J of Nuclear Medicine Technology, first published online February 2, 2018 as doi:10.2967/jnmt.117.204719

Title: Statistics Refresher for Molecular Imaging Technologists Part 2: Accuracy of Interpretation, Significance, and Variance

Short running title: Statistics Refresher Part 2

Author: Mary Beth Farrell, MS, CNMT, NCT, FSNMMI-TS

Affiliation: Intersocietal Accreditation Commission

Corresponding Author/First Author:
Mary Beth Farrell, MS, CNMT, FSNMMI-TS
27 Boxwood Lane
Langhorne, PA 19047
Phone: 443-285-2503
Email: marybethfarrell2016@gmail.com

Word count: 3852

Abstract (343/350 words)

This article is the second part of continuing education series reviewing basic statistics that nuclear medicine and molecular imaging technologists should understand. In this article, the statistics for evaluating interpretation accuracy, significance and variance are discussed. Throughout the article, actual statistics are pulled from published literature.

Part two begins by explaining two methods for quantifying interpretive accuracy: inter-reader and intra-reader reliability. Agreement among readers can simply be expressed by percentage. However, Cohen's kappa is a more robust measure of agreement that accounts for chance. The higher the kappa score, the more agreement between readers. When three or more readers are being compared, Fleiss' kappa is used.

Significance testing determines if the difference between two conditions or interventions is meaningful. Statistical significance is usually expressed using a number called a P-value. Calculation of P-value is beyond the scope of this review. However, knowing how to interpret P-values in important for understanding scientific literature. Generally, a P-value less than 0.05 is considered significant and indicates that the results of the experiment are due to more than just chance.

Variance, standard deviation, confidence intervals and standard error explain the dispersion of data around a mean of a sample drawn from a population. Standard deviation is commonly reported in the literature. A small standard deviation indicates that there is not much variation in the sample data. Many biologic measurements fall into what is referred to as a normal distribution taking the shape of a bell curve. In a normal distribution, 68% of the data will fall within one standard deviation, 95% will fall between two standard deviations, and 99.7% of the data will fall within three standard deviations.

Confidence intervals define the range of possible values within which the population parameter is likely to lie and gives an idea of the precision of the statistic being measured. A wide confidence

interval indicates that if the experiment were repeated multiple times in other samples, the measured statistic would lie within a wide range of possibilities. Confidence intervals rely on the calculation of another metric called the standard error.

Part 1 of the continuing education series, *Statistics Refresher for Molecular Imaging Technologists: Testing the Test,* reviewed the statistics important is in describing the accuracy of a diagnostic procedure. In other words, expressing how well a test distinguishes between two conditions, e.g., disease is present and disease is absent. The ability of a diagnostic test to discriminate is quantified by measures of diagnostic accuracy including sensitivity, specificity, accuracy, positive predictive value, negative predictive value, pre-test probability and post-test probability.

Part 2 of this series will review several additional statistical concepts with which molecular technologists should be familiar. First, accuracy of interpretation will be discussed. Interpretive accuracy is described based on the level of consistency or agreement between observers: inter-reader reliability and intra-reader reliability. Part 2 will also briefly discuss hypothesis testing and significance. Significance testing identifies meaningful differences between two tests and differences due to chance. Finally, Part 2 will review the less fascinating but crucial statistics describing variance, including standard deviation, confidence intervals, and standard error.

Examples from nuclear medicine and molecular imaging literature are used to illustrate each of the specific statistical concepts. It is hoped that the statistical concepts will be more easily understood when described in the context of real-world imaging.

ACCURACY OF INTERPRETATION

Interpretation issues must be considered when evaluating a molecular imaging test. How do you know if an interpretation is accurate? How do you know if the same interpreter would read a scan similarly if presented a second time? Does the test perform to the same sensitivity and specificity among different readers? How often do readers agree in their interpretation of the test? These concepts are discussed in the section below about assessing reader accuracy and agreement.

BLINDED INTERPRETATIONS

The best way for physicians to determine interpreter accuracy is to perform a blinded read experiment. An interpreter is considered "blinded" when they are provided with the images alone, without any medical history, description of clinical symptoms or other diagnostic testing information. Results of the blinded interpretations are then compared to the known results.

