PEARSON NEW INTERNATIONAL EDITION

# Statistics for the Behavioral and Social Sciences: A Brief Course

Aron Coups Aron

Fifth Edition

ALWAYS LEARNING

PEARSON

# Pearson New International Edition

Statistics for the Behavioral
and Social Sciences: A Brief Course

Aron Coups Aron

Fifth Edition

PEARSON

**Pearson Education Limited**
Edinburgh Gate
Harlow
Essex CM20 2JE
England and Associated Companies throughout the world

*Visit us on the World Wide Web at:* www.pearsoned.co.uk



PEARSON

ISBN 10: 1-292-02304-X
ISBN 13: 978-1-292-02304-5

**British Library Cataloguing-in-Publication Data**
A catalogue record for this book is available from the British Library

Printed in the United States of America

# Table of Contents

Glossary
Arthur Aron/Elaine N. Aron/Elliot Coups **1**

Glossary of Symbols
Arthur Aron/Elaine N. Aron/Elliot Coups **7**

1. Displaying the Order in a Group of Numbers Using Tables and Graphs
Arthur Aron/Elaine N. Aron/Elliot Coups **9**

2. The Mean, Variance, Standard Deviation, and z Scores
Arthur Aron/Elaine N. Aron/Elliot Coups **39**

3. Correlation and Prediction
Arthur Aron/Elaine N. Aron/Elliot Coups **75**

4. Some Key Ingredients for Inferential Statistics: The Normal Curve, Sample versus Population, and Probability
Arthur Aron/Elaine N. Aron/Elliot Coups **135**

5. Introduction to Hypothesis Testing
Arthur Aron/Elaine N. Aron/Elliot Coups **163**

6. Hypothesis Tests with Means of Samples
Arthur Aron/Elaine N. Aron/Elliot Coups **193**

7. Making Sense of Statistical Significance: Effect Size and Statistical Power
Arthur Aron/Elaine N. Aron/Elliot Coups **225**

8. Introduction to the t Test: Single Sample and Dependent Means
Arthur Aron/Elaine N. Aron/Elliot Coups **259**

9. The t Test for Independent Means
Arthur Aron/Elaine N. Aron/Elliot Coups **305**

10. Introduction to the Analysis of Variance
Arthur Aron/Elaine N. Aron/Elliot Coups **345**

11. Chi-Square Tests and Strategies When Population Distributions Are Not Normal
Arthur Aron/Elaine N. Aron/Elliot Coups **401**

12. Applying Statistical Methods in Your Own Research Project
Arthur Aron/Elaine N. Aron/Elliot Coups **461**

References
Arthur Aron/Elaine N. Aron/Elliot Coups **475**

Index **481**

# Glossary

**Alpha ($\alpha$).** Probability of a Type I error; same as *significance level*.

**Analysis of variance (ANOVA).** Hypothesis-testing procedure for studies with three or more groups.

**Assumption.** A condition, such as a population's having a normal distribution, required for carrying out a particular hypothesis-testing procedure; a part of the mathematical foundation for the accuracy of the tables used in determining cutoff values.

**Beta ($\beta$).** Same as *standardized regression coefficient*.

**Between-groups degrees of freedom ($df_{Between}$).** Degrees of freedom used in the between-groups estimate of the population variance in an analysis of variance (the numerator of the $F$ ratio); number of scores free to vary (number of means minus 1) in figuring the between-groups estimate of the population variance; same as *numerator degrees of freedom*.

**Between-groups estimate of the population variance ($S^2_{Between}$).** In an analysis of variance, estimate of the variance of the population of individuals based on the variation among the means of the groups studied.

**Between-subjects design.** Research strategy in which each person is tested only once and groups of individuals are compared as, for example, in a $t$ test for independent means.

**Biased estimate.** Estimate of a population parameter that is likely systematically to overestimate or underestimate the true value of the population parameter. For example, $SD^2$ would be a biased estimate of the population variance (it would systematically underestimate it).

**Bimodal distribution.** Frequency distribution with two approximately equal frequencies, each clearly larger than any of the others.

**Categorical variable.** Same as *nominal variable*.

**Ceiling effect.** Situation in which many scores pile up at the high end of a distribution (creating skewness to the left) because it is not possible to have a higher score.

**Cell.** In a factorial research design, a particular combination of levels of the variables that divide the groups. In chi-square, the particular combination of categories for two variables in a contingency table.

**Cell mean.** Mean of a particular combination of levels of the variables that divide the groups in a factorial design in analysis of variance.

**Central limit theorem.** Mathematical principle that the distribution of the sums (or means) of scores taken at random from any distribution of individuals will tend to form a normal curve.

**Central tendency.** Typical or most representative value of a group of scores, such as the mean, median, or mode.

**Change score.** After score minus before score. A kind of *difference score*.

**Chi-square distribution.** Mathematically defined curve used as the comparison distribution in chi-square tests; the distribution of the chi-square statistic.

**Chi-square statistic ($\chi^2$).** Statistic that reflects the overall lack of fit between the expected and observed frequencies; the sum, over all the categories or cells, of the squared difference between observed and expected frequencies divided by the expected frequency.

**Chi-square table.** Table of cutoff scores on the chi-square distribution for various degrees of freedom and significance levels.

**Chi-square test.** Hypothesis-testing procedure that uses the chi-square distribution as the comparison distribution.

**Chi-square test for goodness of fit.** Hypothesis-testing procedure that examines how well an observed frequency distribution of a single nominal variable fits some expected pattern of frequencies.

**Chi-square test for independence.** Hypothesis-testing procedure that examines whether the distribution of frequencies over the categories of one nominal variable are unrelated to (independent of) the distribution of frequencies over the categories of a second nominal variable.

**Comparison distribution.** Distribution used in hypothesis testing. It represents the population situation if the null hypothesis is true. It is the distribution to which you compare the score based on your sample's results. It is made up of the same kinds of numbers as those of the sample's results (such as a sample mean, a difference between sample means, $F$ ratios, or chi-squares).

**Computational formula.** Equation mathematically equivalent to the definitional formula. It is easier to use when figuring by hand, but does not directly show the meaning of the procedure.

**Confidence interval.** Roughly speaking, the region of scores (that is, the scores between an upper and lower value) that is likely to include the true population mean; more precisely, the range of possible population means from which it is not highly unlikely that you could have obtained your sample mean.

**Confidence limit.** Upper or lower value of a confidence interval.

**Contingency table.** Two-dimensional chart showing frequencies in each combination of categories of two nominal variables, as in a chi-square test for independence.

**Conventional levels of significance ($p < .05, p < .01$).** The levels of significance widely used in the behavioral and social sciences.

**Correlation.** Association between scores on two variables.

**Correlation coefficient (*r*).** Measure of the degree of linear correlation between two variables, ranging from −1 (a perfect negative linear correlation) through 0 (no correlation) to +1 (a perfect positive linear correlation); average of the cross-products of *Z* scores of two variables; square root of the proportion of variance accounted for.

**Correlation matrix.** Common way of reporting the correlation coefficients among several variables in a research article; table in which the variables are named on the top and along the side and the correlations among them are all shown (only half of the resulting square, above or below the diagonal, is usually filled in, the other half being redundant).

**Cramer's phi.** Measure of association between two nominal variables; effect-size measure for a chi-square test for independence used with a contingency table that is larger than 2 × 2; square root of result of dividing the chi-square statistic by the product of the number of participants multiplied by the degrees of freedom of the smaller side of the contingency table; also known as *Cramer's V* and sometimes written as $\phi_c$ or $V_c$.

**Criterion variable (usually *Y*).** In prediction (regression), a variable that is predicted; sometimes called *dependent variable*.

**Cross-product of *Z* scores.** The result of multiplying a person's *Z* score on one variable by the person's *Z* score on another variable; for a group of individuals, the average of the cross-products of *Z* scores between two variables is the correlation coefficient for those two variables.

**Curvilinear correlation.** Relationship between two variables that shows up on a scatter diagram as dots following a systematic pattern that is not a straight line; any association between two variables other than a linear correlation.

**Cutoff sample score.** In hypothesis testing, the point on the comparison distribution at which, if reached or exceeded by the sample score, you reject the null hypothesis; sometimes called *critical value*.

**Data transformation.** Mathematical procedure (such as taking the square root) used on each score in a sample, usually done to make the sample distribution closer to normal.

**Decision error.** Incorrect conclusion in hypothesis testing in relation to the real (but unknown) situation, such as deciding the null hypothesis is false when it is really true.

**Definitional formula.** Equation for a statistical procedure directly showing the meaning of the procedure.

**Degrees of freedom (*df*).** Number of scores free to vary when estimating a population parameter; usually part of a formula for making that estimate—for example, in the formula for estimating the population variance from a single sample, the degrees of freedom is the number of scores minus 1.

