The following clearly goes beyond the scope of our course, but may be of interest to some who have more advanced training or specific interest in tests.  You will no doubt see some of these things as you go forward in the program.  I wrote this fairly quickly and with simplification in mind, so I may have 'fudged' some technical details.

**Multivariate Hypothesis Tests**

Used to determine if a collection several variables belong in a model. These are commonly used in a regression context. The F-test for joint significance that is produced by essentially all regression software would be an example.

Example Regression Model:

As said above, the most common use of this test is to see if a regression model contributes anything to the understanding of the dependent variable it is trying to explain.  The general regression model looks like:

Y = B0 + B1 X1 + B2 X2 + … Bk Xk + error

Regression software runs a standard F-test of the null hypothesis:

H0: B1 = B2 = … = Bk = 0

H1: at least one of B1..Bk <> 0

If the data cannot support the rejection of the null hypothesis, that means that the data cannot assert that any of the independent variables actually predict the dependent variable, so the model collectively doesn’t explain anything.

Note that B0 – the term associated with the intercept is NOT included, because the B0 is just the average of what is left over after the other factors have been taken into consideration – that is not really an explanatory factor.

**While that example is the most commonly occurring one, there is a general principle at work here: If you want to test something, you need to build a model that allows the data to express the relationship you are interested in testing and then you check the results to see if the data shows that result.**

For example, one might build a regression model to determine the amount sold based on price and dollars of advertising spent on different channels to address the managerial question “do investment in different channels have the same returns?”

Such a model might look something like:

Sales = B0 + B1 Price + B2 Ad\_Print + B3 Ad\_Budget\_Facebook + B4 Ad\_Budget\_TV + B5 Ad\_Budget\_Signage + Other\_Terms + Error

The test could take the form:

H1: <Conclude they are not equally effective> if <data shows that the impacts are different>

H0: B2 = B3 = B4 = B5

H1: Not (B2 = B3 = B4 = B5)

If one rejects H0, one would conclude that at least one of the B2..B5 terms are not the same as the others, which would imply that the impact of advertising in at least one of the areas was different from the others.  You could then do more analysis to find out which one is different, how and why.

Similar analysis could be done to determine if groups are the same for example, you could build a model for the impact of gender on sales and prices:

Sales = B0 + B1 Income + BM Male + B1M Male\_Income + BA Ad\_budget + Other\_terms + error

Where Male is a dummy variable = 1 if male and 0 otherwise; Male\_Income = Male \* Income.

Here we could test:

H0: BM = BM1 = 0

H1: at least one of BM and BM1 <> 0

Here if you fail to reject the null, you would effectively conclude that the data does not show a difference between the male and non-male elements of your population.  In general this could be because the difference doesn’t exist; the difference exists but you lack the power to detect the difference; or the difference exists but it is captured by another relationship.

**Specification Tests**

Used to determine if models are properly constructed. For example, one might attempt to determine if the error term in a regression model is ‘well behaved’. A test for the existence of heteroscedasticity or autocorrelation would fall into this category.

There must be a zillion specification tests out there.  I will comment on two of them, a test for heteroscedasticity and a test for autocorrelation.

**Heteroskedasticity**:

One common one deals with heteroscedasticity vs. homoscedasticity.  These issues will be addressed in later courses, but for now, it suffices to say that the error terms in a standard regression model are supposed to be independent and identically distributed.  Amongst other things, this means that they have the same variance (which is to say they are homoscedastic).

To understand what this means, you could imagine a simple regression model:

Spend = B0 + B1 Income + Error

Where Spend = Money spent on pizza and Income = weekly income.  Using linear regression to estimate this relationship might give rise to the following function:

Sales = 10 + 0.01 Income + Error

Which would be to say, we predict that an individual’s spend on pizza per week was an average of $10 + 1% of their weekly income.  The function is estimated with an error, which is to be expected.

Now suppose the error’s variability increased with income, so that you have:

Error ~ N(Mu\_Error = 0; Sigma\_Error = 1 + Income / 1000)

This would mean that the spend by high-income earners would be more erratic than those of low income earners.  To test this conceptually we would want to test:

H0: All error terms have same variance

H1: At least some of the error terms have a different variance

Describing the math involved would introduce complexities and detract from the high-level summary I intend here so I will leave that for other classes.

One intuitive implication of having higher error variance for some observations than others is that those high-variance observations have relatively more ‘noise’ than the other observations, so one might think that they should be given less weight in estimating the relationship.  That is exactly what one does when one detects heteroscedasticity.  (For those of you who have taken more advanced courses, this is exactly what is being done in feasible generalized least squares.)

**Autocorrelation:**

Autocorrelation occurs when time series data has patterns where the observation at one point of time affects the value at another time.  These patterns across time can be used to create predictions of future events based on stable patterns.  Using these patterns to make predictions falls into the domain of ‘time series analysis’.  It is a very large field so what I’ll say here is just a high-level simplification.