For example, researchers looked at sensitivity and specificity of F18 Florbetapir Injection (florbetapir), a PET amyloid imaging tracer (Amyvid®; Eli Lilly and Co), in the hands of multiple interpreters. Five independent and blinded nuclear medicine physicians were asked to interpret 46 scans from patients at end of life who expired within 12 months of the amyloid PET. The standard of truth for comparison was pathologic confirmation of amyloid plaque at autopsy, and the dataset included both positive and negative amyloid confirmations. The sensitivity of majority reads across 5 readers was 96% and specificity was 100%.

In the clinical patient care setting, interpreters assume that, using the prescribed interpretation technique and acquiring images properly, the test performs as well for all interpreters as it did in the blinded read experiment.

Patient specific or lesion specific accuracy is difficult to measure in clinical practice without biopsy confirmation, and often cannot be measured at all when the scan is negative and no additional testing or follow up is performed. Nuclear medicine physicians and radiologists can participate in hospital quality initiatives which systematically follow a group of patients and analyze outcomes against imaging results, and thus measure their own accuracy of interpretation. This is routinely done in mammography, for example. National mammography standards require that all radiologists who interpret mammography must do routine checks of accuracy (*1*). However, interpretation is more frequently assessed by comparing interpretations between two readers or by comparisons for the same reader.

INTER-READER AND INTRA-READER RELIABILITY

Agreement among readers is also an important characteristic for a diagnostic test. How would you measure the level of agreement or disagreement among readers? If a scan has excellent accuracy with one expert reader, but additional readers disagree about the findings, the test is less valuable. Inter-reader reliability is the measurement of how frequently interpreters agree with each other. The higher the reliability, the more readers agree with one another and interpretation is more standardized across users. The lower the reliability, the less agreement among readers. Intra-reader reliability is the measure of how consistently one reader interprets the same scan a second time. Ideally intra-reader reliability should be high, meaning that an interpreter reads the same scan the same way every time.

Agreement among a group of readers can be expressed as a percentage of the total reads. If 100 scans are interpreted by two readers who provide a binary result (e.g. positive or negative) and they disagree on 15 scans, the inter-reader agreement would be 85%. Reader agreement between two different tests can be evaluated in the same way. For example, fluciclovine researchers looked at the agreement between fluciclovine and C-11 choline PET interpretations in the same patients by analyzing results from three independent readers. Agreement between the two tests was 61% for reader 1, 67% for reader 2, and 77% for reader three. Another way of stating the result to say that the average agreement between the two tests among three readers was 68% (*2*). (Other ways of measuring interpreter performance are median read among interpreters, or the average read result.)

KAPPA

Inter- and intra-reader reliability among two readers can also be characterized by a correlation measure called Cohen's kappa (usually denoted as Greek letter κ). This statistic measures how well two readers agree given the binary option of positive or negative. Considered to be a more robust measure of agreement than calculation of a simple percentage, kappa considers the fact that some agreement happens by chance (e.g. two readers may be guessing and happen to guess the same answer at the

same time). Kappa scores range from 0 (complete disagreement) to 1 (complete agreement.) The higher the kappa score, the more agreement between readers. A low kappa score indicates less agreement among readers. Kappa scores can be interpreted as follows (*3*):

< 0.20	Poor (little agreement no more likely to occur than by chance)
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Good
0.81-0.99	Very good
1.00	Perfect agreement

When three or more readers are evaluated, a similar statistic called "Fleiss' kappa" is used. Like Cohen's kappa, the closer the Fleiss' kappa value is to 1.0, the closer to perfect agreement among readers (*4*).

An example of the kappa measurement in PET literature is a recent study by Ohira et al (*5)* which looked at inter-reader and intra-reader reliability of PET FDG in patients being referred for cardiac sarcoidosis. The authors measured reader agreement using two interpretation strategies: reading scans in pattern categories (focal, focal on diffuse, no uptake, diffuse, or isolated lateral and /or basal) and in a binary fashion (positive scan consistent with cardiac sarcoidosis or negative scan.) The kappa for two interpreters who used the pattern categories was 0.64, which reflects good agreement. The kappa statistic for two interpreters who used the binary method was 0.85, showing a very good level of agreement. Analysis of intra-reader agreement demonstrated very good agreement for both interpretation methods (0.94 for pattern interpretation and 0.92 for binary read.) From this data we can understand the impact on inter- and intra-reader agreement for two methods of interpretation for PET FDG of cardiac sarcoidosis.