**Denominator degrees of freedom ($df_{Within}$).** Same as *within-groups degrees of freedom*.

**Dependent variable.** Variable considered to be an effect.

**Descriptive statistics.** Procedures for summarizing a group of scores or otherwise making them more understandable.

**Deviation score.** Score minus the mean.

**Difference score.** Difference between a person's score on one testing and the same person's score on another testing; often an after score minus a before score, in which case it is also called a *change score*.

**Dimension.** In a factorial design, one of the grouping variables that is crossed with another grouping variable; in a contingency table, one of the nominal variables.

**Direction of causality.** Path of causal effect; if *X* is thought to cause *Y*, then the direction of causality is from *X* to *Y*.

**Directional hypothesis.** Research hypothesis predicting a particular direction of difference between populations—for example, a prediction that the population from which the sample studied was drawn has a higher mean than the population in general.

**Distribution of differences between means.** Distribution of differences between means of pairs of samples such that for each pair of means, one is from one population and the other is from a second population; the comparison distribution in a *t* test for independent means.

**Distribution of means.** Distribution of means of samples of a given size from a particular population (also called a sampling distribution of the mean); comparison distribution when testing hypotheses involving a single sample of more than one individual.

**Distribution-free test.** Hypothesis-testing procedure making no assumptions about the shape of the populations; approximately the same as a *nonparametric test*.

**Effect size.** In studies involving means of one or two groups, measure of difference (lack of overlap) between populations; the usual standardized effect size measure used in the behavioral and social sciences increases with greater differences between means and decreases with greater standard deviations in the populations, but it is not affected by sample size. There are also conventional effect size measures for other kinds of studies (correlations, analysis of variance, and chi-square test situations); these describe the standardized degree of association in the population.

**Effect size conventions.** Standard rules about what to consider a small, medium, and large effect size, based on what is typical in behavioral and social science research; also known as Cohen's conventions.

**Equal-interval variable.** A variable in which the numbers stand for approximately equal amounts of what is being measured.

**Expected frequency (*E*).** In a chi-square test, number of people in a category or cell expected if the null hypothesis were true.

**Expected relative frequency.** In figuring probabilities, number of successful outcomes divided by the number of total outcomes you would expect to get if you repeated an experiment a large number of times.

***F* distribution.** Mathematically defined curve that is the comparison distribution used in an analysis of variance; distribution of *F* ratios when the null hypothesis is true.

***F* ratio.** In analysis of variance, ratio of the between-groups population variance estimate to the within-groups population variance estimate; score on the comparison distribution (an *F* distribution) in an analysis of variance; also referred to simply as *F*.

***F* table.** Table of cutoff scores on the *F* distribution for various degrees of freedom and significance levels.

**Factorial analysis of variance.** Analysis of variance for a factorial research design.

**Factorial research design.** Way of organizing a study in which the influence of two or more variables is studied at once by setting up the situation so that a different group of people are tested for each combination of the levels of the variables; for example, in a 2 × 2 factorial research design there would be four groups, those high on variable 1 and high on variable 2, those high on variable 1 but low on variable 2, those high on variable 2 but low on variable 1, and those low on variable 1 and low on variable 2.

**Floor effect.** Situation in which many scores pile up at the low end of a distribution (creating skewness to the right) because it is not possible to have any lower score.

**Frequency distribution.** Pattern of frequencies over the various values; what a frequency table or histogram describes.

**Frequency table.** Ordered listing of the number of individuals having each of the different values for a particular variable.

**Grand mean (*GM*).** In analysis of variance, overall mean of all the scores, regardless of what group they are in; when group sizes are equal, mean of the group means.

**Grouped frequency table.** Frequency table in which the number of individuals (frequency) is given for each interval of values.

**Grouping variable.** A variable that separates groups in analysis of variance (and *t* tests); also see *independent variable*.

**Haphazard selection.** Procedure of selecting a sample of individuals to study by taking whoever is available or happens to be first on a list; should not be confused with true random selection.

**Harmonic mean.** Special average influenced more by smaller numbers; in a *t* test for independent means when the number of scores in the two groups differ, the harmonic mean is used as the equivalent of each group's sample size when determining power.

**Heavy-tailed distribution.** Distribution that differs from a normal curve by being too spread out so that a histogram of the distribution would have too many scores at each of the two extremes ("tails").

**Histogram.** Barlike graph of a frequency distribution in which the values are plotted along the horizontal axis and the height of each bar is the frequency of that value; the bars are usually placed next to each other without spaces, giving the appearance of a city skyline.

**Hypothesis.** A prediction, often based on observation, previous research, or theory, that is tested in a research study.

**Hypothesis testing.** Procedure for deciding whether the outcome of a study (results for a sample) support a particular theory or practical innovation (which is thought to apply to a population).

**Independence.** Situation of no relationship between two variables; term usually used regarding two nominal variables in the chi-square test for independence.

**Independent variable.** Variable considered to be a cause, such as what group a person is in for a *t* test or analysis of variance; also see *grouping variable*.

**Inferential statistics.** Procedures for drawing conclusions based on the scores collected in a research study (sample scores) but going beyond them (to conclusions about a population).

**Interaction effect.** Situation in a factorial analysis of variance in which the combination of variables has an effect that could not be predicted from the effects of the two variables individually; the effect of one grouping variable depends on the level of the other grouping variable.

**Interval.** In a grouped frequency table, the range of values that are grouped together. (For example, if the interval size was 10, one of the intervals might be from 10 to 19.)

**Interval size.** In a grouped frequency table, difference between the start of one interval and the start of the next.

**Inverse transformation.** Data transformation using the inverse (1 divided by the number) of each score.

**Level of measurement.** Type of underlying numerical information provided by a measure, such as equal-interval, rank-order, and nominal (categorical).

**Level of significance.** Same as *significance level*.

**Light-tailed distribution.** Distribution that differs from a normal curve by being too peaked or pinched so that a histogram of the distribution would have too few scores at each of the two extremes ("tails").

**Linear correlation.** Relationship between two variables that shows up on a scatter diagram as the dots roughly following a straight line; a correlation coefficient (*r*) unequal to 0.

**Log transformation.** Data transformation using the logarithm of each score.

**Longitudinal study.** A study where people are measured at two or more points in time.

**Long-run relative-frequency interpretation of probability.** Understanding of probability as the proportion of a particular outcome that you would get if the experiment were repeated many times.

**Main effect.** Difference between groups on one grouping variable in a factorial analysis of variance; result for a variable that divides the groups, averaging across the levels of the other variable that divides the groups (sometimes used only for significant differences).

**Marginal mean.** In a factorial design, mean score for all the participants at a particular level of one of the grouping variables; often shortened to *marginal*.

**Mean (*M*).** Arithmetic average of a group of scores; sum of the scores divided by the number of scores.

**Mean of a distribution of means (Population $M_M$).** The mean of a distribution of means of samples of a given size from a particular population; the same as the mean of the population of individuals. (6)

**Median.** Middle score when all the scores in a distribution are arranged from lowest to highest.

**Meta-analysis.** Statistical method for combining effect sizes from different studies.

**Mode.** Value with the greatest frequency in a distribution.

**Multimodal distribution.** Frequency distribution with two or more high frequencies separated by a lower frequency; a bimodal distribution is the special case of two high frequencies.

**Multiple correlation.** Correlation of a criterion variable with two or more predictor variables.

**Multiple correlation coefficient (*R*).** Measure of degree of multiple correlation; positive square root of the proportion of variance accounted for in a multiple regression analysis.

**Multiple regression.** Procedure for predicting scores on a criterion variable from scores on two or more predictor variables.

**Multivariate statistical test.** Statistical procedure involving more than one criterion or outcome variable.

**Negative correlation.** Relation between two variables in which high scores on one go with low scores on the other, mediums with mediums, and lows with highs; on a scatter diagram, the dots roughly follow a straight line sloping down and to the right; a correlation coefficient (*r*) less than 0.

**95% confidence interval.** Confidence interval in which, roughly speaking, there is a 95% chance that the population mean falls within this interval.

**99% confidence interval.** Confidence interval in which, roughly speaking, there is a 99% chance that the population mean falls within this interval.

**No correlation.** No systematic relationship between two variables; also used for a correlation coefficient (*r*) equal to 0.

**Nominal variable.** Variable with values that are categories (that is, they are names rather than numbers); same as *categorical variable*.

**Nondirectional hypothesis.** Research hypothesis that does not predict a particular direction of difference between the population like the sample studied and the population in general.

**Nonparametric test.** Hypothesis-testing procedure making no assumptions about population parameters; approximately the same as *distribution-free test*.

