What an Internist Needs to Know about Statistics

Dale Gardiner MBBS MBioEth, Caroline Hird MBChB MPH

The basics of medical statistics can be readily understood but are often approached by clinicians as if mysterious or forbidding. This may be because statistics was poorly taught at medical school or due to a tendency for articles on statistics to rapidly overcomplicate concepts. It is our hope that this series will be enlightening and provide a solid grounding in the building blocks to all evidence-based medicine.

This series is divided into the three main statistical areas: descriptive statistics as may be used commonly in audit projects, inferential statistics as may be used in therapeutic trials, and diagnostic tests in which sensitivity and specificity are important. Statistical concepts are illustrated with examples predominantly from the critical care literature. The choice of examples, however, should be regarded as non-significant when compared with any personal clinical practice. (There is no p value for this statement.)

In this article, we discuss concepts in descriptive statistics: types of data, measures of central tendency, measures of dispersion, and data distribution. Examine Table 1 and Table 2, an audit data set from a mixed neurological/general intensive care population. It is important to understand how the data are classified before you can interpret the numbers. This has been done for you in the tables. Data can be classified as categorical or numerical, and each of these two can be further subdivided into two subtypes (Figure 1).

**Categorical** data are often non-numerical and consist of data grouped together. Where the groups, or categories, are nominated qualitatively the data are **nominal** (see Figure 1 for examples). There is no value judgment about the worth of each category in nominal data (e.g., being male is not inferior to being female). Nominal data categories are often mutually exclusive. For example, you cannot be both alive and dead.

**Ordinal** data are categorical data where there is a value judgment about the order of the categories. Scoring systems are common examples; they use numbers as labels for categories and to indicate order, and these numbers have no associated units. Two examples are the American Society of Anesthesiologists (ASA) Score and the Glasgow Coma Score (GCS). The higher the ASA Score or the lower the GCS, the worse the patient is, indicating order. The division between each category in ordinal data is not clearly mathematically defined. The ASA Score is assigned by patient history, whereas the GCS is defined by clinical examination. Neither relies on a mathematically measurable variable, nor can you apply simple mathematical analysis, such as multiplication or subtraction. This can be seen more clearly if we consider that it cannot be said that an ASA Score 4 patient is twice as sick as an ASA Score 2 patient. The ASA could have chosen colours (e.g., red for bad, green for good) instead of numbers for its grading system to describe the same ordinal data.

When categorical data are presented in a table, as in Table 1, it is important to understand how the data are classified before you can interpret the numbers. This has been done for you in the tables. Data can be classified as categorical or numerical, and each of these two can be further subdivided into two subtypes (Figure 1).**

**Categorical** data are often non-numerical and consist of data grouped together. Where the groups, or categories, are nominated qualitatively the data are **nominal** (see Figure 1 for examples). There is no value judgment about the worth of each category in nominal data (e.g., being male is not inferior to being female). Nominal data categories are often mutually exclusive. For example, you cannot be both alive and dead.

**Ordinal** data are categorical data where there is a value judgment about the order of the categories. Scoring systems are common examples; they use numbers as labels for categories and to indicate order, and these numbers have no associated units. Two examples are the American Society of Anesthesiologists (ASA) Score and the Glasgow Coma Score (GCS). The higher the ASA Score or the lower the GCS, the worse the patient is, indicating order. The division between each category in ordinal data is not clearly mathematically defined. The ASA Score is assigned by patient history, whereas the GCS is defined by clinical examination. Neither relies on a mathematically measurable variable, nor can you apply simple mathematical analysis, such as multiplication or subtraction. This can be seen more clearly if we consider that it cannot be said that an ASA Score 4 patient is twice as sick as an ASA Score 2 patient. The ASA could have chosen colours (e.g., red for bad, green for good) instead of numbers for its grading system to describe the same ordinal data.

When categorical data are presented in a table, as in Table 1...
1, it is best to give the absolute number or percentage occurring in each category. Occasionally, one sees an article calculating the mean or median for categorical data. Mean and median are measures of central tendency, which should be reserved for numerical data (Table 3). This is because you can’t have a mean or median ethnicity (nominal), and you can’t really have (although it is often spoken as if you can) a mean or median GCS (ordinal). But, for the patient cohort given in Table 1, we can tell you that the calculated mean, median, and mode for GCS is 12.3, 14, and 15, respectively. Since it uses numbers, statistics allows you to calculate many things; you can, but you shouldn’t.

As we have seen above, in ordinal data each value has an uneven separation, and this makes the mean meaningless in this context. (A few outliers will skew the result dramatically; it relies on a detailed understanding of the GCS to interpret – a small difference in GCS may, for a patient, determine a large difference in outcome.) Better to present the data as percentages, as were given in Table 1; 15% of the admissions (where GCS was documented) have severe head injury, as defined by a GCS ≤8 at admission. The art of statistics is in presenting data appropriately and not in stirring the pot in the hope of generating interesting numbers.