SIGNIFICANCE TESTING

When two tests or interventions are compared, how can we know if the data is meaningful? Statistical significance, usually represented by a number called a "P-value," is a way of making sure that the experimental result, or differences between two measurements, are not just due to chance. Calculation of P-values is beyond the scope of this paper, but knowing how to interpret a P-value is important for understanding scientific literature. A P-value that is less than 0.05 indicates that the results of the experiment are due to more than just chance. Put another way, the research hypothesis is that there is a difference between A and B, and the null hypothesis (generally the opposite of what you are interested in finding out) is that there is no difference between A and B. If the P-value is less than 0.05, it means that the null hypothesis is rejected and the measured difference between A and B is most likely real. A P-value that is larger than 0.05 means that there is not enough information available to reject the null hypothesis, and therefore the measured difference between A and B could be due to random chance (*6).*

To demonstrate P-values and significance, the results from a study published by Brayshaw, Mosley, and Currie in 2016 are presented in Figure 1. (*7*) The investigators evaluated the effect of hamburgers cooking nearby on the uptake of tracer in the stomach during myocardial perfusion imaging. The results demonstrated that women had significantly higher stomach counts than men associated with olfactory stimulation. By looking at the bar graphs alone it appears that counts per pixel were higher in both the heart and stomach for women. However, the difference in counts per pixel for women and men was only statistically significant for the stomach (stomach p=.022; stomach background corrected p=.018). This means that the null hypothesis (i.e. no difference between counts in the stomach for women vs. men) must be rejected. In this example, the P-value helps to illustrate the differences between the stomach and heart.

VARIANCE

To illustrate variance, standard deviation, confidence intervals, and standard error imagine an experiment to calculate the average weight of male patients coming into your department for thyroid ablation. You collect data on 50 patients. The average weight is 175 pounds and the range is 100-250 pounds. You could stop there and report the average weight and range, but how reliable is that measurement and how confident are you that 175 pounds is a good description of a typical patient in your sample? Figure 2 demonstrates how the distribution of weights can be very different although the average is the same.

To better characterize your data, the next step is to calculate the variance – describe how far from the mean your sample weights lie. Variance of a sample is the sum of all squared differences from the mean, divided by the sample size minus 1.

Variance = $\sum(X_i - X)^2 / (n-1)$

To accomplish this, the first step is to determine absolute differences from the mean for each patient. For the patient who weighed 100 pounds, subtract 100 from 170 for a difference of 70 pounds, and 70 squared is 4900. For the patient who weighed 250 pounds, the difference from the mean is 170-250 pounds, or 80 pounds, and 80 squared is 6400, etc. After adding up all the squared differences, you divide the total by the sample size of 50 - 1 to determine the sample variation from the mean. For illustration sake, assume the sample variance is calculated as 35839. As a stand-alone number, this is not useful to the average reader of statistics; however, this number is used in the determination of standard deviation.

STANDARD DEVIATION

Standard deviation (SD) is a commonly cited statistic in medical literature used to measure the dispersion or variability in the sampling data. When we calculate the standard deviation of a sample, we are using it as an estimate of the variability of the population from which the sample was drawn. A small

standard deviation indicates that there is not much variation in the sample data for this experiment, and the calculated statistic is a precise characterization of the sample. A large standard deviation means that the data has a wide variability.

Mathematically, standard deviation of a sample is the square root of the variance.

SD sample = square root of sample variation

In our weight experiment from above, the standard deviation of our sample can be expressed as the square root of 35839, or 26.8. This means that the sample mean and standard deviation are 175 ± 26.8 pounds.

Biologic measurements, such as weight, fall into what is referred to as a "normal" distribution. This means that if you plot the data from an infinite number of samples, the results will take the form of a standard curve, referred to as a bell curve because of its shape. The mean of the group will form the peak of the curve, and all other data will cluster around that mean in a predictable pattern. In a normal distribution, 95% of the data will fall within 1.96 standard deviations of the mean (usually rounded up to 2), and the remaining 5% will be scattered at the low or high end of the range (see Figure 3). Using our patient weight example above, we know that 95% of patients will fall within approximately two standard deviations, or 53.6 pounds (26.7 x 2) above and 53.6 below. Therefore an accurate description of a typical patient in the sample is between 116.4 and 223.6 pounds. Note that, in this particular example, describing the population with the range (i.e. 100-250 pounds), while accurate, is not as useful as describing the average plus the standard deviation. Both statistics are accurate, but one is more meaningful to the researcher.

