	Type of Data			
Goal	**Measurement (from Gaussian Population)**	**Rank, Score, or Measurement (from Non-Gaussian Population)**	**Binomial (Two Possible Outcomes)**	**Survival Time**
Describe one group	Mean, SD	Median, interquartile range	Proportion	Kaplan Meier survival curve
Compare one group to a hypothetical value	One-sample *t* test	Wilcoxon test	Chi-square or Binomial test **	
Compare two unpaired groups	Unpaired *t* test	Mann-Whitney test	Fisher's test (chi-square for large samples)	Log-rank test or Mantel-Haenszel*
Compare two paired groups	Paired *t* test	Wilcoxon test	McNemar's test	Conditional proportional hazards regression*
Compare three or more unmatched groups	One-way ANOVA	Kruskal-Wallis test	Chi-square test	Cox proportional hazard regression**
Compare three or more matched groups	Repeated-measures ANOVA	Friedman test	Cochrane Q**	Conditional proportional hazards regression**
Quantify association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients**	
Predict value from another measured variable	Simple linear regression or Nonlinear regression	Nonparametric regression**	Simple logistic regression*	Cox proportional hazard regression*
Predict value from several measured or binomial variables	Multiple linear regression* or Multiple nonlinear regression**		Multiple logistic regression*	Cox proportional hazard regression*

One sample t-test

A one sample t-test allows us to test whether a sample mean (of a normally distributed interval variable) significantly differs from a hypothesized value. For example, using the data file, say we wish to test whether the average writing score (**write**) differs significantly from 50. We can do this as shown below.

One sample median test

A one sample median test allows us to test whether a sample median differs significantly from a hypothesized value. We will use the same variable, **write**, as we did in the one sample t-test example above, but we do not need to assume that it is interval and normally distributed (we only need to assume that **write** is an ordinal variable). We will test whether the median writing score (**write**) differs significantly from 50.

Binomial test

A one sample binomial test allows us to test whether the proportion of successes on a two-level categorical dependent variable significantly differs from a hypothesized value. For example, using the data file, say we wish to test whether the proportion of females (**female**) differs significantly from 50%, i.e., from .5. .

Chi-square goodness of fit

A chi-square goodness of fit test allows us to test whether the observed proportions for a categorical variable differ from hypothesized proportions. For example, let's suppose that we believe that the general population consists of 10% Hispanic, 10% Asian, 10% African American and 70% White folks. We want to test whether the observed proportions from our sample differ significantly from these hypothesized proportions.

Two independent samples t-test

An independent samples t-test is used when you want to compare the means of a normally distributed interval dependent variable for two independent groups. For example, using the data file, say we wish to test whether the mean for **write** is the same for males and females.

Wilcoxon-Mann-Whitney U test

The Wilcoxon-Mann-Whitney test is a non-parametric analog to the independent samples t-test and can be used when you do not assume that the dependent variable is a normally distributed interval variable (you only assume that the variable is at least ordinal). We will use the same data file and the same variables in this example as we did in the independent t-test example above and will not assume that **write**, our dependent variable, is normally distributed.

Chi-square test

A chi-square test is used when you want to see if there is a relationship between two categorical variables. Using the data file, let's see if there is a relationship between the type of school attended (**schtyp**) and students' gender (**female**). Remember that the chi-square test assumes the expected value of each cell is five or higher. This assumption is easily met in the examples below. However, if this assumption is not met in your data, please see the section on Fisher's exact test below.

Fisher's exact test

The Fisher's exact test is used when you want to conduct a chi-square test, but one or more of your cells has an expected frequency of five or less. Remember that the chi-square test assumes that each cell has an expected frequency of five or more, but the Fisher's exact test has no such assumption and can be used regardless of how small the expected frequency is. In the example below, we have cells with observed frequencies of two and one, which may indicate expected frequencies that could be below five, so we will use Fisher's exact test.

One-way ANOVA

A one-way analysis of variance (ANOVA) is used when you have a categorical independent variable (with two or more categories) and a normally distributed interval dependent variable and you wish to test for differences in the means of the dependent variable broken down by the levels of the independent variable. For example, using the data file, say we wish to test whether the mean of **write** differs between the three program types (**prog**). The command for this test would be:

Kruskal Wallis test

The Kruskal Wallis test is used when you have one independent variable with two or more levels and an ordinal dependent variable. In other words, it is the non-parametric version of ANOVA and a generalized form of the Mann-Whitney test method since it permits 2 or more groups. We will use the same data file as the one way ANOVA example above and the same variables as in the example above, but we will not assume that **write** is a normally distributed interval variable.

