# I Basic Statistical Concepts

# 1 Understanding Some Basic Statistical Concepts

## 1.1 Sample vs. Population

A concept that may not have been well explained in your statistics class of long ago, is sample vs. population estimates. If your mission is to measure weight on all patients with diabetes in the world (or what statisticians call the "universe"), we must measure the **mean** and **standard deviation** (SD) in a representative **sample** of patients that estimates the mean and variance of weight in the **universe** ($\mu$ or **standard error**, SE). Thus, we make estimates of population weight through a sample of diabetic patients, for example. By examining samples of data, we are always estimating characteristics of the universe. Ideally, these samples are randomly selected from the population and thus, if we were to choose another random sample from the same population, the results would be approximately equal, within the acceptable bounds of random error. This is why the sample selection becomes critical; the sample must be as representative of the population being studied as possible.

## 1.2 General Data Management Considerations

If you have a binary variable, i.e., a variable with only two choices, such as gender, in a dataset, it is necessary to always enter the binary code, or leave it blank if it is missing. Too often I have received datasets for analysis that contain a 1 for "yes" and blank for "no," meaning to me that all of the blanks are missing. However, when I asked the author of the dataset about these missing values: "did the patient have cardiovascular disease (CVD)?," she said "oh, no, a blank means that they did not have the CVD." Those missing fields were corrected to receive a value of 0, while the truly missing

values were simply left blank. Make no assumptions if you see a lot of missing data.

Another tragic example was when a post-doc decided not to consult the codebook to define who was on study drug and who was on placebo, so he switched the assignments, and in his post-doctoral dissertation, he reported that patients on the study drug did not benefit while the controls did benefit, when indeed the reverse was true. Always document variable codes in a codebook or in the database itself.

Laboratory personnel also need to have access to statisticians, or they should possess a foundational and functional level of knowledge in statistics in order to understand, apply, and interpret their laboratory results and keep their instruments calibrated. Or, they are advised to speak to a statistician at the very start of study development. There can be struggles between laboratory personnel and statisticians/epidemiologists when it comes to data handling or interpretation. As one who helped a lab to optimize the performance of their assays, I would sometimes experience comments like, "oh, we eliminated the outliers," or when data are missing for a certain field, they enter a QNS (quantity not sufficient) instead of leaving it blank, or entering dates as month/day/year in some of the fields and then day/month/year in others, thereby throwing off the date format recognition of my analysis software. So, my best advice is to speak to your statistician before designing your study and your database to discuss:

- study design
- developing the hypothesis and the null hypothesis
- sample size. . . sample size. . . sample size. . . sample size. . .
- Also, develop a data codebook (including strict formats for dates!).
- Keep ALL data and don't throw out the outliers!

## 1.3 Central Limit Theorem

Another related concept is the **Central Limit Theorem**, which simply posits that when one continually draws samples from a population and measures their HbA1c levels, for example, the HbA1c values from multiple subjects will eventually take on a normal distribution as one keeps sampling, that is, when plotted the values will take on a bell-shaped distribution that is centered around the mean and the median of the distribution, but only if

the values are capable of being normally distributed in the first place. For example, you wouldn't expect a logarithmically distributed variable to eventually take on a normal distribution after repeated sampling because it will always be logarithmically distributed. More on that in Section 2.2.

## 1.4 Parametric vs. Non-parametric Analyses

We will learn about different types of analyses to perform on different types of data, but the initial question to ask is: "Are the data normally distributed?" If yes, use the parametric statistics toolbox. If they are otherwise, i.e., not normally distributed, use the non-parametric toolbox.

**Parametric statistics** are a set of statistical procedures that are conducted on normally distributed, continuous variables. Parametric statistics are generally more robust than non-parametric statistics, so it becomes understandable why efforts are often made to "normalize" non-normally distributed data (see Chapter 4) before subjecting them to parametric statistics. For example, HIV viral load must be $\log_{10}$ transformed to assimilate a normal distribution before being analyzed using the parametric *T*-test.

**Non-parametric statistics** are a set of statistical procedures that are performed on non-normally distributed data like binary, ordinal, and nominal variables, or on continuous variables that are not normally distributed.

Borrowed and adapted from Tanya Hoskin, a statistician in the Mayo Clinic Department of Health Sciences Research who provides consultations through the Mayo Clinic CTSA BERD Resource,[1] Table 1.1 elucidates which test to use in different circumstances by giving laboratory and clinical examples.

