

## Jottings

**D**id you notice that the EBM website has been updated? If not, please have a look. Some careful research by the BMJ on what folk use and how they use it has led to a much more functional website. It may not look more beautiful, but navigating is much easier, more intuitive, and faster. For example, you can now search from the home page and get to the full search engine in 1 click. And the entire homepage provides always 1-click access to most of the important functions and links. Check the Online Tools: in addition to the emailed table of contents, there is now a citation tracker that allows you to track particular articles or issues. So please have a look—suggestions welcome.

Of course the journal is not the only website that discusses evidence-based medicine. If you type evidence-based medicine into Google, you get a staggering 3 million+ entries. The number one entry is the Centre for Evidence-Based Medicine—[www.cebm.net](http://www.cebm.net). From here you can download various EBM tools, including a full (and free) version of CATmaker and the CEBM workshops schedule.

The core of the EBM journal is the abstracts: the valid and clinically relevant new research that we find from over 100

screened journals. However, we also aim to help readers with developments and ideas about the practice and teaching evidence-based medicine. One area of repeated requests is more about understanding statistics. So in this issue is the first of a new series of short and painless notes on statistics that we hope will be useful for both readers and teachers. The series aims to introduce the basic principles of statistics from the perspective of a reader rather than a doer of research. So we will focus on principles and meaning rather than formula and calculations. This first article provides an overview of measurement scales and their summary statistics. Subsequent notes will focus on measures of association (eg, relative risk and odds ratio), statistical inference using confidence intervals and p values, precision and bias (including sample size and power), the evaluation of diagnostic tests (predictive values), and survival analysis. Let us know if there is a statistical issue you would particularly like covered in the series.

THE EDITORS

## Measurement scales and their summary statistics

### MEASUREMENT SCALES

In any study, observations on each individual are made. These observations vary both between and within individuals and are thus referred to as “variables.” We may summarise the data collected in a study either numerically, in the form of summary statistics, or in tabular or graphical form. The advantage of the first is that individual statistics (such as means or proportions) can be used to summarise the data simply; on the other hand all, or most, of the data can be presented in a table or figure. The appropriate summary method (as well as the statistical analysis) depends on the type of variable and its measurement scale. For example, only if the distribution is approximately normal (symmetrical and bell-shaped) should the mean be used to summarise the data.

There are essentially 2 main types of measurement scales: categorical and numerical (table 1). Categorical variables have a set of labels for category membership (eg, diabetic, non-diabetic); numerical variables are either a count (eg, number of GP visits), a measure on a particular instrument (eg, blood pressure), or a summary score (eg, SF-36 score).

### Tabular and graphical presentation

Tables and graphs can present a distribution simply (table 1). For a single categorical variable, the frequency of observations in each category can be tabulated. The graphical equivalent is a bar graph or bar chart. For a numerical

variable, a histogram is the simplest way of presenting the data. In order to present the data in a table, unless the scale is very narrow, categories will need to be created representing the number of observations within particular group intervals. The number of observations within each interval is presented in the frequency table, allowing the calculation of both relative frequency (the percentage of observations in each category) and cumulative relative frequency (the percentage of observations in that category or below it).

### Summary statistics

Both categorical and numerical data can be summarised using summary statistics (table 2). Appropriate summary statistics for categorical data are the number of observations, and their proportion or percentage, in each category. Numerical data are summarised using an “average” value, such as the mean or median, together with a measure of the spread of the observations around this value, such as the range or standard deviation. The mode is only rarely used. The mean and the standard deviation are the most informative measures since they use all the data in their calculation. They should, however, only be used for normally distributed numerical variables since any skewness in the data will also distort the values of the mean and the standard deviation. Non-normally distributed variables should be summarised using the median and either the range or interquartile range.

Table 1. Types of variable and the graphical and tabular presentation of individual variables (univariate data)