**Normal curve.** Specific, mathematically defined, bell-shaped frequency distribution that is symmetrical and unimodal; distributions observed in nature and in research commonly approximate it.

**Normal curve table.** Table showing percentages of scores associated with the normal curve; the table usually includes percentages of scores between the mean and various numbers of standard deviations above the mean and percentages of scores more positive than various numbers of standard deviations above the mean.

**Normal distribution.** Frequency distribution following a normal curve.

**Null hypothesis.** Statement about a relation between populations that is the opposite of the research hypothesis; a statement that in the population there is no difference (or a difference opposite to that predicted) between populations; a contrived statement set up to examine whether it can be rejected as part of hypothesis testing.

**Numerator degrees of freedom ($df_{\text{Between}}$).** Same as *between-groups degrees of freedom.*

**Numeric variable.** Variable whose values are numbers (as opposed to a nominal variable).

**Observed frequency (*O*).** In a chi-square test, number of individuals actually found in the study to be in a category or cell.

**One-tailed test.** Hypothesis-testing procedure for a directional hypothesis; situation in which the region of the comparison distribution in which the null hypothesis would be rejected is all on one side (or tail) of the distribution.

**One-way analysis of variance.** Analysis of variance in which there is only one grouping variable (as distinguished from a factorial analysis of variance, such as a two-way or three-way analysis of variance).

**Ordinal variable.** Same as *rank-order variable*.

**Outcome.** Term used in discussing probability for the result of an experiment (or almost any event, such as a coin coming up heads or it raining tomorrow).

**Outlier.** Score with an extreme value (very high or very low) in relation to the other scores in the distribution.

**Parametric test.** Ordinary hypothesis-testing procedure, such as a *t* test or an analysis of variance, that requires assumptions about the shape or other parameters (such as the variance) of the populations.

**Perfect correlation.** Relationship between two variables that shows up on a scatter diagram as the dots exactly following a straight line; correlation of $r = 1$ or $r = -1$; situation in which each person's *Z* score on one variable is exactly the same as that person's *Z* score on the other variable.

**Phi coefficient ($\phi$).** Measure of association between two dichotomous nominal variables; square root of division of chi-square statistic by *N*; equivalent to correlation of the two variables if they were each given numerical values (for example, of 1 and 0 for the two categories); effect-size measure for a chi-square test for independence with a $2 \times 2$ contingency table.

**Pooled estimate of the population variance ($S^2_{\text{Pooled}}$).** In a *t* test for independent means, weighted average of the estimates of the population variance from two samples (each estimate weighted by a proportion consisting of its sample's degrees of freedom divided by the total degrees of freedom for both samples).

**Population.** Entire group of people to which a researcher intends the results of a study to apply; the larger group to which inferences are made on the basis of the particular set of people (sample) studied.

**Population mean.** Mean of the population (usually not known).

**Population parameter.** Actual value of the mean, standard deviation, and so on, for the population (usually population parameters are not known, though often they are estimated based on information in samples).

**Population standard deviation.** Standard deviation of the population (usually not known).

**Population variance.** Variance of the population (usually not known).

**Positive correlation.** Relation between two variables in which high scores on one go with high scores on the other, mediums with mediums, and lows with lows; on a scatter diagram, the dots roughly follow a straight line sloping up and to the right; a correlation coefficient (*r*) greater than 0.

**Power.** Same as *statistical power*.

**Power table.** Table for a hypothesis-testing procedure showing the statistical power of a study for various effect sizes and sample sizes.

**Prediction model.** Formula or rule for making predictions; that is, formula for predicting a person's score on a criterion variable based on the person's score on one or more predictor variables.

**Predictor variable (usually *X*).** In prediction (regression), variable that is used to predict scores of individuals on another variable.

**Probability (*p*).** Expected relative frequency of a particular outcome; the proportion of successful outcomes to all outcomes.

**Proportion of variance accounted for ($r^2$ or $R^2$).** Measure of association between variables used when comparing associations found in different studies or with different variables; correlation coefficient squared or multiple correlation coefficient squared; in the context of correlation it represents the proportion of the total variance in one variable that can be explained by the other variable; measure of effect size for analysis of variance.

**Protected *t* tests.** In analysis of variance, *t* tests among pairs of means after finding that the *F* for the overall difference among the means is significant.

**Quantitative variable.** Same as *numeric variable*.

**Random selection.** Method for selecting a sample that uses truly random procedures (usually meaning that each person in the population has an equal chance of being selected); one procedure is for the researcher to begin with a complete list of all the people in the population and select a group of them to study using a table of random numbers; should not be confused with haphazard selection.

**Rank-order test.** Hypothesis-testing procedure that uses rank-ordered scores.

**Rank-order transformation.** Changing a set of scores to ranks (for example, so that the lowest score is rank 1, the next lowest rank 2, etc.).

**Rank-order variable.** Numeric variable in which the values are ranks, such as class standing or place finished in a race; also called *ordinal variable*.

**Raw score.** Ordinary score (or any other number in a distribution before it has been made into a *Z* score or otherwise transformed).

**Rectangular distribution.** Frequency distribution in which all values have approximately the same frequency.

**Regression coefficient.** Number multiplied by a person's score on the predictor variable as part of a prediction model.

**Repeated measures analysis of variance.** Analysis of variance in which each individual is tested more than once so that the levels of the grouping variable(s) are different times or types of testing for the same people.

**Repeated-measures design.** Research strategy in which each person is tested more than once; same as *within-subjects design*.

**Research hypothesis.** Statement in hypothesis testing about the predicted relation between populations (usually a prediction of a difference between population means).

**Sample.** Scores of the particular group of people studied; usually considered to be representative of the scores in some larger population.

**Sample statistic.** Descriptive statistic, such as the mean or standard deviation, figured from the scores in a particular group of people studied.

**Scatter diagram.** Graph showing the relationship between two variables: the values of one variable (often the predictor variable) are along the horizontal axis and the values of the other variable (often the criterion variable) are along the vertical axis; each score is shown as a dot in this two-dimensional space; also called *scatterplot* or *scattergram*.

**Scatterplot.** Same as *scatter diagram*.

**Score.** Particular person's value on a variable.

**Shape of a distribution of means.** Contour of a histogram of a distribution of means, such as whether it follows a normal curve or is skewed; in general, a distribution of means will tend to be unimodal and symmetrical and is often normal.

**Significance level.** Probability of getting statistical significance if the null hypothesis is actually true; probability of a Type I error.

**Skew.** Extent to which a frequency distribution has more scores on one side of the middle as opposed to being perfectly symmetrical.

**Skewed distribution.** Distribution in which the scores pile up on one side of the middle and are spread out on the other side; distribution that is not symmetrical.

**Square-root transformation.** Data transformation using the square root of each score.

**Squared deviation score.** Square of the difference between a score and the mean.

**Standard deviation (*SD*).** Square root of the average of the squared deviations from the mean; the most common descriptive statistic for variation; approximately the average amount that scores in a distribution vary from the mean.

**Standard deviation of a distribution of differences between means ($S_{\text{Difference}}$).** In a *t* test for independent means, square root of the variance of the distribution of differences between means.

**Standard deviation of a distribution of means (Population $SD_M$, $S_M$).** Square root of the variance of the distribution of means; same as *standard error (SE)*.

**Standard deviation of a distribution of means of difference scores ($S_M$).** In a *t* test for dependent means, square root of the variance of the distribution of mean differences.

**Standard error (*SE*).** Same as *standard deviation of a distribution of means*; also called *standard error of the mean (SEM)*.

**Standard score.** Same as *Z score*.

**Standardized regression coefficient (beta, β).** Regression coefficient in a prediction model using *Z* scores.

**Statistical power.** Probability that the study will give a significant result if the research hypothesis is true.

**Statistically significant.** Conclusion that the results of a study would be unlikely if in fact the sample studied represents a population that is no different from the population in general; an outcome of hypothesis testing in which the null hypothesis is rejected.

**Statistics.** A branch of mathematics that focuses on the organization, analysis, and interpretation of a group of numbers.

**Sum of squared deviations.** Total over all the scores of each score's squared difference from the mean.

**Symmetrical distribution.** Distribution in which the pattern of frequencies on the left and right side are mirror images of each other.

***t* distribution.** Mathematically defined curve that is the comparison distribution used in a *t* test.

***t* score.** On a *t* distribution, number of standard deviations from the mean (like a *Z* score, but on a *t* distribution).

***t* table.** Table of cutoff scores on the *t* distribution for various degrees of freedom, significance levels, and one- and two-tailed tests.

***t* test.** Hypothesis-testing procedure in which the population variance is unknown; it compares *t* scores from a sample to a comparison distribution called a *t* distribution.