## 1.5 How to Calculate Some Basic Measures of Central Tendency

The mean, median, and mode are common indices used to describe the characteristics of a sample. They are simple to calculate and give some useful information on how sample values, like age and gender, are distributed; the indices can also be used to compare age and gender between different populations. However, the value of these indices has limitations, and misuse can yield misleading information. Following the definitions and the way to calculate the mean, median, and mode (below), two

**Table 1.1** The selection of appropriate statistical tests is dependent on data type

| Analysis type | Example | Parametric procedure | Non-parametric procedure |
|---|---|---|---|
| Compare means between two distinct and independent groups | Mean systolic blood pressure for patients on placebo vs. patients on study drug | Two-sample *T*-test | Wilcoxon rank-sum test |
| Compare two quantitative measurements taken from the same individual | Cell viability before vs. after 3 days in −80F freezer | Paired *T*-test | Wilcoxon signed-rank test |
| Compare means between three or more distinct/ independent groups | We want to compare the baseline ages of 3 groups: drug #1 vs. drug #2 vs. placebo | Analysis of variance (ANOVA) | Kruskal–Wallis test |
| Estimate the degree of association between two quantitative values | Viral particles in urine vs. saliva specimens | Pearson correlation | Spearman–Rank correlation |

examples of populations are illustrated to make the point of the use and misuse of these measures.

### 1.5.1 Mean

The **mean** is calculated as the sum of the values divided by the number of the values. Why are means important? Because they give us a single summary value that describes one measure of the data that is useful for comparing across two or more populations.

Calculation of the mean:

What is the mean of 2, 3, 6, and 10?
**Answer:** $(2+3+6+10)/4 = 21/4 = 5.25$.

However, if you have a distribution such as the following:

What is the mean of 20 + 2 + 500 + 2500?

**Table 1.2** Frequency distribution of Measured Blood Loss

| MBL | | | | |
|---|---|---|---|---|
| | | **Frequency** | **Percent** | **Cumulative Percent** |
| Valid | 1.50 | 1 | 4.3 | 5.3 |
| | 1.54 | 1 | 4.3 | 10.5 |
| | 1.64 | 1 | 4.3 | 15.8 |
| | 1.78 | 1 | 4.3 | 21.1 |
| | 3.55 | 1 | 4.3 | 26.3 |
| | 3.70 | 1 | 4.3 | 31.6 |
| | 3.78 | 1 | 4.3 | 36.8 |
| | 4.27 | 1 | 4.3 | 42.1 |
| | 4.93 | 1 | 4.3 | 47.4 |
| | 6.41 | 1 | 4.3 | 52.6 |
| | 7.21 | 1 | 4.3 | 57.9 |
| | 7.39 | 1 | 4.3 | 63.2 |
| | 7.53 | 1 | 4.3 | 68.4 |
| | 7.84 | 1 | 4.3 | 73.7 |
| | 8.02 | 1 | 4.3 | 78.9 |
| | 8.63 | 1 | 4.3 | 84.2 |
| | 11.28 | 1 | 4.3 | 89.5 |
| | 13.43 | 1 | 4.3 | 94.7 |
| | 68.83 | 1 | 4.3 | 100.0 |
| | Total | 19 | 82.6 | |
| Missing | | 4 | 17.4 | |
| Total | | 23 | 100.0 | |

You might decide that finding the mean of this distribution may be meaningless since there is just so much space between the values. A way to reduce the space is by "normalizing" these values before taking the mean of them (see Section 1.4).

### 1.5.2 Median

The **median** is the midpoint of a frequency distribution where 50% of values fall below it and 50% of values fall above it. The median can be estimated by constructing a frequency distribution table.

As can be seen from Table 1.2, the midpoint of "Measured Blood Loss" can be found by looking at the cumulative frequency and finding that the

50% mark of the distribution falls somewhere between 4.93 and 6.41. These two values are almost equally distant from 50%, so we can approximate the median by taking the mean of these two values: (4.93 + 6.41)/2 = 5.67 = the median.

### 1.5.3 Mode

The **mode** is the number that is repeated most frequently in a distribution. If there is a tie between two values, the distribution is said to be bimodal. If three or more values are tied, it is said to be multimodal. Mode can also be used in cases of ordinal variables like race; which race is most prevalent in the ordinal continuum of the race values?

Calculation of the mode:

What is the mode of 1, 4, 2, 4, 7, 5, 6, 4, 3, 7, 4, 4, 7, 7, and 7?

**Answer:** 4 and 7. This is a bimodal distribution.

Now, let's examine two different samples and see how the mean, median, and mode represent the samples and their usefulness.

#### Mean, Median, and Mode

Group #A: There are 200 individuals in a study that describes blood glucose levels. The data show that the blood glucose values seem to be normally distributed (see Introduction) from 50 to 150 mg/mL. That is, there are persons with low values, mid-values, and high values. If the data are plotted, they show a bell-shaped curve that is normally distributed (Figure 1.1).

Group #B: There are also 200 individuals from a different area and the distribution of blood glucose levels is assessed. In this group, it can be noted that about two-thirds of persons have very low glucose values (<50 mg/mL), while the upper third have very high levels (>150 mg/mL); there are very few persons with mid-range glucose values.

