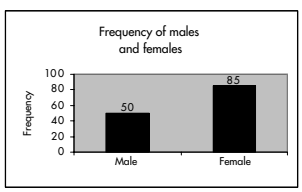
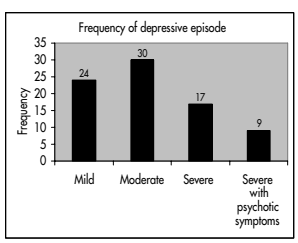
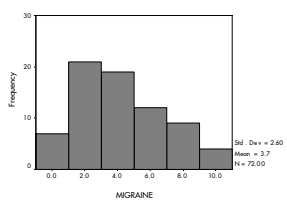
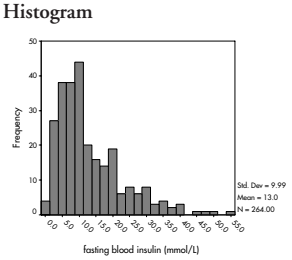
Introduction to Biostatistics: Part 1. Measurement scales and their summary statistics

This is the first in a series of editorials that will be appearing in *ACP Journal Club*. The aim of the series is to introduce readers to the basic principles of statistics to enable effective evaluation of research evidence and data presented in clinical journals. This first article provides an overview of measurement scales and their summary statistics. Subsequent notes will focus on measures of association (e.g., absolute and relative risk), measuring statistical uncertainty using confidence intervals and *P* values, precision and bias (including sample size and power), the evaluation of diagnostic tests (predictive values), correlation and regression (interpreting scatter plots and coefficients), meta-analysis (interpreting forest plots), and survival analysis.

MEASUREMENT SCALES

In any study, observations are made on each individual. These observations vary both between and within individuals and are thus referred to as "variables." We may summarize 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 method is that individual statistics (such as means or proportions) can be used to summarize 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 summarize the data.

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

	Variable type	Description	Graphical presentation	Tabular presentation	
Categorical	2 categories	**Dichotomous measure**—describes presence or absence of an attribute (e.g., pregnant/not pregnant; smoker/nonsmoker; male/female)	**Bar graph** 	**Frequency table** The list of possible values and the number of times each occurs. The relative frequency is the percentage of times that each value occurs.	**Dichotomous** Sex / Frequency / Relative frequency Male 50 37.0% Female 85 63.0% Total 135 100.0%
	> 2 categories (**nominal or ordinal**)	**Nominal data**—several categories with no natural order (e.g., blood groups—A/B/AB/O; marital status—married/single/divorced/separated/widowed) **Ordinal data**—several categories with a natural order but with no clear "units" (e.g., depressive episode; mild/moderate/severe/severe with psychotic symptoms)	**Bar graph** 		**Nominal/ordinal** 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	**Discrete**	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 (e.g., 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 to avoid implying continuity.	**Histogram** 	**Frequency distribution** Values are grouped into nonoverlapping intervals. The *cumulative* relative frequency 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/mo while 22% report fewer than 2.	**Discrete/continuous** Migraines last month / Frequency (relative frequency) / Cumulative relative frequency 0-1 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%)
	Continuous	Continuous variables are observations where the only restriction is the accuracy of the measuring instrument (e.g., height, blood pressure, body mass index [BMI]). They may be transformed to a categorical variable (e.g., BMI <20, 20–24.9, 25–29.9, 30+).	**Histogram** 		

The 2 main types of measurement scales are categorical and numerical (Table 1). Categorical variables have a set of labels for category membership (e.g., diabetic and nondiabetic); numerical variables are a count (e.g., number of physician visits), a measure on a particular instrument (e.g., blood pressure), or a summary score (e.g., SF-36 score).

TABULAR AND GRAPHICAL PRESENTATION

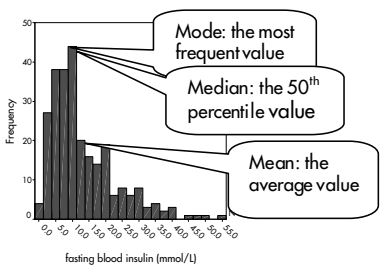
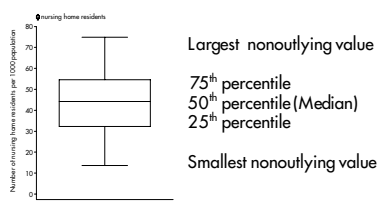
Tables and graphs can present a distribution simply (Table 1). For a single categorical variable, the frequency of observations in each category can be tabulated. The graphical equivalent is a bar graph or bar chart. For a numerical variable, a histogram is the simplest way to present the data. To present the data in a table, unless the scale is very narrow, categories 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 summarized 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 summarized 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 range or standard deviation. The mode is only rarely used. The mean and 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 (see Comments in Table 2) also distorts the values of the mean and standard deviation. Nonnormally distributed variables should be summarized using the median and either the range or interquartile range.

Stuart Carney, MB, ChB, MPH, MRCPsych
Department of Psychiatry
Helen Doll, BSc, Dip App Stats, MSc
Department of Public Health
University of Oxford
Oxford, England, UK

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

	Summary measure	Description	Graphical presentation	Comments
Measures of central tendency	**Mean**	Average value: The sum of all the observations divided by the number of observations	**Histogram** 	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.
	Median	Middle value: 50th percentile when all observations are ordered from smallest to largest		
	Mode	Most common value: More appropriate for categorical data. Continuous data often have no mode (and sometimes have > 1 mode).		
Measures of dispersion or spread	**Range**	Difference between the largest value and the smallest value	**Box plot** 	The range is very sensitive to skewness and to outliers. Generally increases with sample size.
	Interquartile range	Difference between 75th percentile and 25th percentile. Encompasses the middle 50% of the observations.		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.
	Standard deviation	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.		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.