***t* test for dependent means.** Hypothesis-testing procedure in which there are two scores for each person and the population variance is not known; it determines the significance of a hypothesis that is being tested using difference or change scores from a single group of people; also called paired sample *t* test, *t* test for correlated means, *t* test for matched samples, and *t* test for matched pairs.

***t* test for independent means.** Hypothesis-testing procedure in which there are two separate groups of people tested and in which the population variance is not known.

***t* test for a single sample.** Hypothesis-testing procedure in which a sample mean is being compared to a known population mean and the population variance is unknown.

**Theory.** A set of principles that attempt to explain one or more facts, relationships, or events; behavioral and social scientists often derive specific predictions (hypotheses) from theories that are then examined in research studies.

**True experiment.** A study in which participants are randomly assigned (say, by flipping a coin) to a particular level of a grouping or predictor variable and then measured on another variable.

**Two-tailed test.** Hypothesis-testing procedure for a nondirectional hypothesis; the situation in which the region of the comparison distribution in which the null hypothesis would be rejected is divided between the two sides (or tails) of the distribution.

**Two-way analysis of variance.** Analysis of variance for a two-way factorial research design.

**Two-way factorial research design.** Factorial design with two variables that each divide the groups.

**Type I error.** Rejecting the null hypothesis when in fact it is true; getting a statistically significant result when in fact the research hypothesis is not true.

**Type II error.** Failing to reject the null hypothesis when in fact it is false; failing to get a statistically significant result when in fact the research hypothesis is true.

**Unbiased estimate of the population variance ($S^2$).** Estimate of the population variance, based on sample scores, that has been corrected so that it is equally likely to overestimate or underestimate the true population variance; the correction used is dividing the sum of squared deviations by the sample size minus 1, instead of the usual procedure of dividing by the sample size directly.

**Unimodal distribution.** Frequency distribution with one value clearly having a larger frequency than any other.

**Value.** Number or category that a score can have.

**Variable.** Characteristic that can have different values.

**Variance ($SD^2$).** Measure of how spread out a set of scores are; average of the squared deviations from the mean; standard deviation squared.

**Variance of a distribution of differences between means ($S^2_{\text{Difference}}$).** One of the numbers figured as part of a *t* test for independent means; it equals the sum of the variances of the distributions of means associated with each of the two samples.

**Variance of a distribution of means (Population $SD^2_M$, $S^2_M$).** Variance of the population divided by the number of scores in each sample.

**Variance of a distribution of means of difference scores ($S_M^2$).** One of the numbers figured as part of a *t* test for dependent means; it equals the estimated population variance of differences scores divided by the sample size.

**Weighted average.** Average in which the scores being averaged do not have equal influence on the total, as in figuring the pooled variance estimate in a *t* test for independent means.

**Within-groups degrees of freedom ($df_{\text{Within}}$).** Degrees of freedom used in the within-groups estimate of the population variance in an analysis of variance (the denominator of the *F* ratio); number of scores free to vary (number of scores in each group minus 1, summed over all the groups) in figuring the within-groups population variance estimate; same as *denominator degrees of freedom*.

**Within-groups estimate of the population variance ($S_{\text{Within}}^2$).** In analysis of variance, estimate of the variance of the distribution of the population of individuals based on the variation among the scores within each of the actual groups studied.

**Within-subject design.** Same as *repeated-measures design*.

***Z* score.** Number of standard deviations a score is above (or below, if it is negative) the mean of its distribution; it is thus an ordinary score transformed so that it better describes that score's location in a distribution.

***Z* test.** Hypothesis-testing procedure in which there is a single sample and the population variance is known.

# Glossary of Symbols

**$\alpha$** Significance level, such as .05 or .01; probability of a Type I error.

**$\beta$** Standardized regression coefficient.

**$\Sigma$** Sum of; add up all the scores following.

**$\phi$** Phi coefficient; effect size in a chi-square test for independence with a 2 × 2 contingency table.

**$\phi_c$** Cramer's phi, effect size in chi-square test for independence with a contingency table larger than 2 × 2.

**$\chi^2$** Chi-square statistic.

***df*** Degrees of freedom.

**$df_1$, $df_2$, and so on.** Degrees of freedom for the first group, second group, and so on.

**$df_{\text{Between}}$** Between-groups (numerator) degrees of freedom in analysis of variance.

**$df_{\text{Smaller}}$** Degrees of freedom for the nominal variable (the row or column in the contingency table) with the smaller number of categories in a chi-square test for independence.

**$df_{\text{Total}}$** Total degrees of freedom over all groups.

**$df_{\text{Within}}$** Within-groups (denominator) degrees of freedom in analysis of variance.

***F* ratio** In an analysis of variance, ratio of the between-groups population variance estimate to the within-groups population variance estimate.

***GM*** Grand mean; mean of all scores in the analysis of variance.

***M*** Mean.

**$M_1$, $M_2$, and so on.** Mean of the first group, second group, and so on.

***n*** Number of scores within each group in analysis of variance.

***N*** Total number of scores.

**$N_1$, $N_2$, and so on.** Number of scores in the first group, second group, and so on.

**$N_{\text{Categories}}$** Number of categories in a chi-square test for goodness of fit.

**$N_{\text{Columns}}$** Number of columns in a contingency table.

**$N_{\text{Groups}}$** Number of groups in an analysis of variance.

**$N_{\text{Rows}}$** Number of rows in a contingency table.

***p*** Probability.

***r*** Correlation coefficient.

**$r^2$, $R^2$** Proportion of variance accounted for.

***R*** Multiple correlation coefficient.

***S*** Unbiased estimate of the population standard deviation.

**$S^2$** Unbiased estimate of the population variance.

**$S_1^2$, $S_2^2$, and so on.** Unbiased estimate of the population variance based on scores in the first sample, second sample, and so on.

**$S^2_{\text{Between}}$** Between-groups estimate of the population variance.

**$S_{\text{Difference}}$** Standard deviation of the distribution of differences between means.

**$S^2_{\text{Difference}}$** Variance of the distribution of differences between means.

***SE*** Standard error (standard deviation of the distribution of means, $S_M$).

**$S_M$** Standard deviation of the distribution of means based on an estimated population variance, same as standard error (*SE*).

**$S_M^2$** Variance of a distribution of means based on an estimated population variance in a *t* test or as estimated from the variation among means of groups in the analysis of variance.

**$S^2_{M_1}$, $S^2_{M_2}$, and so on.** Variance of the distribution of means based on a pooled population variance estimate, corresponding to the first sample, second sample, and so on.

**$S_{\text{Pooled}}$** Pooled estimate of the population standard deviation.

**$S^2_{\text{Pooled}}$** Pooled estimate of the population variance.

**$S^2_{\text{Within}}$** Within-groups estimate of the population variance.

***SD*** Standard deviation.

**$SD^2$** Variance.

***t* score** Number of standard deviations from the mean on a *t* distribution.

***X*** Score on a particular variable; in prediction (regression), *X* is usually the predictor variable.

**$X_1$, $X_2$, and so on.** First predictor variable, second predictor variable, and so on.

***Y*** Score on a particular variable; in prediction (regression), *Y* is usually the criterion variable.

***Z*** Number of standard deviations from the mean.

**$Z_X$** *Z* score for variable *X*.

**$Z_Y$** *Z* score for variable *Y*.

*This page intentionally left blank*

# Displaying the Order in a Group of Numbers Using Tables and Graphs

## Chapter Outline


- The Two Branches of Statistical Methods
- Some Basic Concepts
- Kinds of Variables
- Frequency Tables
- Histograms
- Shapes of Frequency Distributions
- Frequency Tables and Histograms in Research Articles
- Learning Aids
  - *Summary*
  - *Key Terms*
  - *Example Worked-Out Problems*
  - *Practice Problems*
  - *Using SPSS*


We imagine you to be as unique as the other students we have known who have taken this course. Some of you are highly scientific sorts; others are more intuitive. Some of you are fond of math; others are less so, or even afraid of it. Whatever your style, we welcome you.

Given that you *can* learn statistics, you still have to decide if you want to make the effort it will require. Why would you want to do that, except to meet a requirement of your major? (Not a very energizing motive.) First, you will be far better equipped to read research articles in your major. Second, you'll be on your way to being able to do your own research if you so choose. Third, you will improve both your reasoning and your intuition. Fourth, each time you finally grasp something you were struggling with, you will feel great.

Formally, **statistics** is a branch of mathematics that focuses on the organization, analysis, and interpretation of a group of numbers. But really what is statistics? Think of statistics as a tool that extends a basic thinking process employed by every human: You observe a thing; you wonder what it means or what caused it; you have an insight or make an intuitive guess; you observe again, but now in detail, or you try making some little changes in the process to test your intuition. Then you face the eternal problem: was your hunch confirmed or not? What are the chances that what you observed this second time will happen again and again so that you can announce your insight to the world as something probably true?