CONFIDENCE INTERVALS

Confidence intervals are another tool to help us to understand the strength of a statistic. What is the difference between confidence interval and standard deviation? A confidence interval (usually denoted as CI) defines a range of possible values within which our population parameter is likely to lie,

and gives an idea of the precision of the statistic being measured. While standard deviation describes the attributes of the individual data points that go into the sample statistic, confidence interval describes the range of results that would occur if the experiment were repeated with a different sample of the population. A wide confidence interval indicates that, if the experiment were repeated multiple times in other samples, the measured statistic would lie within a wide range of possibilities. This would indicate a lack of precision in the measurement. A narrow confidence interval means that the result is relatively more precise, and if the experiment were repeated the range of likely results is close to the original calculation. Confidence intervals can be applied to any statistic, such as mean or kappa, etc.

STANDARD ERROR

Confidence intervals rely upon calculation of another metric called standard error or standard error of the mean. Standard error (denoted as SE or SEM), like standard deviation, is a measurement of variance or dispersion from the mean. While standard deviation describes the variation between individuals and the calculated sample mean, standard error is a measurement of uncertainty in the mean statistic itself. Referring to our patient weight experiment above, the population of 50 patients is only a small sample of all patients who get thyroid ablation. The mean and standard deviation of weight is assumed to approximate the population as a whole, but if another hospital repeats the experiment, or if the experiment is repeated with 50 additional patients, the mean may be a different number. Standard error helps us to understand how reliable the mean measurement of the sample is compared to the mean of the entire population. In other words, standard error of the mean describes how much variation there would be in the calculated mean if you repeated the entire experiment again and again, and calculated an average every time. The formula for standard error is:

SE = SD/square root of sample size

Using our patient weight experiment, we could calculate the standard error of the mean as 26.8 divided by the square root of 50 (7.07) which equals 3.8. Standard error by itself is not typically an

informative statistic and is rarely cited; however standard error provides an important basis for group statistics, for example as part of a calculation of confidence interval (*6, 8)*.

The mathematical definition for 95% confidence interval (95% CI) is

(mean - 1.96 x SE) to (mean + 1.96 x SE)

Using our weight experiment as raw data with a mean of 175 pounds and standard error of the mean of 3.8, the 95% confidence limit for the calculated mean would be derived from:

(175 – 1.96 x 3.8) to (175 + 1.96 x 3.8)

(175 - 7.45) to (175 + 7.45)

167.5 to 182.5

In this example, therefore, these confidence limits tell us that if we repeated our weight experiment 100 times, we can expect 95 experiments to result in a calculated mean between 162.5 and 177.5 pounds. Confidence intervals can be calculated for other percentages such as 99% confidence or 90% confidence; however, most examples in the medical literature use 95% confidence intervals.

An example of published confidence intervals is seen in the prescribing information for florbetapir (Amyvid USPI). Inter-reader reliability was measured, and the resulting fleiss' kappa was 0.83, with a 95% confidence interval of 0.78 to 0.88. The kappa itself indicates very good agreement, but the confidence intervals tell us that there is a 95% likelihood of repeated experiments resulting in a kappa within the range of 0.78 to 0.88, all within the range of good to very good agreement. If the kappa for a hypothetical test was .83, but the confidence interval was very wide (e.g. 0.4-0.9) you would be concerned that repeating the experiment could result in a kappa ranging from fair agreement (0.4) to very good agreement (0.9.) In this way, confidence intervals help us to know how reliable the statistic would be if repeated.

While significance (P value) and confidence intervals test two different things, there is a strong relationship between the two. If the calculated 95% confidence interval for a difference between two groups or tests does not include zero, meaning the range does not extend from a negative value to a positive value (e.g., -0.60 to -0.1 or 0.02 to 0.30 , the hypothesis test will be significant (e.g. $p<0.05$). If the confidence interval includes zero (e.g., -0.3 to 0.4), then there will not be statistical significance in the comparison (e.g. $p>0.05$). This is why confidence intervals sometimes contain more clinically relevant information than P-values. Presenting a 95% confidence interval indicates whether the result is statistically significant at the 5% level, but it also provides important information about how well the measurement would hold up under repeated testing (*8*).

CONCLUSION

The goal of this basic statistics continuing education series was to provide a refresher for nuclear medicine and molecular imaging technologists. The statistics reviewed are common statistics found in the literature that technologists should understand. Part one of the series reviewed statistics used to describe characteristics of diagnostic imaging tests: sensitivity, specificity, and predictive value. Part two discussed statistics used in evaluating interpretation accuracy, significance, and variance. Throughout the series, actual statistics are pulled from published literature in the hope that the statistical concepts would more easily come to life. It is possible that a third part may be added to this series reviewing more complex concepts such as difference testing, risk, correlation, and survival analysis.