Paired t-test

A paired (samples) t-test is used when you have two related observations (i.e. two observations per subject) and you want to see if the means on these two normally distributed interval variables differ from one another. For example, using the data file we will test whether the mean of **read** is equal to the mean of **write**.

Wilcoxon signed rank sum test

The Wilcoxon signed rank sum test is the non-parametric version of a paired samples t-test. You use the Wilcoxon signed rank sum test when you do not wish to assume that the difference between the two variables is interval and normally distributed (but you do assume the difference is ordinal). We will use the same example as above, but we will not assume that the difference between **read** and **write** is interval and normally distributed.

Correlation

A correlation is useful when you want to see the linear relationship between two (or more) normally distributed interval variables. For example, using the data file we can run a correlation between two continuous variables, **read** and **write**.

Non-parametric correlation

A Spearman correlation is used when one or both of the variables are not assumed to be normally distributed and interval (but are assumed to be ordinal). The values of the variables are converted in ranks and then correlated. In our example, we will look for a relationship between **read** and **write**. We will not assume that both of these variables are normal and interval .

Simple linear regression

Simple linear regression allows us to look at the linear relationship between one normally distributed interval predictor and one normally distributed interval outcome variable. For example, using the data file, say we wish to look at the relationship between writing scores (**write**) and reading scores (**read**); in other words, predicting **write** from **read**.

Multiple regression

Multiple regression is very similar to simple regression, except that in multiple regression you have more than one predictor variable in the equation. For example, using the data file we will predict writing score from gender (**female**), reading, math, science and social studies (**socst**) scores.

CHOOSING BETWEEN PARAMETRIC AND NONPARAMETRIC TESTS: Choosing between parametric and nonparametric tests is sometimes easy. You should definitely choose a parametric test if you are sure that your data are sampled from a population that follows a Gaussian distribution (at least approximately). You should definitely select a nonparametric test in three situations:

- The outcome is a rank or a score and the population is clearly not Gaussian. Examples include class ranking of students, the Apgar score for the health of newborn babies (measured on a scale of 0 to IO and where all scores are integers), the visual analogue score for pain (measured on a continuous scale where 0 is no pain and 10 is unbearable pain), and the star scale commonly used by movie and restaurant critics (* is OK, ***** is fantastic).
- Some values are "off the scale," that is, too high or too low to measure. Even if the population is Gaussian, it is impossible to analyze such data with a parametric test since you don't know all of the values. Using a nonparametric test with these data is simple. Assign values too low to measure an arbitrary very low value and assign values too high to measure an arbitrary very high value. Then perform a nonparametric test. Since the nonparametric test only knows about the relative ranks of the values, it won't matter that you didn't know all the values exactly.
- The data ire measurements, and you are sure that the population is not distributed in a Gaussian manner. If the data are not sampled from a Gaussian distribution, consider whether you can transformed the values to make the distribution become Gaussian. For example, you might take the logarithm or reciprocal of all values. There are often biological or chemical reasons (as well as statistical ones) for performing a particular transform.

ONE- OR TWO-SIDED P VALUE?

With many tests, you must choose whether you wish to calculate a one- or two-sided P value (same as one- or two-tailed P value). The difference between one- and two-sided P values was discussed in Chapter 10. Let's review the difference in the context of a t test. The P value is calculated for the null hypothesis that the two population means are equal, and any discrepancy between the two sample means is due to chance. If this null hypothesis is true, the one-sided P value is the probability that two sample means would differ as much as was observed (or further) in the direction specified by the hypothesis just by chance, even though the means of the overall populations are actually equal. The two-sided P value also includes the probability that the sample means would differ that much in the opposite direction (i.e., the other group has the larger mean). The two-sided P value is twice the one-sided P value.

A one-sided P value is appropriate when you can state with certainty (and before collecting any data) that there either will be no difference between the means or that the difference will go in a direction you can specify in advance (i.e., you have specified which group will have the larger mean). If you cannot specify the direction of any difference before collecting data, then a two-sided P value is more appropriate. If in doubt, select a two-sided P value.

If you select a one-sided test, you should do so before collecting any data and you need to state the direction of your experimental hypothesis. If the data go the other way, you must be willing to attribute that difference (or association or correlation) to chance, no matter how striking the data. If you would be intrigued, even a little, by data that goes in the "wrong" direction, then you should use a two-sided P value. For reasons discussed in Chapter 10, I recommend that you always calculate a two-sided P value.