Statistics is a method of pursuing truth. As a minimum, statistics can tell you the likelihood that your hunch is true in this time and place and with these sorts of people. This pursuit of truth, or at least its future likelihood, is the essence of science. It is also the essence of human evolution and survival. Think of the first research questions: What will the mammoths do next spring? What will happen if I eat this root? It is easy to see how the early accurate "researchers" survived. You are here today because your ancestors exercised brains as well as brawn. Do those who come after you the same favor: Think carefully about outcomes. Statistics is one good way to do that.

Behavioral and social scientists usually use a computer and statistical software to carry out statistical procedures. However, the best way to develop a solid understanding of statistics is to learn how to do the procedures by hand (with the help of a calculator—it's not the multiplying and adding that you learn from, but the going through all the steps). In order to minimize the amount of figuring you have to do, we use relatively small groups of numbers in examples and practice problems. We hope that this will also allow you to focus more on the *underlying principles and logic* of the statistical procedures, rather than on the computations for each practice problem (such as subtracting 3 from 7 and then dividing the result by 2 to give an answer of 2). Having said that, we also recognize the importance of learning how to do statistical procedures on a computer, as you most likely would when conducting your own research. SPSS statistical software is commonly used by behavioral and social scientists to carry out statistical analyses. Check with your instructor to see if you have access to SPSS at your institution.

**statistics** A branch of mathematics that focuses on the organization, analysis, and interpretation of a group of numbers.

**descriptive statistics** Procedures for summarizing a group of scores or otherwise making them more understandable.

## The Two Branches of Statistical Methods

There are two main branches of statistical methods:

1. **Descriptive statistics:** Behavioral and social scientists use descriptive statistics to summarize and describe a group of numbers from a research study.

2. **Inferential statistics:** Behavioral and social scientists use inferential statistics to draw conclusions and make inferences that are based on the numbers from a research study but that go beyond these numbers. For example, inferential statistics allow researchers to make inferences about a large group of individuals based on a research study in which a much smaller number of individuals took part.

**inferential statistics** Procedures for drawing conclusions based on the scores collected in a research study but going beyond them.

## BOX 1 Math Anxiety, Statistics Anxiety, and You: A Message for Those of You Who Are Truly Worried about This Course

Let's face it: Many of you dread this course, even to the point of having a full-blown case of "statistics anxiety" (Zeidner, 1991). If you become tense the minute you see numbers, we need to talk about that right now.

First, this course is a chance for a fresh start with digits. Your past performance in (or avoidance of) geometry, trigonometry, calculus, or similar horrors need not influence in any way how well you comprehend statistics. This is largely a different subject.

Second, if your worry persists, you need to determine where it is coming from. Math or statistics anxiety, test anxiety, general anxiety, and general lack of confidence each seems to play its own role in students' difficulties with math courses (Cooper & Robinson, 1989; Dwinell & Higbee, 1991).

*Is your problem mainly math/statistics anxiety?* An Internet search will yield many wonderful books and Web sites to help you. We highly recommend Sheila Tobias's *Overcoming Math Anxiety* (1995). Tobias, a former math avoider herself, suggests that your goal be "math mental health," which she defines as "the willingness to learn the math you need when you need it" (p. 12). (Could it be that this course in statistics is one of those times?)

Tobias (1995) explains that math mental health is usually lost in elementary school, when you are called to the blackboard, your mind goes blank, and you are unable to produce the one right answer to an arithmetic problem. What confidence remained probably faded during timed tests, which you did not realize were difficult for everyone except the most proficient few.

Conquer your math anxiety by thinking back to when it might have started. Before that you were doing as well as others, so this is not about your lack of ability, but about some bad experiences or inadequate teaching that left you feeling inferior to others who were better prepared. Starting now, treat yourself more gently. Do not expect to understand things instantly. You are still recovering.

Start out confident. To do well, you only need a semester of high school algebra, but do review it, perhaps online, so that your math brain is warmed up. Go to class faithfully and do keep yourself alert during it. Do your statistics homework first, before you are tired. Ask for help from teachers, teaching assistants, or anyone you know in the class. Form a study group. If you don't understand something, try explaining it to someone who understands it even less. You may figure it out in the process. Read ahead in a relaxed way, so that what you hear in class is already familiar.

*Is your problem test anxiety?* Anxiety produces arousal, and one of the best understood relationships in psychology is between arousal and performance. Whereas moderate arousal helps performance, too much or too little dramatically decreases cognitive capacity and working memory. Things you have learned become harder to recall. Your mind starts to race, creating more anxiety, more arousal, and so on. There are many ways to reduce anxiety and arousal in general, such as learning to breathe properly and to take a brief break to relax deeply. Your counseling center should be able to help you or direct you to some good books on the subject. Again, many Web sites deal with reducing anxiety.

To reduce test anxiety specifically, overprepare for the first few tests, so that you go in with the certainty that you cannot possibly fail, no matter how aroused you become. Do this especially for the first test, because there will be no old material to review and a good score on your first test will make you less anxious about the next. Do a practice test, and make it as similar to a real test as possible, making a special effort to duplicate the aspects that bother you most. Once you think you are well prepared, set yourself a time limit for solving some homework problems. Make yourself write out answers fully and legibly. This may be part of what makes you feel slow during a test. If the presence of others bothers you—the sound of their scurrying pencils while yours is frozen in midair—do your practice test with others in your course. Even make it an explicit contest to see who can finish first.

For the test itself, reduce the effects of anxiety by jotting down in the margins any formulas you are afraid you will forget. Glance through the whole test so that you know how to pace yourself. Read the instructions carefully. Start with questions that you can answer easily, and if some seem impossible, don't worry—you still may pass even if you do not answer them correctly. Try to ignore the people around you. Some people may seem to be working faster, but they may not be working as well as you.

*Is your problem a general lack of confidence?* Is there something else in your life causing you to worry or feel bad about yourself? Then we suggest that it is time you tried your friendly college counseling center.

*Lastly, could you be highly sensitive?* High sensitivity, reactivity, or responsiveness is an innate trait found in about 15–20% of humans (and at least a hundred other species; Wolf, Van Doorn, & Weissing, 2008). Individuals with the trait naturally adopt a strategy of reflecting before acting, rather than rushing into things. They process information more deeply and notice subtleties, as seen in brain imaging studies (Jagiellowicz et al., 2010). This makes them especially intuitive, creative, and conscientious (Aron, 1996; Aron & Aron, 1997). They also have stronger emotional reactions and are more affected by their environment and events in their childhood (Aron, Aron, & Davies, 2005; Belsky, Bakermans-Kranenburg, & Van IJzendoorn, 2007; Ellis, Essex, & Boyce, 2005). Being aware of so much, they become overstimulated sooner than others, so that college life itself can be a challenge for them, even though they have the potential to perform exceptionally well when not overaroused.

You might want to find out if you are a highly sensitive person (at *http://www.hsperson.com*). If you are, appreciate the trait's assets and make some allowances for its one disadvantage, this tendency to become easily overaroused. It has to affect your performance on tests. What matters is what you actually know, which is probably quite a bit. This simple act of self-acceptance—that you are *not* less smart but *are* more sensitive—may in itself help ease your arousal when trying to express your statistical knowledge.

So good luck to all of you. We wish you the best both while taking this course and in your lives.

In this chapter, we focus on descriptive statistics. This topic is important in its own right, but it also prepares you to understand inferential statistics.

In this chapter, we introduce you to some basic concepts, then you learn to use tables and graphs to describe a group of numbers. The purpose of descriptive statistics is to make a group of numbers easy to understand. As you will see, tables and graphs help a great deal.

## Some Basic Concepts

### Variables, Values, and Scores

As part of a larger study (Aron, Paris, & Aron, 1995), researchers gave a questionnaire to 151 students in an introductory statistics class during the first week of class. One question asked was, "How stressed have you been in the last 2½ weeks, on a scale of 0 to 10, with 0 being *not at all stressed* and 10 being *as stressed as possible?*" (How would *you* answer?) In this study, the researchers used a survey to examine students' level of stress. Some other methods that researchers use to study stress include creating stress with laboratory tasks (like having to be videotaped giving a talk for humans or swimming in water for rats) and measuring stress-related hormones or brain changes.

In the current example, level of stress is a **variable,** which can have **values** from 0 to 10, and the value of any particular person's answer is the person's **score.** If you answered 6, your score is 6; your score has a value of 6 on the variable called level of stress.