Interpretation of the mean and median in the two groups: when Groups A and B are combined, Population 1 appears to have HbA1c levels that are

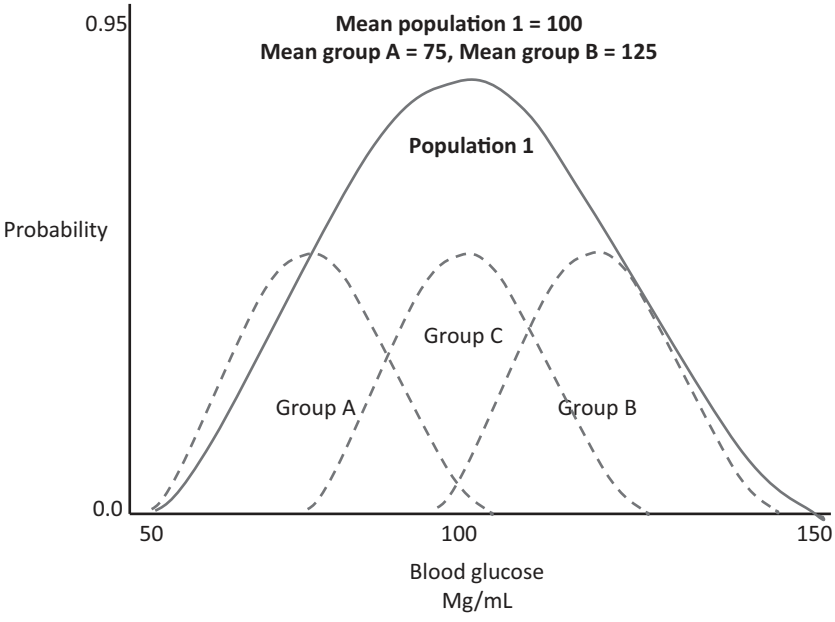


**Figure 1.1** Dissimilar distributions of blood glucose levels.

normally distributed; the median is central and likely lies very close to the mean of 100. So, seeing that the population data are normally distributed, one may automatically consider running parametric analyses. However, when broken into Groups A and B and C, the group means of Groups A and B (i.e., 75 and 125), and likely their group medians, have shifted away from the population mean of 100 and the group distributions also appear to be statistically different from each other. Another visual observation is that Group C is likely not statistically different from Groups A and B. In this situation, a non-parametric statistic should be considered when the distributions of blood glucose in Groups A, B, or C are not normally distributed after being stratified from the population, even though the distribution of Population 1 appears to be normally distributed.

In summary, the characteristics of data are extremely important to understand, and therefore simple measures should not be used exclusively; other statistical tools must be considered to fully and correctly describe the data. These include the standard deviation, the coefficient of variation, the range, and the interquartile range.

### 1.5.4 Range

The lowest number and the highest number of a sorted distribution designates the **range** of the distribution. Ranges are useful when speaking of normal and abnormal ranges for a biological characteristic.

Calculation of the range:

**Example for the Clinician**

Using the Measured Blood Loss distribution in Table 1.2, the smallest number in the distribution of numbers is 1.50 and the largest number is 68.83. That is the range of values in the Measured Blood Loss distribution.

### 1.5.5 Interquartile Range

The **interquartile range** (IQR) is at the 25th and 75th percentile of the distribution. This is a useful set of numbers because it presents a little more information about how the data are distributed at a more granular level than the range. In normal situations, the 25th and 75th percentiles of distributions may not be conveniently obvious, as is shown in Figure 1.2,

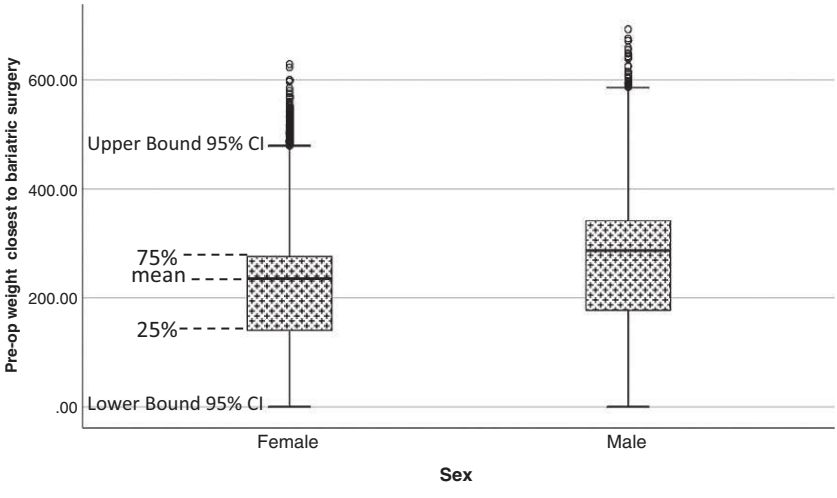


Figure 1.2 Boxplot showing comparison of weight loss by gender.