PAIRED OR UNPAIRED TEST?

When comparing two groups, you need to decide whether to use a paired test. When comparing three or more groups, the term paired is not apt and the term repeated measures is used instead.

Use an unpaired test to compare groups when the individual values are not paired or matched with one another. Select a paired or repeated-measures test when values represent repeated measurements on one subject (before and after an intervention) or measurements on matched subjects. The paired or repeated-measures tests are also appropriate for repeated laboratory experiments run at different times, each with its own control.

You should select a paired test when values in one group are more closely correlated with a specific value in the other group than with random values in the other group. It is only appropriate to select a paired test when the subjects were matched or paired before the data were collected. You cannot base the pairing on the data you are analyzing.

FISHER'S TEST OR THE CHI-SQUARE TEST?

When analyzing contingency tables with two rows and two columns, you can use either Fisher's exact test or the chi-square test. The Fisher's test is the best choice as it always gives the exact P value. The chi-square test is simpler to calculate but yields only an approximate P value. If a computer is doing the calculations, you should choose Fisher's test unless you prefer the familiarity of the chi-square test. You should definitely avoid the chi-square test when the numbers in the contingency table are very small (any number less than about six).

When the numbers are larger, the P values reported by the chi-square and Fisher's test will he very similar.

The chi-square test calculates approximate P values, and the Yates' continuity correction is designed to make the approximation better. Without the Yates' correction, the P values are too low. However, the correction goes too far, and the resulting P value is too high. Statisticians give different recommendations regarding Yates' correction. With large sample sizes, the Yates' correction makes little difference. If you select Fisher's test, the P value is exact and Yates' correction is not needed and is not available.

REGRESSION OR CORRELATION?

Linear regression and correlation are similar and easily confused. In some situations it makes sense to perform both calculations. Calculate linear correlation if you measured both X and Y in each subject and wish to quantity how well they are associated. Select the Pearson (parametric) correlation coefficient if you can assume that both X and Y are sampled from Gaussian populations. Otherwise choose the Spearman nonparametric correlation coefficient. Don't calculate the correlation coefficient (or its confidence interval) if you manipulated the X variable.

Calculate linear regressions only if one of the variables (X) is likely to precede or cause the other variable (Y). Definitely choose linear regression if you manipulated the X variable. It makes a big difference which variable is called X and which is called Y, as linear regression calculations are not symmetrical with respect to X and Y. If you swap the two variables, you will obtain a different regression line. In contrast, linear correlation calculations are symmetrical with respect to X and Y. If you swap the labels X and Y, you will still get the same correlation coefficient.

What is a null hypothesis?

When statisticians discuss P values, they use the term null hypothesis. The null hypothesis simply states that there is no difference between the groups. Using that term, you can define the P value to be the probability of observing a difference as large or larger than you observed if the null hypothesis were true.

Statistical significance in science

The term significant is seductive, and easy to misinterpret. Using the conventional definition with alpha=0.05, a result is said to be statistically significant when the result would occur less than 5% of the time if the populations were really identical. It is easy to read far too much into the word significant because the statistical use of the word has a meaning entirely distinct from its usual meaning. Just because a difference is statistically significant does not mean that it is biologically or clinically important or interesting result that is not statistically significant (in the first experiment) may turn out to be very important.

If a result is statistically significant, there are two possible explanations: The populations are identical, so there really is no difference. By chance, you obtained larger values in one group and smaller values in the other. Finding a statistically significant result when the populations are identical is called making a Type I error. If you define statistically significant to mean “P<0.05”, then you’ll make a Type I error in 5% of experiments where there really is no difference. The populations really are different, so your conclusion is correct. The difference may be large enough to be scientifically interesting. Or it may be tiny and trivial. "Extremely significant" results Intuitively, you may think that P=0.0001 is more statistically significant than P=0.04. Using strict definitions, this is not correct. Once you have set a threshold P value for statistical significance, every result is either statistically significant or is not statistically significant. Degrees of statistical significance are not distinguished.

Some statisticians feel very strongly about this. Many scientists are not so rigid, and refer to results as being "barely significant", "very significant" or "extremely significant". Prism summarizes the P value using the words in the middle column of this table. Many scientists label graphs with the symbols of the third column. These definitions are not entirely standard. If you report the results in this way, you should define the symbols in your figure legend.

P value	Wording	Summary
>0.05	Not significant	ns
0.01 to 0.05	Significant	*
0.001 to 0.01	Very significant	**
< 0.001	Extremely significant	***