**variable** Characteristic that can have different values.

**value** Number or category that a score can have.

**score** Particular person's value on a variable.

More formally, a variable is a condition or characteristic that can have different values. In short, it can *vary.* In our example, the variable is level of stress. It can have

**Table 1** Some Basic Terminology

| Term | Definition | Examples |
|---|---|---|
| Variable | Condition or characteristic that can have different values | Stress level; age; gender; religion |
| Value | Number or category | 0, 1, 2, 3, 4; 25, 85; female; Catholic |
| Score | A particular person's value on a variable | 0, 1, 2, 3, 4; 25, 85; female; Catholic |

values of 0 through 10. Height is a variable, social class is a variable, score on a creativity test is a variable, number of people absent from work on a given day is a variable, dosage of a medication is a variable, political party preference is a variable, and class size is a variable.

A value is just a number, such as 4, −81, or 367.12. A value can also be a category, such as female or male or the country in which you live (Canada, the United States, Australia, etc.).

Finally, on any variable, each person has a particular number or score that is that person's value on the variable. For example, your score on the stress variable might have a value of 6. Another student's score might have a value of 8.

Behavioral and social science research is about variables, values, and scores (see Table 1). The formal definitions are a bit abstract. In practice, you will find that what we mean when we use these words is usually clear.

## Kinds of Variables

Most of the variables behavioral and social scientists use are like those in the stress ratings example: The scores are numbers that tell you how much there is of the thing being measured. In the stress ratings example, the higher the number is, the more stress there is. We call this kind of variable a **numeric variable.** Numeric variables are also called *quantitative variables.*

Behavioral and social scientists use two main kinds of numeric variables: equal-interval variables and rank-order variables. The kind of variable used most often is a variable in which the numbers stand for approximately equal amounts of what is being measured. This is called an **equal-interval variable.** Take grade point average (GPA). This is a roughly equal-interval variable. For example, the difference between a GPA of 2.5 and 2.8 means about as much of a difference as the difference between that of 3.0 and 3.3 (each is a difference of 0.3 of a GPA). Most behavioral and social scientists also consider scales like the 0 to 10 stress ratings as roughly equal interval. So, for example, a difference between stress ratings of 4 and 6 means about as much as the difference between 7 and 9.

The other kind of numeric variable behavioral and social scientists often use is where the numbers only stand for relative rankings. This is called a **rank-order variable.** (Rank-order variables are also called *ordinal variables.*) An example is rank in one's graduating class. Notice that with a rank-order variable, the difference between one number and the next does not always mean the same amount of the underlying thing being measured. For example, the difference between being second and third in your graduating class could be a very unlike amount of difference in underlying GPA than the difference between being eighth and ninth. A rank-order variable provides less information than an equal-interval variable. It is less precise. However, behavioral and social scientists often use rank-order variables because they are the only information available.

**numeric variable** Variable whose values are numbers. Also called *quantitative variable.*

**equal-interval variable** A variable in which the numbers stand for approximately equal amounts of what is being measured.

**rank-order variable** Numeric variable in which the values are ranks, such as class standing or place finished in a race. Also called *ordinal variable.*

**Table 2** Levels of Measurement

| Level | Definition | Example |
|---|---|---|
| Equal-interval | Numeric variable in which differences between values correspond to differences in the underlying thing being measured | Stress level, age |
| Rank-order | Numeric variable in which values correspond to the relative position of things measured | Class standing, position finished in a race |
| Nominal | Variable in which the values are categories | Gender, religion |

There is also a kind of variable that is not about numbers at all, but that refers just to names or categories. This is called a **nominal variable.** The term *nominal* comes from the idea that its values are names. (Nominal variables are also called *categorical variables* because their values are categories.) For example, for the nominal variable gender, the values are female and male. A person's "score" on the variable gender is one of these two values. Similarly, hair color has values, such as brown, black, and so forth.

These different kinds of variables are based on different **levels of measurement** (see Table 2). Researchers sometimes have to decide how they will measure a particular variable. For example, they might use an equal-interval scale, a rank-order scale, or a nominal scale. The level of measurement selected affects the type of statistics that can be used with a variable. We focus mostly on numeric equal-interval variables. However, rank-order and nominal variables also are fairly common in the behavioral and social sciences.

**How are you doing?**

1. A father rates his daughter as a 2 on a 7-point scale (from 1 to 7) of crankiness. In this example, (a) what is the variable, (b) what is the score, and (c) what is the range of possible values?
2. What is the difference between a numeric and a nominal variable?
3. Name the kind of variable for each of the following variables: (a) a person's nationality (Mexican, French, Japanese, etc.), (b) a person's score on a standardized IQ test, (c) a person's place on a waiting list (first in line, second in line, etc.).

**Answers**

1. (a) crankiness, (b) 2, (c) 1 to 7.
2. A numeric variable has values that are numbers that tell you the degree or extent of what the variable measures; a nominal variable has values that are different categories and have no particular numerical order.
3. (a) nominal, (b) equal-interval, (c) rank-order.

**nominal variable** Variable with values that are categories (that is, they are names rather than numbers). Also called *categorical variable.*

**level of measurement** Type of underlying numerical information provided by a measure, such as equal-interval, rank-order, and nominal (categorical).

## Frequency Tables

### An Example

Let's return to the stress rating example. Recall that in this study, students in an introductory statistics class during the first week of the course answered the question "How stressed have you been in the last 2½ weeks, on a scale of 0 to 10, with 0 being *not at* *all stressed* and 10 being *as stressed as possible?*" The actual study included scores from 151 students. To ease the learning for this example, we are going to use a representative subset of scores from 30 of these 151 students (this also saves you time if you want to try it for yourself). The 30 students' scores (their ratings on the scale) are:

8, 7, 4, 10, 8, 6, 8, 9, 9, 7, 3, 7, 6, 5, 0, 9, 10, 7, 7, 3, 6, 7, 5, 2, 1, 6, 7, 10, 8, 8.

Looking through all these scores gives some sense of the overall tendencies. But this is hardly an accurate method. One solution is to make a table showing how many students used each of the 11 values that the ratings can have (0, 1, 2, etc., through 10). That is, the number of students who used each particular rating is the *frequency* of that value. We have done this in Table 3. We also figured the percentage each value's frequency is of the total number of scores. Tables like this sometimes give only the raw-number frequencies and not the percentages, or only the percentages and not the raw-number frequencies.[1]

Table 3 is called a **frequency table** because it shows how frequently (how many times) each rating number was used. A frequency table makes the pattern of numbers easy to see. In this example, you can see that most of the students rated their stress around 7 or 8, with few rating it very low.

**Table 3** Frequency Table of Number of Students Rating Each Value of the Stress Scale

| Stress Rating | Frequency | Percent |
|---|---|---|
| 0 | 1 | 3.3 |
| 1 | 1 | 3.3 |
| 2 | 1 | 3.3 |
| 3 | 2 | 6.7 |
| 4 | 1 | 3.3 |
| 5 | 2 | 6.7 |
| 6 | 4 | 13.3 |
| 7 | 7 | 23.3 |
| 8 | 5 | 16.7 |
| 9 | 3 | 10.0 |
| 10 | 3 | 10.0 |

*Source:* Data based on Aron et al. (1995).

## How to Make a Frequency Table

There are four steps in making a frequency table.

❶ **Make a list of each possible value down the left edge of a page, starting from the lowest and ending with the highest.** In the stress rating results, the list goes from 0, the lowest possible rating, up to 10, the highest possible rating. Note that even if one of the ratings between 0 and 10 had not been used, you would still include that value in the listing, showing it as having a frequency of 0. For example, if no one gave a stress rating of 2, you still include 2 as one of the values on the frequency table.

❷ **Go one by one through the scores, making a mark for each next to its value on your list.** This is shown in Figure 1.

❸ **Make a table showing how many times each value on your list was used.** To do this, add up the number of marks beside each value.

❹ **Figure the percentage of scores for each value.** To do this, take the frequency for that value, divide it by the total number of scores, and multiply by 100. You usually will need to round off the percentage. We recommend that you round percentages to one decimal place. Note that because of the rounding, your percentages will not usually add up to exactly 100% (but the total should be close).

## Frequency Tables for Nominal Variables

The preceding steps assume you are using numeric variables, the most common situation. However, you can also use a frequency table to show the number of scores in each value (that is, for each category) of a nominal variable. For example, researchers

**frequency table** Ordered listing of the number of individuals having each of the different values for a particular variable.

[1]In addition, some frequency tables include, for each value, the total number of scores with that value and all values preceding it. These are called *cumulative frequencies* because they tell how many scores are accumulated up to this point on the table. If percentages are used, cumulative percentages also may be included. Cumulative percentages give, for each value, the percentage of scores up to and including that value. The cumulative percentage for any given value (or for a score having that value) is also called a *percentile.* Cumulative frequencies and cumulative percentages help you see where a particular score falls in the overall group of scores.

**TIP FOR SUCCESS**

When doing Step ❷, cross off each score as you mark it on the list. This should help you to avoid mistakes, which are common in this step.

