

Chapter 4

Normal Distributions

A frequency distribution of interval or ratio scale data has a preponderance of values around the mean with progressively fewer observations towards the extremes of the range of values. If the frequency of observations, n , is large, then the plot of the frequency (Y-axis) on the value of the observation (X-axis) is "bell shaped." Such distributions are generally termed **normal distributions** or **Gaussian distributions**.

The location and shape of these curves is dictated by the mean (μ) and the standard deviation (σ). Two such normal distributions with different μ s and σ s are displayed in Figure 4.1. The location on the X-axis is controlled by μ , and normal distributions with larger μ s are located upscale of those with smaller μ s (Figure 4.1). The shape of the "bell," its dispersion or spread, is controlled by σ , and a normal distribution with larger σ appears flatter and more dispersed than one with smaller σ (Figure 4.1),

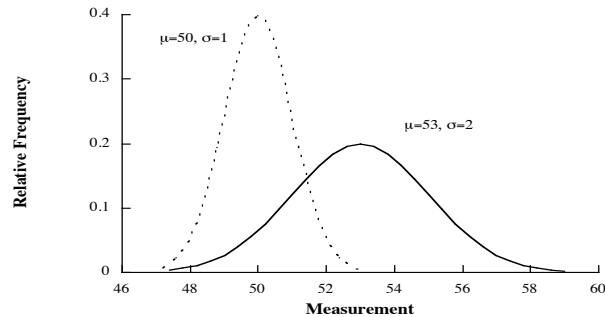


Figure 4.1. Two normal distributions.

A normal curve with a $\mu = 0$ and $\sigma = 1$ is said to be a **standardized normal curve**. Data from any normal distribution can be standardized by using the equation of

$$Z = \frac{(X_i - \mu)}{\sigma} \quad [4.1]$$

where X_i is an individual value. Basically, subtracting μ locates the standardized normal curve over a Z value of 0 on the independent (X) axis, and dividing by σ sets the dispersion of the curve to a standard one Z. Carrying out this calculation is normalizing or standardizing the individual value X_i , and Z is called a **normal deviate** or a standard score. Both of the data sets shown in Figure 4.1 are converted to one standardized normal curve shown in Figure 4.2.

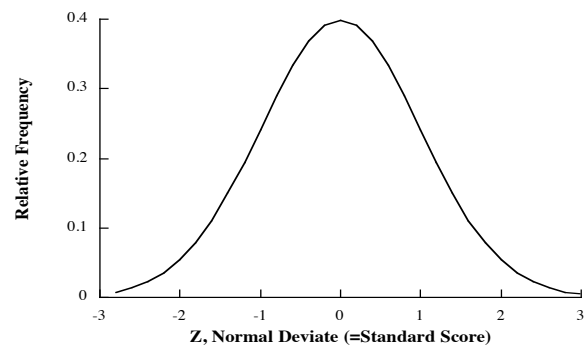


Figure 4.2. Standardized normal curve.

Another way to think of the standardized normal curve is to imagine constructing a huge known population with some variable, say length, which has a known μ and σ . Then imagine taking an individual from the population, measuring its length, and then calculating the Z score. Imagine doing this a zillion, zillion times, and then plotting on the Y-axis the frequency of times that a particular Z score on the X-axis is observed. If the values on the Y-axis, the frequencies, are divided by the total number of Z scores obtained, then that axis becomes relative frequency. The resulting plot would be identical to Figure 4.2.

Proportions of a Normal Distribution

The area under a standardized normal curve can be set to 1, and the proportion of the normal distribution lying between any values of Z can be determined. These proportions are probabilities, and are packaged in tabular form (Table 1, Appendix B). Table 4.1 presents a subset of the probability table for normal distributions for illustration of its usage. This table gives the proportion of the standardized normal distribution that lies beyond a given value of Z , and it can be used to determine what proportion of measurements in a normal distribution lie between or beyond a variety of limits.

