EPI-546: Fundamentals of Epidemiology and Biostatistics

Course Notes - Lecture 2 - Descriptive Statistics

Normality, Abnormality, and Medical Measurement

Mat Reeves, PhD and Michael Brown, MD

I. Normality vs. abnormality

II. Sampling
   a. Random vs. systematic error
   b. Statistical inference

III. Variation in Clinical Data
   a. Biological and measurement variation
   b. Validity and precision

IV. Statistical aspects of variability
   a. Measures of variation
      Standard deviation
   b. Measures of agreement
      Correlation, Kappa

I. Normal vs. Abnormal

Chapter 2 in FF is informative and easy to follow. The online lecture summarizes some of the key concepts.

II. Sampling

It is very difficult if not impossible to obtain data from every member of a population – case in point is the U.S. census which attempts to contact every household in the country. In year 2000, its response rate was only 65%. So a more practical approach is to take a representative sample of the underlying population and then draw inferences about the population from this sample. Taking a sample always involves an element of random variation or error – sampling statistics are essentially about characterizing the nature and magnitude of this random error.

Random error can be defined as the variation that is due to “chance” and is an inherent feature of “sampling” and statistical inference. However, in clinical medicine we also recognize that random error can occur due to the process of measurement and/or the biological phenomenon itself. For example, a given blood pressure measurement may vary because of random error in the measurement tool (sphygmomanometer) and due to the “natural” biological variation in the underlying blood pressure.
While the field of statistics is essentially concerned with the characterization of random error, the degree to which any data is affected by *systematic error or bias* is probably more important. Systematic error can be defined as *any process that acts to distort data or findings from their true value*. In epidemiology, we classify bias as either *selection bias, measurement bias or confounding bias* (and we will refer to these sources of bias frequently during this course and in EPI-547). It is important to note that traditional “statistics” for the most part addresses only random error and NOT systematic bias. Thus, a significant *P* value or a precise confidence interval cannot tell you whether the underlying data is accurate i.e., unbiased.

*Statistical Inference* is the process whereby one draws conclusions regarding a population from the results observed in a sample taken from that population. There are two different but complementary categories of statistical inference: estimation and hypothesis testing. *Estimation* is concerned with estimating the specific value of an unknown population parameter, while *hypothesis testing* is concerned with making a decision about a hypothesized value of an unknown population parameter. These concepts will be further explored in Lecture 4 and in Chapter 10 of FF.

**III. Variation in Clinical Data**

A. *Biological and Measurement Variation*

Clinical information is no different from any other source of data - it has inherent variability which can create substantial difficulties. All types of clinical data whether concerning the patient’s history, physical exam, laboratory results, or response to treatment may change even over the shortest of time intervals. In the broadest sense, variation may be grouped into two categories: variation in the actual entity being measured (*biological variation*) and variation due to the measurement process (*measurement variation*).

1. *Biologic Variation:*

   The causes and origins of biologic variation are endless; variation derives from the dynamic nature of physiology, homeostasis and pathophysiology, as well as genetic differences and differences in the way individuals react to changing environments such as those induced by disease and treatment. Biologic variation can be further sub-divided into within (*intra-person*) and between (*inter-person*) variability. For example, your blood pressure shows a high degree of intra-person variability- it is changing by the hour or even minute in response to many stimuli, such as time of day, posture, physical exertion, emotions, and that last shot of expresso. Biologic variation also occurs because of differences between subjects (inter-person variability). Fortunately there is enough variation between individuals, compared to the degree of within-person variability, that after several repeat blood pressure measurements it is possible to determine the typical (average) blood pressure value of an individual patient and classify them as to their hypertension status i.e., normal, pre-hypertension, or hypertension (Stage I, II or III). Different variables have different amounts of within and between subject variation which can have important clinical consequences.
Regardless of the source of biological variation, its net effect is to add to the level of random error in any measurement process. A common method of reducing the impact of biological variation is to take repeated measurements of a variable or phenomenon - as in the above example of blood pressure. Finally, note that the presence of biological variation is *sine qua non* for epidemiologists to define factors that are associated with disease or outcomes i.e., risk factors. If everyone in the population has the same value or outcome then it is impossible to study the disease process.

2. *Measurement Variation:*
Measurement variation is derived from the measurement process itself. It may be caused by inaccuracy of the instrument (*instrument error*) or the person operating the test (*operator error*). Measurement variation can introduce both *random* error into the data as well as *systematic* error or bias, especially when the test requires some human judgment. Systematic differences between laboratories is one reason that it is vital for each lab to establish its own “reference” ranges. Variation due to different observers reading the same test (e.g., radiologists reading the same x-ray) is referred to as *inter-observer variability*, whereas variability resulting from the same observer reading the same test (e.g., one radiologist reading the same x-ray at different times) is referred to as *intra-observer variability*. Approaches to reduce the impact of measurement bias include the use of specific operational standards e.g., assay standards for laboratory instruments, or the use of explicit operational procedures as in the example of blood pressure measurement (i.e., seated position, appropriate cuff size, identification of specific Korotkoff sounds, repeated measurements etc). Random variation due to measurement variation can again be lessened by taking repeated measurements of a variable or phenomenon.