References

1. Home | Occupational Safety and Health Administration. OSHAgov. 2017. Available at: https://www.osha.gov/. Accessed October 29, 2017.
2. Blue Earth Diagnostics, Ltd. (2016). Axumin™ Fluciclovine F 18 Injection. Prescribing information. [Online]. Available: www.axumin.com. [2017, 12 Sep].
3. Altman DG (1991). *Practical statistics for medical research*. London: Chapman and Hall.
4. McHugh ML. Interrater reliability: The kappa statistic. *Biochem Med*. 2012; 22:276-282.
5. Ohira H, McArdle B, deKemp RA, et al. Inter- and intraobserver agreement of 18F-FDG PET/CT imaging interpretation in patients referred for assessment of cardiac sarcoidosis. *J Nucl Med*. 2017; 58(8):1324-1329.
6. Campbell MJ, Machin D, Walters SJ. *Medical Statistics: A textbook for the health sciences*. 4th Ed. The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England: John Wiley & Sons, Ltd.
7. Brayshaw G, Mosley S, Currie G. Increased gastric activity on myocardial perfusion imaging. *J Nucl Med.* 2016; 44:195-198.
8. Altman DG and Bland JM. Statistics notes. Standard deviations and standard errors. *BMJ*. 2005; 331:903.

Figures

Figure 1. Results from a previous study published by Brayshaw, Mosely and Currie in 2016 are graphed to demonstrate significant P-values. The study evaluated tracer uptake in the stomach during myocardial perfusion imaging when hamburgers were cooked nearby. The results demonstrate significant differences (significant P-values) between women and men for stomach counts but not heart counts.

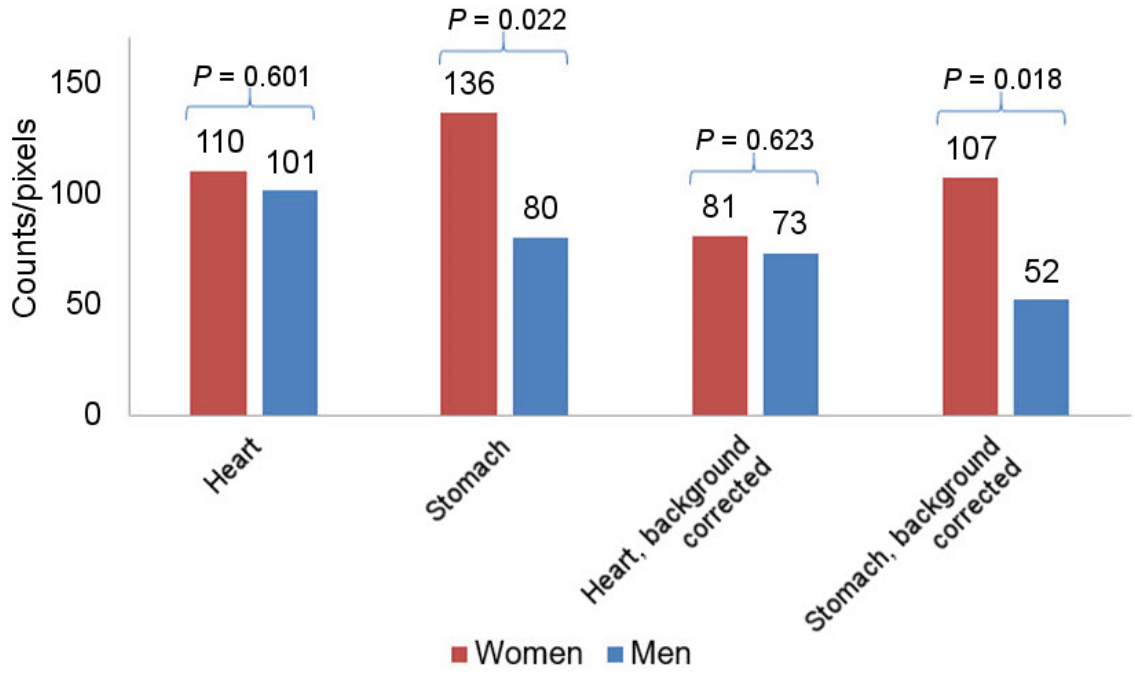

Figure 2. The two graphs demonstrate variation in the data for two samples although the range (100-250 pounds) and mean (175 pounds) were the same for both populations. A) Demonstrates a bell-shaped curve B) Demonstrates a bimodal (two peak) distribution.