Table 4.1. Selected proportion of a normal distribution that lies beyond a given normal deviate (Z). For example, the proportion of a normal distribution for which Z is greater than or equal to 1.02 is 0.1539; symbolically this is expressed as $P(Z \geq 1.02) = 0.1539$.

Z	<u>0</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	Z
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641	0.0
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776	0.5
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379	1.0
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183	2.0

As an example of how normal deviates can be used, assume a population of fish with a mean (μ) oxygen consumption of 10 ml O₂ min⁻¹ and standard deviation (σ) of 1 ml O₂ min⁻¹. Consider a single value of oxygen consumption, at 11 ml O₂ min⁻¹. The standard score, Z , for this X would be 1.00 because $Z = (11 - 10)/1$ from formula 4.1. Consultation of Table 4.1 indicates that 0.1587 of the distribution is larger than a Z of 1.00. This can be stated, as "the probability of obtaining a Z greater or equal to 1.00 is 0.1587." It also can mathematically noted as $P(Z \geq 1.00) = 0.1587$. And, remember that the Z of 1.00 was derived from the oxygen consumption of 11 ml O₂ min⁻¹. Thus, for this population, $P(X_i \geq 11 \text{ ml O}_2 \text{ min}^{-1}) = P(Z \geq 1.00) = 0.1587$. This probability can be interpreted in a number of different ways. It could be said that the probability of drawing at random from this population one individual with an oxygen consumption equal to or greater than 11 ml O₂ min⁻¹ is 0.1587. Or 15.87% of the population would be expected to have an oxygen consumption of 11 or greater. Conversely, 84.13% of the population would be expected to have an oxygen consumption less than 11.

Consider another example for this population. What's the probability of having an oxygen consumption that is lower than 9.5 ml O₂ min⁻¹? Here $Z = (9.5 - 10)/1 = -0.5$, and $P(X_i < 9.5) = P(Z < -0.5) = 0.3085$ (or 30.85%). Thus, the probability is 0.3085, and 30.85% of the population would be expected to have oxygen consumption values of 9.5 ml O₂/min or less. Conversely, 69.15% would be expected to have an oxygen consumption larger than 9.5 ml O₂/min.

Finally, consider what proportion of the population is enclosed between the mean plus or minus the standard deviation (Figure 3.3). For our population numerically that would encompass a range of oxygen consumption's from 9 to 11 ml O₂ min⁻¹. At the mean of 10, $Z=0$ and 0.5 of the population has a higher value as indicated by Table 4.1. At one standard deviation, $Z=1$ and 0.1587 of the population has a higher value (Table 4.1). The proportion of

the population between Z of 0 and 1 would then be $0.3413 (= 0.5 - 0.1587)$. Thus, 0.3413 is the amount of area under the curve from the mean to *plus* one standard deviation of the mean.

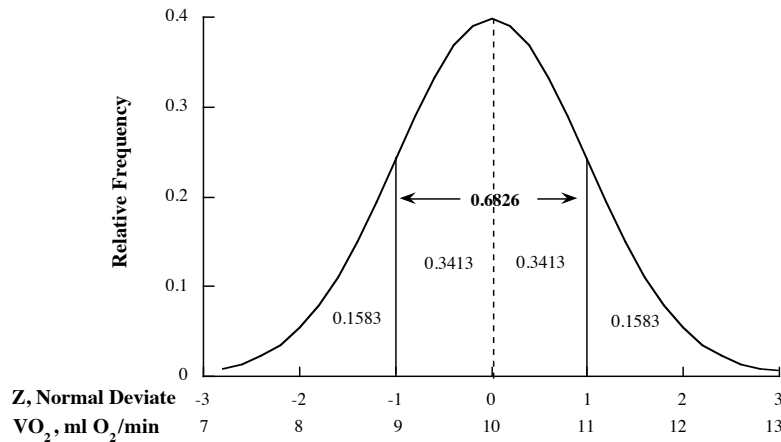


Figure 4.3. Standardized normal distribution of oxygen consumption ($\text{ml O}_2 \text{ min}^{-1}$) with $\mu=10$, $\sigma=1$ showing the probabilities for the mean \pm one standard deviation.