Note that all variation is additive, so that the net observed variation is a result of the culmination of various individual sources. This is shown nicely by the following figure adapted from the Fletcher text:
B. Validity and Precision

*Validity* refers to the degree to which a measurement process tends to measure what is intended i.e., how accurate is the measurement. A valid instrument will, on average, be close to the underlying true value, and is therefore free of any *systematic error* or *bias*. Graphically, validity can be depicted as a series of related measurements that cluster around the true value (see Fig 2.2). For some clinical data, such as laboratory variables, validity can easily be determined by comparing the observed measurement to an accepted “gold standard”. However, for other clinical data, such as pain, nausea or anxiety there are no obvious gold standard measures available. In this case, it is common to develop instruments that are thought to measure some specific phenomena or *construct*. These constructs are then used to develop a *clinical scale* that can be used to measure the phenomenon in practice. The validity of the instrument or scale can then be evaluated in terms of *content validity* (i.e., the extent to which the instrument includes all of the dimensions of the construct being measured – this is also called *face validity*), *construct validity* (i.e., the degree to which the scale correlates with other known measures of the phenomenon) and *criterion validity* (i.e., the degree to which the scale predicts a directly observable phenomenon).
For dichotomous data, validity is usually expressed in terms of sensitivity and specificity (see lecture 5). There are several different statistical methods for expressing the validity of continuous data, including presenting the mean and standard deviation of the difference between the surrogate measure and the gold standard, as well as correlation and regression analysis.

Fig 2.2 Schematic representation of validity and precision (for you to enjoy completing…)

![Diagram showing valid and precise, valid and imprecise, invalid but precise, invalid and imprecise]

*Precision (or reliability or reproducibility)* refers to the extent that repeated measurements of a phenomenon tend to yield the same results - regardless of whether they are correct or not. There is therefore no comparison to a reference or gold standard measure. Precision refers to the lack of random error - the degree of precision or reliability is inversely related to the amount of random error - the more error the less precise the instrument. Precision is quantified by various standard statistical measures of dispersion such as standard deviation, variance and range. Graphically, precision can be depicted as the degree to which a series of related measurements cluster together (Fig 2.2).

Random variation can be classified according to whether there is one or multiple observers or instruments i.e., *intra-observer variability* vs. *inter-observer variability*, respectively. Using the target analogy of Fig 2.2, inter-observer reliability refers to the scatter from different observers shooting at the same target, while intra-observer reliability refers to the scatter of shots from one shooter.

For measurements that do not involve a direct observation e.g., self-administered questionnaires, reliability can be assessed using the test-retest method, where respondents answer the same question at two different times. This approach measures a form of intra-observer reliability, where the respondent is acting as the same observer on two separate occasions. The exact statistical approach used to quantify reliability depends on the type of data measured — Kappa for categorical data and intra-class correlation for interval data.
IV. Statistical aspects of variability

A. Measures of variation

1. Variance and Standard Deviation

The variability or precision of a measurement is expressed by the standard deviation (SD). The SD represents the absolute value of the average difference of individual values from the mean, and is calculated by taking the square root of the variance.

\[ SD = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} \]

Assuming a normal distribution, one standard deviation either side of the mean includes 68% of the total number of observations, while 2 SD’s include 95%.

Example:

The SD of serum cholesterol is 15 mg/dl. Americans now have an average cholesterol value of 205 mg/dl, thus the values for the middle 68% of the population would be expected to vary from 190 to 220 mg/dl (mean ± 1 SD)

B. Measures of agreement

1. Correlation (r)

The reliability of a continuous measurement (i.e., interval data) can be expressed by the correlation coefficient (r) between two sets of measurements. The correlation coefficient (r) measures the strength of the linear relationship between two continuous variables: r ranges from -1 to +1, with zero representing no relationship.

If information can be obtained on the actual true values, then the correlation can be regarded as a test of validity between the “truth” and an imperfect measurement. However, true values are rarely available, so in most cases, the correlation between two measures assesses reliability i.e., the extent to which the results can be replicated by two measurements.