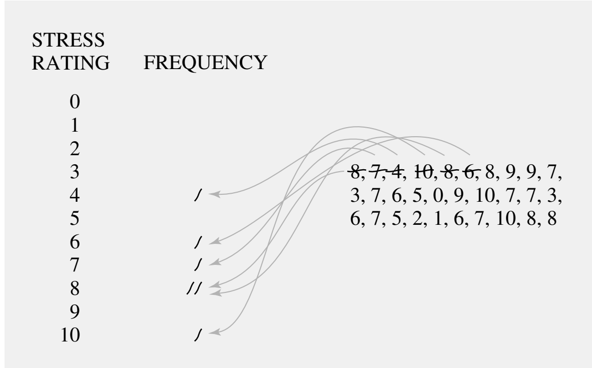



**Figure 1** Making a frequency table for the stress ratings scores. (Data based on Aron, Paris, and Aron, 1995.)

**Table 4** Frequency Table for a Nominal Variable: Closest Person in Life for 208 Students

| Closest Person | Frequency | Percent |
|---|---|---|
| Family member | 33 | 15.9 |
| Nonromantic friend | 76 | 36.5 |
| Romantic partner | 92 | 44.2 |
| Other | 7 | 3.4 |

*Source:* Data from Aron, Aron, and Smollan (1992).

(Aron, Aron, & Smollan, 1992) asked 208 students to name the closest person in their lives. As shown in Table 4, 33 students selected a family member, 76 a nonromantic friend, 92 a romantic partner, and 7 selected some other person. As you can see in Table 4, the values listed on the left-hand side of the frequency table are the values (the categories) of the variable.

## Another Example

McLaughlin-Volpe, Aron, and Reis (2001) had 94 first- and second-year university students keep a diary of their social interactions for a week during the regular semester. Each time a student had a social interaction lasting 10 minutes or longer, the student would fill out a card. The card included questions about who were the other people in the interaction and about various aspects of the conversation. Excluding family and work situations, the number of social interactions of 10 minutes or longer over a week for these 94 students were as follows:

48, 15, 33, 3, 21, 19, 17, 16, 44, 25, 30, 3, 5, 9, 35, 32, 26, 13, 14, 14, 47, 47, 29, 18, 11, 5, 19, 24, 17, 6, 25, 8, 18, 29, 1, 18, 22, 3, 22, 29, 2, 6, 10, 29, 10, 21, 38, 41, 16, 17, 8, 40, 8, 10, 18, 7, 4, 4, 8, 11, 3, 23, 10, 19, 21, 13, 12, 10, 4, 17, 11, 21, 9, 8, 7, 5, 3, 22, 14, 25, 4, 11, 10, 18, 1, 28, 27, 19, 24, 35, 9, 30, 8, 26.

Now, let's follow our four steps for making a frequency table.

❶ **Make a list of each possible value down the left edge of a page, starting from the lowest and ending with the highest.** The lowest possible number of interactions is 0. In this study, the highest number of interactions could be any number. However, the highest actual number in this group was 48; so we can use 48 as the highest value. Thus, the first step is to list these values down a page. (It might be good to use several columns so that you can have all the scores on a single page.)

❷ **Go one by one through the scores, making a mark for each next to its value on your list.** Figure 2 shows the results of this step.

❸ **Make a table showing how many times each value on your list was used.** Table 5 is the result.

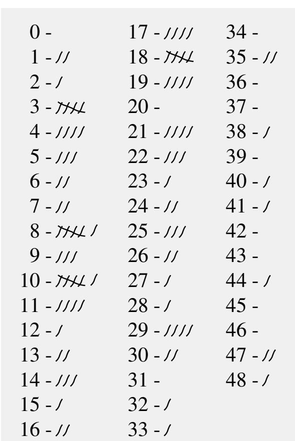



**Figure 2** Making a frequency table of students' social interactions over a week. (Data from McLaughlin-Volpe et al., 2001.)

**Table 5** Frequency Table for Number of Social Interactions During a Week for 94 College Students

| Score | Frequency | Score | Frequency | Score | Frequency |
|---|---|---|---|---|---|
| 0 | 0 | 17 | 4 | 34 | 0 |
| 1 | 2 | 18 | 5 | 35 | 2 |
| 2 | 1 | 19 | 4 | 36 | 0 |
| 3 | 5 | 20 | 0 | 37 | 0 |
| 4 | 4 | 21 | 4 | 38 | 1 |
| 5 | 3 | 22 | 3 | 39 | 0 |
| 6 | 2 | 23 | 1 | 40 | 1 |
| 7 | 2 | 24 | 2 | 41 | 1 |
| 8 | 6 | 25 | 3 | 42 | 0 |
| 9 | 3 | 26 | 2 | 43 | 0 |
| 10 | 6 | 27 | 1 | 44 | 1 |
| 11 | 4 | 28 | 1 | 45 | 0 |
| 12 | 1 | 29 | 4 | 46 | 0 |
| 13 | 2 | 30 | 2 | 47 | 2 |
| 14 | 3 | 31 | 0 | 48 | 1 |
| 15 | 1 | 32 | 1 | | |
| 16 | 2 | 33 | 1 | | |

*Source:* Data from McLaughlin-Volpe et al. (2001).

**TIP FOR SUCCESS**

Be sure to check your work by adding the frequencies for all of the scores. This sum should equal the total number of scores you started with.

❹ **Figure the percentage of scores for each value.** We have *not* done so in this example because with so many categories, it would not help much for seeing the pattern of scores. However, if you want to check your understanding of this step, the first five percentages would be 0.0%, 2.1%, 1.1%, 5.3%, and 4.3%. (These are the percentages for frequencies of 0, 2, 1, 5, and 4, rounded to one decimal place.)

## Grouped Frequency Tables

Sometimes there are so many possible values that a frequency table is too awkward to give a simple picture of the scores. The last example was a bit like that, wasn't it? The solution is to make groupings of values that include all values in a certain range. For example, consider our stress example. Instead of having a separate frequency figure for the group of students who rated their stress as 8 and another for those who rated it as 9, you could have a combined category of 8 and 9. This combined category is a range of values that includes these two values. A combined category like this is called an **interval.** This particular interval of 8 and 9 has a frequency of 8 (the sum of the 5 scores with a value of 8 and the 3 scores with a value of 9).

A frequency table that uses intervals is called a **grouped frequency table.** Table 6 is a grouped frequency table for the stress ratings example. (However, in this example, the full frequency table has only 11 different values. Thus, a grouped frequency table was not really necessary.) Table 7 is a grouped frequency table for the 94 students' numbers of social interactions over a week.

A grouped frequency table can make information even more directly understandable than an ordinary frequency table can. Of course, the greater understandability of a grouped frequency table is at a cost. You lose information about the breakdown of frequencies in each interval.

**Table 6** Grouped Frequency Table for Stress Ratings

| Stress Rating Interval | Frequency | Percent |
|---|---|---|
| 0–1 | 2 | 6.7 |
| 2–3 | 3 | 10.0 |
| 4–5 | 3 | 10.0 |
| 6–7 | 11 | 36.7 |
| 8–9 | 8 | 26.7 |
| 10–11 | 3 | 10.0 |

*Source:* Data based on Aron et al. (1995).

**interval** In a grouped frequency table, the range of values that are grouped together. (For example, if the interval size is 10, one of the intervals might be from 10 to 19.)

**grouped frequency table** Frequency table in which the number of individuals (frequency) is given for each interval of values.

**Table 7** Grouped Frequency Table for Number of Social Interactions During a Week for 94 College Students

| Interval | Frequency | Percent |
|---|---|---|
| 0–4 | 12 | 12.8 |
| 5–9 | 16 | 17.0 |
| 10–14 | 16 | 17.0 |
| 15–19 | 16 | 17.0 |
| 20–24 | 10 | 10.6 |
| 25–29 | 11 | 11.7 |
| 30–34 | 4 | 4.3 |
| 35–39 | 3 | 3.2 |
| 40–44 | 3 | 3.2 |
| 45–49 | 3 | 3.2 |

*Source:* Data from McLaughlin-Volpe et al. (2001).