Obviously, 0.3413 is also the area under the curve from the mean to *minus* one standard deviation of the mean. The two values combined are 0.6826 . Thus, it can be said that the mean plus or minus (\pm) the standard deviation of any population includes 68.26% of the population. Or, specifically in the population we are considering 68.26% of the individuals would be expected to have an oxygen consumption of 9 to $11 \text{ ml O}_2 \text{ min}^{-1}$. It can also be said that a) 31.74% will have a value outside that range, b) 15.87% will have a value lower than 9 , and c) 15.87% will have a value greater than 11 .

It can be determined similarly that a) the mean \pm twice the standard deviation includes approximately 95% of the population and b) the mean \pm three times the standard deviation includes about 99% . Interested students should be able to calculate the exact percentage for these values.

Central Limit Theorem

If random samples of size n are drawn from a normal population, the *means* of these samples also will form a normal distribution (see Figure 4.4). The distribution of means from a non-normal population will tend towards normality as the sample size (n) increases. The measurement of dispersion (*e.g.*, standard deviation) of the sampling distribution of means will decrease as the sample size increases, and is computed as

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad [4.2]$$

where $\sigma_{\bar{x}}$ is the **standard deviation of the mean**. The standard deviation of a parameter or a statistic is called a standard error; thus $\sigma_{\bar{x}}$ is the standard error of the mean or simply the **standard error**.

Just as a distribution of individual measurements, X_i , can be standardized so can a distribution of means. This is done by computing the normal deviate as

$$Z = \frac{(\bar{x} - \mu)}{\sigma_{\bar{x}}} \quad [4.3]$$

where \bar{x} = the mean of the sample. Here Z can be used to answer questions concerning the probability of obtaining a particular mean based on random samples of n measurements from a population with a known μ and σ .

In order to compute $\sigma_{\bar{x}}$, the parametric standard error of the mean, the parametric standard deviation, σ , must be known. In most biological cases σ of a population is not known and must be calculated based on a random sample from the population. The best estimate of σ is the sample standard deviation, s , and it can be used to compute the statistic standard error of the mean, $s_{\bar{x}}$, which is the best estimate of $\sigma_{\bar{x}}$. Thus, the statistic standard error of the means, $s_{\bar{x}}$, is computed as

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \quad [4.4]$$

where n is the size of the sample used to compute s .

Normal distributions are illustrated in Figure 4.4 for individuals and for means for a population with $\mu=53$ and $\sigma=2$. The widest distribution indicated ($n=1$) is a distribution of individual measurements, and the vertical line at 55 indicates the mean plus one standard deviation for this distribution. The next widest distribution ($n=4$) was obtained by sampling four individuals, determining the mean for the four values, and repeating this many times. This distribution is narrower than the first one because it is based on means of four measurements and not on individual measurements. The vertical line at 54 indicates its mean plus one standard deviation. In this case, the standard deviation is really a "standard deviation of a mean" which is more popularly called a "standard error" computed from equation 4.2.

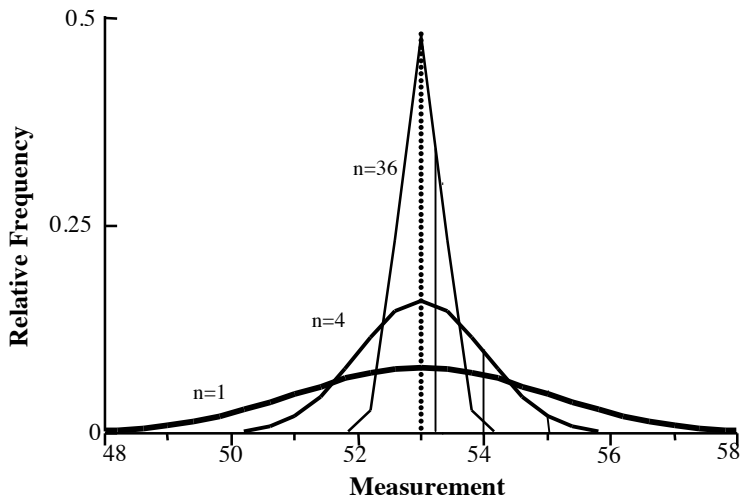


Figure 4.4. Normal distributions of a population with $\mu=53$, $\sigma=2$. The distribution noted as $n=1$ is a distribution of individuals measured one at a time. The other two distributions are distribution of means based on sample sizes of 4 and 36. See text for further explanation.