If measures are obtained from two observers, then the extent of agreement between the two would reflect the between-observer variability (or reliability). However, it should be noted that it is possible to have high values of r, yet have little direct agreement between two observers or instruments. For example, in measuring blood cholesterol, a perfect r (1.0) can occur if laboratory A results are always exactly 10 mg/dl points higher than laboratory B. The correlation co-efficient is also commonly used as a measure of reliability in test-retest studies - where the same instrument is applied to the same population at a later time (a measure of intra-rater or within-person variability).
2. **Categorical data - Kappa**

For categorical or qualitative data, reliability can be characterized using the *kappa statistic* \((k)\). Kappa has the useful property of correcting for the degree of chance in the overall level of agreement, and is therefore preferred over other measures like the commonly used *overall percent agreement*. The ability of kappa to adjust for chance agreement is especially important in clinical data, because the prevalence of the particular condition being evaluated affects the likelihood that observers will agree purely due to chance. This chance agreement must be adjusted for, otherwise false reassurances can occur. As an example of this phenomenon, if 2 people each repeatedly toss a coin, there are only 4 possible results i.e., HH (i.e., head, head), TT (i.e., tail, tail), HT, and TH. The probability \((p)\) of each result is \(\frac{1}{4}\), so the overall agreement between the two coins (due to chance alone) is 0.5 (i.e., sum of \(p(\text{HH})\) and \(p(\text{TT})\)). The influence of the underlying prevalence of the attribute or condition being measured on the overall percent agreement is shown in the following table:

**Overall percent agreement due to chance for a binary attribute**

<table>
<thead>
<tr>
<th>Prevalence of the attribute</th>
<th>Overall percent agreement*</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>82%</td>
</tr>
<tr>
<td>0.3</td>
<td>58%</td>
</tr>
<tr>
<td>0.5</td>
<td>50%</td>
</tr>
<tr>
<td>0.7</td>
<td>58%</td>
</tr>
<tr>
<td>0.9</td>
<td>82%</td>
</tr>
</tbody>
</table>

* \(P(e)\) calculated by multiplying the marginal totals of a 2x2 table

The kappa \((k)\) statistic is calculated as:

\[
k = \frac{P_o - P_e}{1 - P_e}
\]

where, \(P_o\) = the total proportion of observations on which there is agreement.

\(P_e\) = the proportion of agreement expected by chance alone.

Thus \(k\) is the ratio of the actual agreement attributable to the reproducibility of the observations \((i.e., P_o - P_e)\), compared to the maximum possible value \((1 - P_e)\). Or,

\[
k = \frac{\text{Actual agreement beyond chance}}{\text{Potential agreement beyond chance}}
\]

The following diagram explains the logic of the kappa statistic. In this example chance corrected agreement (Kappa) = \((0.75 - 0.50)/(1 - 0.50) = 0.25/0.50 = 50\%\).
The simplest clinical application of Kappa is in the measurement of inter-rater agreement whereby two observers evaluate the same series of patients and classify them according to some particular dichotomous condition (e.g., disease present or absent). As an example, the following data is generated from 2 radiologists who independently reviewed 150 mammograms and classified each patient as to whether they had an abnormality:

<table>
<thead>
<tr>
<th>OBSERVER B</th>
<th>OBSERVER A</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>69</td>
<td>15</td>
</tr>
<tr>
<td>No</td>
<td>18</td>
<td>48</td>
</tr>
<tr>
<td>TOTALS</td>
<td>87</td>
<td>63</td>
</tr>
</tbody>
</table>

Observer A thought 87 patients (or 58%) had an abnormality, while Observer B thought 84 patients did (i.e., 56%), but they agreed only 78% of the time (the observed proportion of agreement \( P_o \) is calculated as \((69 + 48)/150 = 0.78\) or 78%). However, the proportion of agreement expected due to chance \( P_e \) was 0.51 or 51%, which is estimated by calculating the “expected” numbers in cells a and d from the product of the marginal totals (i.e., 87 \times 84/150 = 48.72 [for cell a], and 63 \times 66/150 = 27.72 [for cell d], and then calculating the proportion of agreement by dividing the sum of these two cells by the total number of subjects i.e., 48.72 + 27.72 / 150, which equals 0.5096 or 51%). Thus \( k \) can be estimated as:

\[
k = \frac{P_o - P_e}{1 - P_e} = \frac{0.78 - 0.51}{1 - 0.51} = 0.27 = 0.55 \text{ or } 55%
\]

Michael Brown MD, Mathew Reeves, PhD
© Department of Epidemiology, Michigan State Univ.
Like the correlation coefficient, kappa varies in value from -1 to +1, however the interpretation is different. A value of zero denotes agreement that is no better than chance, while a negative value denotes agreement that is worse than chance (i.e., fulminant disagreement!). The following guide to interpreting the strength of agreement shown by kappa has been developed. In our example - 55% represents a moderate degree of agreement between the two radiologists.

<table>
<thead>
<tr>
<th>Value of k</th>
<th>Strength of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0</td>
<td>Poor</td>
</tr>
<tr>
<td>0 - 0.20</td>
<td>Slight</td>
</tr>
<tr>
<td>0.21 - 0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41 - 0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61 - 0.80</td>
<td>Substantial</td>
</tr>
<tr>
<td>0.81 - 1.0</td>
<td>Almost perfect</td>
</tr>
</tbody>
</table>