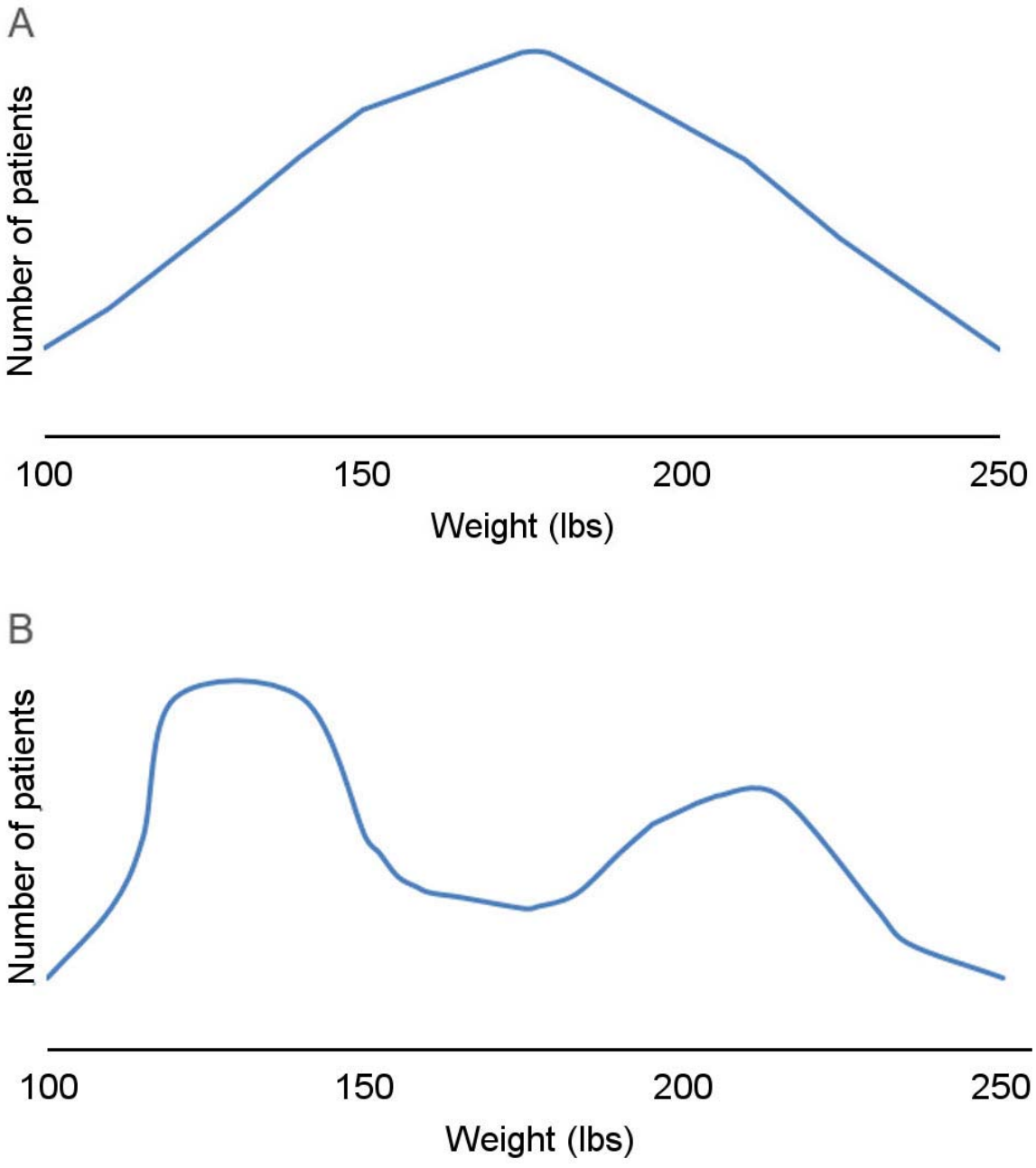

Figure 3. A normal bell-shaped distribution of data. The peak of the curve is the mean (purple line). One standard deviation below and above the mean (green lines) represents 68% of the data. Two standard deviations below and above the mean (orange lines) represents 95% of the data, and three standard deviations below and above the mean (red lines) represents 99.7% of the data.