The next narrowest distribution ($n=36$) is based on means of samples of 36 individuals. It is, of course, much narrower than the other two and the vertical line at 53.33

indicates where its mean plus one standard error lies. In this case, the standard error is 0.33 ($=2/\sqrt{36}$). Finally, the dotted vertical line at 53 indicates a $\mu=53$. It would be the distribution obtained based on the means obtained of *every* individual in the population.

Based upon these data, the following statements can be made:

- 68% of the *individuals* would have values falling between 51 and 55.
- 68% of *means* taken from samples of 4 individuals would fall between 52 and 54.
- 68% of *means* taken from samples of 36 individuals would fall between 52.67 and 53.33
- We are 68% confident that any one *individual* measured would have a value between 51 and 55.
- We are 95% confident that any one mean obtained from a sample of $n=4$ would have a value between 51 and 55.

Can you think of other statements that might be made?

The **standard error** is an estimate of variability of means that is sample-size dependent, decreasing with increased sample size. In contrast, the **standard deviation** is an estimate of variability of the items in the sample (e.g. individual organisms) and is not sample-size dependent. The standard error of the mean is used to describe variability of means of some measurement while the sample standard deviation is used to describe variability of the individual items being measured in a sample. Standard error of the mean is most frequently used by biologists interested in evolutionary problems (e.g. physiological ecologists) since populations biologically evolve while individuals don't. Conversely, biologists interested in individuals (e.g. medical researchers) use standard deviation.

Problem Set – Normal Distributions

- 4.1) A normally distributed population of lemming body weights has a mean of 63.5g and a standard deviation of 12.2g.
- What proportion of this population is 78.0g or larger?
 - What proportion of this population is 78.0g or smaller?
 - If there are 1000 weights in the population, how many of them are 78.0g or larger?
 - What is the probability of choosing at random from this population a weight less than 41.0g?

For questions 4.3 to 4.11 use the following information. The blood glucose of a population of lesser dodos had blood glucose levels with $\mu=65$ and $\sigma=25$ with the units being mg per 100 ml (= mg/dL).

- 4.3) What proportion of the population has a blood glucose level greater than or equal to 85 mg/dL?
- 4.4) What proportion of the population has a blood glucose level less than 45 mg /dL?
- 4.5) What proportion of the population has a blood glucose level above 45 but less than 85 mg /dL?
- 4.6) If a random sample is made of 20 animals, what is the probability that the sample mean obtained will be above 70 mg /dL?
- 4.7) A glucose concentration equal to or exceeding ____ (what value) would be expected to be exhibited by 5% of the population?
- 4.8) It has been determined that lesser dodos with blood glucose concentrations that lie outside the middle 70% of the population die at younger ages; therefore, they need to be treated for abnormal glucose concentrations. At what lower and upper concentration should treatment be initiated?
- 4.9) As a college professor teaching a difficult lower division course you have decided on the following distribution of grades: A, upper 10%; B, next 20%; C, next 50%; D, next 15%; F, lowest 5%. The class (population) average number of points at the end of the quarter is 923 with a standard deviation of 101. Determine the number of points needed for each letter grade.

YOU WILL NOT TURN IN THE ANSWERS TO THESE PROBLEMS, BUT WE WILL GO OVER THEM IN CLASS ON MONDAY OCTOBER 5. PLEASE BE PREPARED TO WORK THE PROBLEMS ON THE WHITE BOARD IF ASKED.