When setting up a grouped frequency table, it makes a big difference how many intervals you use. There are guidelines to help researchers with this, but in practice it is done automatically by the researcher's computer (see this chapter's "Using SPSS" section for instructions on how to create frequency tables using statistical software). Thus, we will not focus on it. However, should you have to make a grouped frequency table on your own, the key is to experiment with the interval size until you come up with one that is a round number (such as 2, 3, 5, or 10) and that creates about 5–15 intervals. Then, when actually setting up the table, be sure you set the start of each interval to a multiple of the interval size and the top end of each interval to the number just below the start of the next interval. For example, Table 6 uses six intervals with an interval size of 2. The intervals are 0–1, 2–3, 4–5, 6–7, 8–9, and 10–11. Note that each interval starts with a multiple of 2 (0, 2, 4, 6, 8, 10) and the top end of each interval (1, 3, 5, 7, 9) is the number just below the start of the next interval (2, 4, 6, 8, 10). Table 7 uses 10 intervals with an interval size of 5. The intervals are 0–4, 5–9, 10–14, 15–19, and so on, with a final interval of 45–49. Note that each interval starts with a multiple of 5 (0, 5, 10, 15, etc.) and that the top end of each interval (4, 9, 14, 19, etc.) is the number just below the start of the next interval (5, 10, 15, 20, etc.).

### How are you doing?

1. What is a frequency table?
2. Why would a researcher want to make a frequency table?
3. Make a frequency table for the following scores: 5, 7, 4, 5, 6, 5, 4
4. What does a grouped frequency table group?

**Answers**

**1.** A frequency table is a systematic listing of the number of scores (the frequency) of each value in the group studied.

**2.** A frequency table makes it easy to see the pattern in a large group of scores.

**3.**

| Value | Frequency | Percent |
|---|---|---|
| 4 | 2 | 28.6 |
| 5 | 3 | 42.9 |
| 6 | 1 | 14.3 |
| 7 | 1 | 14.3 |

**4.** A grouped frequency table groups the frequencies of adjacent values into intervals.

## Histograms

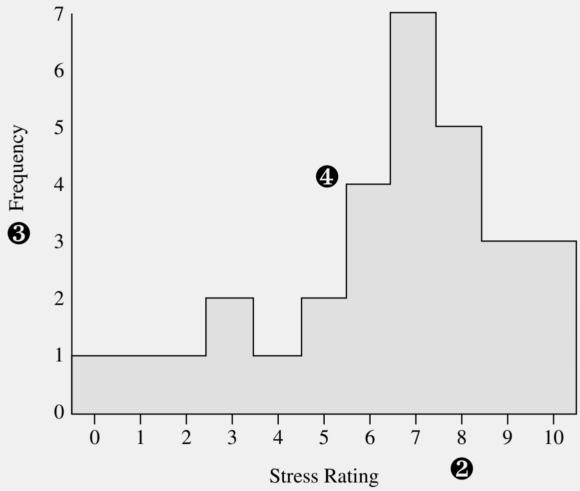
**histogram** Barlike graph of a frequency distribution in which the values are plotted along the horizontal axis and the height of each bar is the frequency of that value; the bars are usually placed next to each other without spaces, giving the appearance of a city skyline.

A graph is another good way to make a large group of scores easy to understand. A picture may be worth a thousand words but it is also sometimes worth a thousand numbers. A straightforward approach is to make a graph of the frequency table. One kind of graph of the information in a frequency table is a kind of bar chart called a **histogram.** In a histogram, the height of each bar is the frequency of each value in the frequency table. Ordinarily, in a histogram all the bars are put next to each other with no space in between. The result is that a histogram looks a bit like a city skyline. Figure 3 shows two histograms based on the stress ratings example, one based on

(a) Frequency Table

❶

| STRESS RATING | FREQUENCY |
|---|---|
| 0 | 1 |
| 1 | 1 |
| 2 | 1 |
| 3 | 2 |
| 4 | 1 |
| 5 | 2 |
| 6 | 4 |
| 7 | 7 |
| 8 | 5 |
| 9 | 3 |
| 10 | 3 |




(b) Grouped Frequency Table

❶

| STRESS RATING INTERVAL | FREQUENCY |
|---|---|
| 0 – 1 | 2 |
| 2 – 3 | 3 |
| 4 – 5 | 3 |
| 6 – 7 | 11 |
| 8 – 9 | 8 |
| 10 – 11 | 3 |

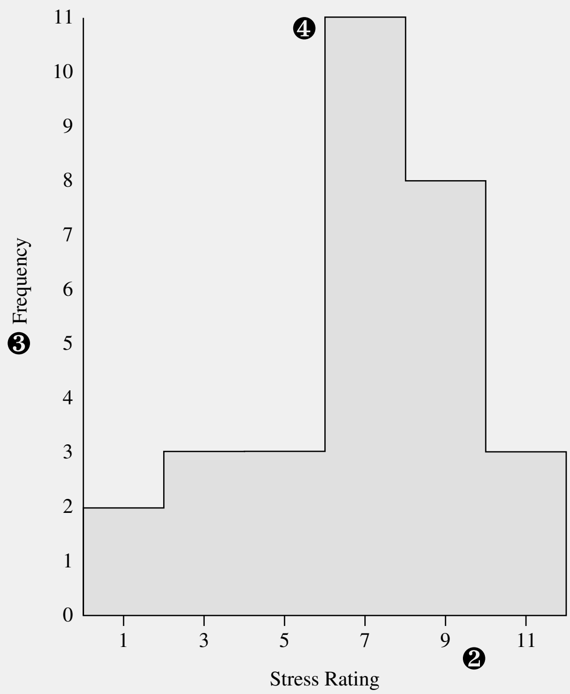



**Figure 3** Four steps in making a histogram based on (a) a frequency table and (b) a grouped frequency table for the stress ratings example. (Data based on Aron, Paris, and Aron, 1995.) ❶ Make a frequency table. ❷ Put the values along the bottom of the page. ❸ Make a scale of frequencies along the left edge of the page. ❹ Make a bar for each value.

| INTERVAL | FREQUENCY |
|---|---|
| 0 – 4 | 12 |
| 5 – 9 | 16 |
| 10 – 14 | 16 |
| 15 – 19 | 16 |
| 20 – 24 | 10 |
| 25 – 29 | 11 |
| 30 – 34 | 4 |
| 35 – 39 | 3 |
| 40 – 44 | 3 |
| 45 – 49 | 3 |

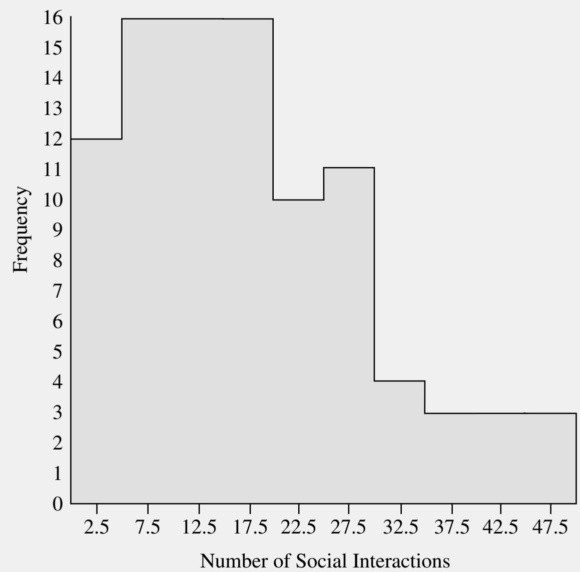



**Figure 4** Histogram for number of social interactions during a week for 94 college students, based on grouped frequencies. (Data from McLaughlin-Volpe et al., 2001.)

the ordinary frequency table and one based on the grouped frequency table. Figure 4 shows a histogram based on the grouped frequency table for the example of the numbers of students' social interactions in a week.

*How to Make a Histogram* There are four steps in making a histogram.

❶ **Make a frequency table (or grouped frequency table).**

❷ **Put the values along the bottom of the page.** The numbers should go from left to right, from lowest to highest. If you are making a histogram from a grouped frequency table, the values you put along the bottom of the page are the interval midpoints. The midpoint of an interval is halfway between the start of that interval and the start of the next highest interval. So, in Figure 4, the midpoint for the 0–4 interval is 2.5, because 2.5 is halfway between 0 (the start of the interval) and 5 (the start of the next highest interval). For the 5–9 interval, the midpoint is 7.5, because 7.5 is halfway between 5 (the start of the interval) and 10 (the start of the next highest interval). Do this for each interval. When you get to the last interval, find the midpoint between the start of the interval and the start of what would be the next highest interval. So, in Figure 4, the midpoint for the 45–49 interval is halfway between 45 (the start of the interval) and 50 (the start of what would be the next interval), which is 47.5.

❸ **Make a scale of frequencies along the left edge of the page.** The scale should go from 0 at the bottom to the highest frequency for any value.

❹ **Make a bar for each value.** The height of each bar is the frequency of the value over which it is placed. For each bar, make sure that the middle of the bar is above its value.

**TIP FOR SUCCESS**

Now try this yourself! Work out the interval midpoints for the grouped frequency table shown in Table 6. Your answers should be the same as the values shown along the bottom of Figure 3b.

**TIP FOR SUCCESS**

You will probably find it easier to make histograms if you use graph paper.