Variable type	Description	Graphical presentation	Tabular presentation																													
Categorical	<p><b>Dichotomous measure</b>—describes presence or absence of an attribute (eg, pregnant/not pregnant; smoker/non-smoker; male/female)</p>	<p><b>Bar Graph</b></p> <p>Frequency of males and females</p>	<p><b>Frequency table</b> The list of possible values and the number of times each occurs.</p> <p>The <i>relative frequency</i> is the percentage of times that each value occurs.</p>																													
	<p><b>Nominal data</b>—several categories with no natural order (eg, blood groups—A/B/AB/O; marital status—married/single/divorced/separated/widowed).</p> <p><b>Ordinal data</b>—several categories with a natural order but with no clear “units,” (eg, depressive episode: mild/moderate/severe/severe with psychotic symptoms).</p>	<p><b>Bar Graph</b></p> <p>Frequency of depressive episode</p>	<p><b>Dichotomous</b></p> <table border="1"> <thead> <tr> <th>Sex</th> <th>Frequency</th> <th>Relative frequency</th> </tr> </thead> <tbody> <tr> <td>Male</td> <td>50</td> <td>37.0%</td> </tr> <tr> <td>Female</td> <td>85</td> <td>63.0%</td> </tr> <tr> <td>Total</td> <td>135</td> <td>100.0%</td> </tr> </tbody> </table> <p><b>Nominal/ordinal</b></p> <table border="1"> <thead> <tr> <th>Depressive episode</th> <th>Frequency</th> <th>Relative frequency</th> </tr> </thead> <tbody> <tr> <td>Mild</td> <td>24</td> <td>30.0%</td> </tr> <tr> <td>Moderate</td> <td>30</td> <td>37.5%</td> </tr> <tr> <td>Severe</td> <td>17</td> <td>21.3%</td> </tr> <tr> <td>Severe +</td> <td>9</td> <td>11.3%</td> </tr> <tr> <td>Total</td> <td>80</td> <td>100.0%</td> </tr> </tbody> </table>	Sex	Frequency	Relative frequency	Male	50	37.0%	Female	85	63.0%	Total	135	100.0%	Depressive episode	Frequency	Relative frequency	Mild	24	30.0%	Moderate	30	37.5%	Severe	17	21.3%	Severe +	9	11.3%	Total	80
Sex	Frequency	Relative frequency																														
Male	50	37.0%																														
Female	85	63.0%																														
Total	135	100.0%																														
Depressive episode	Frequency	Relative frequency																														
Mild	24	30.0%																														
Moderate	30	37.5%																														
Severe	17	21.3%																														
Severe +	9	11.3%																														
Total	80	100.0%																														
Numerical	<p>Discrete variables are observations that can only take whole numerical values, e.g. 0, 1, 2, 3, 4, 5, 6+ migraines last month. They may be transformed to an ordinal variable (eg, 0–1, 2–3, 4–5, 6+ migraines last month). Ideally, a histogram of discrete data should use non-touching lines rather than touching bars so as not to imply continuity.</p>	<p><b>Histogram</b></p> <p>MIGRAINE</p>	<p><b>Frequency distribution</b> Values are grouped into non-overlapping intervals.</p> <p>The <i>cumulative relative frequency</i> is the percentage of observations that have a value either in that interval or below it. For example, in the adjacent table, 72% of the individuals report fewer than 6 migraines/month while 22% report fewer than 2.</p>																													
	<p>Continuous variables are observations where the only restriction is the accuracy of the measuring instrument (eg, height, blood pressure, BMI). They may be transformed to a categorical variable (eg, BMI &lt; 20, 20–24.9, 25–29.9, 30+).</p>	<p><b>Histogram</b></p> <p>fasting blood insulin (mmol/l)</p>	<p><b>Discrete/continuous</b></p> <table border="1"> <thead> <tr> <th>Migraines last month</th> <th>Frequency (Relative frequency)</th> <th>Cumulative frequency</th> </tr> </thead> <tbody> <tr> <td>1–0</td> <td>16 (22.2%)</td> <td>22.2%</td> </tr> <tr> <td>2–3</td> <td>22 (30.6%)</td> <td>52.8%</td> </tr> <tr> <td>4–5</td> <td>14 (19.4%)</td> <td>72.2%</td> </tr> <tr> <td>6–7</td> <td>13 (18.1%)</td> <td>90.3%</td> </tr> <tr> <td>8+</td> <td>7 (9.7%)</td> <td>100.0%</td> </tr> <tr> <td>Total</td> <td>72 (100.0%)</td> <td></td> </tr> </tbody> </table>	Migraines last month	Frequency (Relative frequency)	Cumulative frequency	1–0	16 (22.2%)	22.2%	2–3	22 (30.6%)	52.8%	4–5	14 (19.4%)	72.2%	6–7	13 (18.1%)	90.3%	8+	7 (9.7%)	100.0%	Total	72 (100.0%)									
Migraines last month	Frequency (Relative frequency)	Cumulative frequency																														
1–0	16 (22.2%)	22.2%																														
2–3	22 (30.6%)	52.8%																														
4–5	14 (19.4%)	72.2%																														
6–7	13 (18.1%)	90.3%																														
8+	7 (9.7%)	100.0%																														
Total	72 (100.0%)																															

Table 2. Numerical summary measures—of location (central tendency) and spread

	Summary measure	Description	Graphical presentation	Comments
Measures of central tendency	<b>Mean</b>	Average value: the sum of all the observations divided by the number of observations.		<p>The mean is sensitive to skewness and to outliers (extreme values) since it uses all values in its calculation. Thus, for a distribution that is skewed to the right (as is the distribution of fasting blood insulin shown alongside) the mean is larger than the median, being itself skewed to the right by the extreme values in the tail. For a distribution that is skewed to the left the mean is accordingly smaller than the median. Thus, only for approximately symmetrical distributions (or for distributions that have been transformed to Normality) is the mean a good measure of central tendency.</p>
	<b>Median</b>	Middle value—50 <sup>th</sup> percentile—when all observations are ordered from smallest to largest.		
	<b>Mode</b>	Most common value. More appropriate for categorical data. Continuous data often have no mode (and sometimes more than one mode).		
Measure of dispersion or spread	<b>Range</b>	Difference between the largest value and the smallest value.		<p>The range is very sensitive to skewness and to outliers. Generally increases with sample size.</p> <p>The interquartile range is a better measure of spread when the distribution is skewed or when there are outliers. Is less sensitive to sample size.</p> <p>The standard deviation is also sensitive to skewness and to outliers. In a Normal distribution the standard deviation should be no more than half the size of the mean—this provides a good indication of Normality.</p>
	<b>Interquartile range</b>	Difference between 75 <sup>th</sup> percentile and 25 <sup>th</sup> percentile. Encompasses the middle 50% of the observations.		
	<b>Standard deviation</b>	Quantifies the amount of spread or variability about the mean—the average deviation from the mean. The variance is the square of the standard deviation.		