, etc...).

To end, our work permits us to calculate another important interval of confidence; which is the interval of confidence of the mean:

IC of mean at 95% = m + 1.96 SEM

To explain the interval of confidence of mean, let's imagine that the previous study was repeated, whether by the same researcher or by others, on another comparable group of patients. No two studies will ever calculate the same mean, SD and SEM values and one should ask himself: what is the true mean birth weight (M) of all newborns of comparable mothers? For sure no one can have the definite answer, but one can always calculate a range (i.e. an interval of confidence) which would embrace (M).

In other words, one should expect that the calculated mean birth weight in other comparable studies will have a 95% chance to be within the range of m + 1.96 SEM = 3196.8 + 92.1 = between 3104.7 and 3288.9 gm. Also the more I include cases in my study; the lower would be my SEM and the narrower and more approaching to the real values, would be my calculated interval of confidence. Remember that unlike that of a subject, the IC of the mean does not necessitate the normal distribution of the studied variable (vide-supra).

A confidence interval is typically reported in the following way: "The mean birth weight was 3196.8 gm (95% CI, 3104.7 to 3288.9 gm)"

This means that even though the observed birth weight was 3196.8 gm, the data are statistically compatible with a true birth weight difference as small as 3104.7 gm or as large as 3288.9 gm. True differences that lie outside the 95% confidence interval are not impossible; they merely have less statistical evidence supporting them than values within it. The choice of 95% as the standard convention is somewhat arbitrary and corresponds to the use of a threshold of $P < 0.05$ for statistical significance.

Statistics

References:

1- Schwartz D. Les fluctuations d'échantillonnage d'une moyenne. In: Schwartz D, ed. Méthodes statistiques a l'usage des médecins et des biologistes. 3rd edition, Paris: Médecines-Sciences Flammarion; 1969: 103-116.
2- Ahmed HASSOUNA: Statistics for clinicians: the 4 basic indices. J Egypt Soc Cardiothorac Surg 2005; 13 (1): 14-6.
3- Altman DG and Bland JM. Statistics notes. The normal distribution. BMJ 1995; 310: 298.
4- Goodman SN and Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med 1994; 121 (3): 200-6.
5- Schwartz D. Les fluctuations d'échantillonnage d'une moyenne. In: Schwartz D, ed. Méthodes statistiques a l'usage des médecins et des biologistes. 3rd edition, Paris: Médecines-Sciences Flammarion; 1969: 103-116.
6- Ahmed HASSOUNA: Statistics for clinicians: the 4 basic indices. J Egypt Soc Cardiothorac Surg 2005; 13 (1): 14-6.
7- Altman DG and Bland JM. Statistics notes. The normal distribution. BMJ 1995; 310: 298.
8- Goodman SN and Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med 1994; 121 (3): 200-6.