**Numerical** data always involves numbers. Numerical data have order, units, and mathematical divisions between each value, allowing simple (and more complex) mathematical analysis. Numerical data are divided into discrete and continuous data (see Figure 1 for examples). **Discrete** data are the simplest form of numerical data. Typically, the values in discrete data have a regular mathematical division from the next (usually whole number data) and can be counted (e.g., the number of patients with ischemic heart disease).

**Continuous** data are the most common type of data in medical settings and research. Continuous data are usually measured (e.g., with a measuring tape, Dinamap, 

---

Table 3. Definition and Measures of Central Tendency

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central tendency</td>
<td>The tendency for a set of values to gather around the middle of the set Generally measured by mean, median, and mode</td>
</tr>
<tr>
<td>Mean</td>
<td>Average $\bar{x} = \frac{\sum x}{n}$ (sum of all values $x$ over the number of values $n$) Should be applied to continuous data if normally distributed</td>
</tr>
<tr>
<td>Median</td>
<td>Middle value of an ordered sample of numerical values Extreme values do not affect the median as much as the mean, for example, length of stay, house prices Usually applied to numerical data (unless normally distributed)</td>
</tr>
<tr>
<td>Mode</td>
<td>Value that occurs most frequently Can be used for skewed numerical data or categorical data</td>
</tr>
</tbody>
</table>

---

Figure 1. Types of data.
thermometer), and there are a potential infinite number of values along a continuum depending on the accuracy of the measuring tool.1,3

When numerical data are presented in a table, as was shown in Table 2, it is good to provide a measure of central tendency (see Table 3) and a measure of dispersion (Table 4). These measures help to describe the data and can provide more insight into the data. Table 2 gives three measures of central tendency and three measure of dispersion for age and length of stay. This is not typical and was given for example purposes only. Usually only the most appropriate measure is given. Typically, for discrete data, the most appropriate measures to describe the data are median and interquartile range, whereas for continuous data, mean and standard deviation are the most appropriate measures.

Numerical data can be summarized graphically in a histogram; this displays the frequency of a value occurring (y-axis) versus the value itself (x-axis) (Figure 2 shows an example using the same data set as Table 2). The difference between a histogram and a bar chart, which is used for graphically summarizing categorical data, is that in a bar chart the distance between each category on the x-axis is not of equal mathematical division. The shape of the histogram illustrates the way the data are distributed and gives guidance as to what sort of statistical technique may be appropriate to apply. If the data fit a “normal” Gaussian distribution, they are often continuous and mean and standard deviation are the appropriate measures to describe the data (Table 5).

The length-of-stay data shown in Figure 2 are skewed and do not fit a normal distribution. This is often the case in medical data sets, and there are a number of statistical options. Generally, it is often best to describe the data using median and interquartile range. Describing the data in this way is usually adequate for most audit and service evaluation investigations, and still allows for some statistical testing of significance if desired (to be described in a future article). Alternatively, statistical techniques exist to make non-normally distributed data more normally distributed (although this is not always appropriate).

Knowing what is appropriate is the art of statistics. There is no doubt that these distinctions and exceptions can be
confusing. Benjamin Disraeli, a 19th century British prime minister, once said, “There are three kinds of lies – lies, damned lies and statistics.” The key is to define the question you want answered and understand how the data were measured. In different circumstances, age, for example, may be continuous or discrete (whole years) or categorical (those less than 16 years categorized as children, those 16–25 as young adults, etc). Even in the simple calculations for length of stay in Table 2, it was first required to decide if an admission of less than 1 full day should be measured as 0 days or 1 day; this decision alters any statistical calculation (using hours instead of days would make the data more continuous but more complicated).

Statisticians have many tools at their disposal for transforming and describing data sets. For the rest of us, Figure 1 is a safe place to start.

### References

### Table 5. Data Distribution

<table>
<thead>
<tr>
<th>Normal or Near-Normal Distributions</th>
<th>Non-normal Distributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Many statistical techniques make assumptions about the data being normal distributed. (although this is not always appropriate).</td>
<td>Statistical techniques exist to make non-normally about the data distributed data more normally distributed</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Normal Gaussian Distribution</th>
<th>Skewed Data Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Normal Gaussian Distribution" /></td>
<td><img src="image" alt="Skewed Data Distribution" /></td>
</tr>
</tbody>
</table>

- **Symmetrical bell-shaped pattern**
- **Many biological processes follow this pattern (e.g., height, weight)**
- **Mean, median, and mode are equal**
- **Mean is toward the longer tail**
- **Mode is toward the shorter tail**
- **Median is between the mean and mode (Figure 2 data are skewed)**

---

Gardiner and Hird