Suppose that many people prefer to buy a car when the warranty period ends – which we could assume takes place 5 years into car ownership.  We could imagine that people want to buy cars within the few months around 5 years into ownership.

Now imagine that, for whatever reason, the economy is doing very well, so people who are thinking about buying a new car buy them a bit earlier than they typically would.  Not everyone, of course, but perhaps 50% of people who would have bought a car in a 3-month period buy it in the first of those three months.  This would result in a small ‘spike’ in car sales in the first month of that quarter.

This spike would have two impacts.  Since ‘too many people’ bought cars in the first month of that quarter, we would expect sales to be a bit lower for the rest of that quarter.  This would mean that the higher than anticipated sales in the first month would be followed by a somewhat muted but lower-than-anticipated sales for the rest of the quarter.  This would be a negative autocorrelation (see what I did there: ‘auto-correlation' and car sales!)

Beyond those immediate effects, you might expect that 5 years later, many of those people who bought cars early would again be on the market for a new car.  This could show up in the data as a spike in sales in five years’ time when warranties end.  You would expect the spike to be smaller than the original one, but positive.  This would be a positive autocorrelation.

Testing for and modelling autocorrelation and other time-series patterns is a complex business, but conceptually it works the same way: you develop a model to allow the data to express the relationship you believe might be there, then you test to see if the relationship actually is there.  If the relationship exists, you adjust your model or modelling approach to take advantage of that relationship.

**Tests for Structural Change**

Used to determine if time or geographic differences change the way that explanatory factors predict outcomes. For example, the relationship between the price of a house and its number of rooms, size and location may have changed significantly with the financial crisis of 2008.

The easiest tests for structural change are simply multivariate tests as described above, only here the data is grouped in terms of before and after the structural change.  So where we were looking for the difference between ‘male’ and ‘non-males’ above, alternatively, we could build a model that allowed all the terms before the 2008 financial crisis to change after 2008 and then test if any of them did.

Consider the model:

Sales = B0 + B1 Age + B2 Income + Error

You might suspect that the relationship changed after 2008, so to test that you could develop the model:

Sales = B0 + B1 Age + B2 Income + B0\_d + B1\_d Age + B2\_d Income + Error

You could then run the test:

H0: B0\_d = B1\_d = B2\_d = 0

H1: at least one of B0\_d .. B2\_D <> 0

If you reject H0, then at least one of the parameters has changed.  The Chow test is the standard way of comparing groups that might have different relationships.  The change does not need to be over time, it could be between groups, such as Canadians vs. Americans.

**Tests for Parameters other than the Mean**

We have hinted at this already when we discussed the weighted vs. pooled tests for two independent variables. That would be a test for variance, but tests for other parameters exist as well.

One area where this might come up is in the test for ‘equality of variance’ in hypothesis testing.  (Note this is equivalent to testing whether the standard deviations are the same, but the mathematics of the test are specified in terms of variances.)

When doing two-population tests with independent data, you have to assume equal variance or unequal variance.  When you are uncertain, I suggest assuming unequal variance because it is a bit more robust than assuming equal variance.  But in principle you could test it.

Suppose you are testing two populations, X and Y with at est:

H0: Mu\_X = Mu\_Y

H1: Mu\_X <> Mu\_Y

To test if the variances are the same, you would set up the test with:

H0: Variance\_X = Variance\_Y

H1: Variance\_X <> Variance\_Y

The technical details again are not relevant to us since (a) there are several versions of tests like this and (B) one or more versions of these tests are built into most sophisticated software.  The tests will produce a p-value that can be interpreted in the same way that we have for other tests and can guide the choice of equal vs. unequal variance in your main test.

**Special purpose tests**

These have been developed for a wide variety of characteristics. For example, there are tests for characteristics like autocorrelation or cointegration.

There are any number of special purpose tests.  When I was doing my MA in Economics, all those years ago, I was working on ‘structural change consistent tests for cointegration’, so I’ll tell you a bit about that.

Cointegration is property that some time-series data has.  It pertains to a situation where two or more data sets have a random component over time (e.g. a random walk) but where the data actually moves in a way that they do not get too far apart.

For example, the price of stocks is often described as following a random walk (normally with a drift) so two stocks may appear to be moving randomly.  In spite of that apparent random movement, the stock’s prices may have some hidden mechanism that prevents them from getting too far apart.  Each individual stock price change appears random but somehow they never drift too far from each other – that may well be a result of cointegration.

I sometimes imagine this as a man walking a dog using one of those retractable leashes.  The dog, let’s call him 'Banksy', might move one way or another to sniff or explore something at any point in the walk.  The man walking the dog might also step one way or another.  Each appears to be following his own path with randomness.  The best guess of where either of them will be in the immediate future is pretty much where they are now plus one random step down the path.

If randomly walking was the entire description, you should expect that those random steps would accumulate and the dog and the man would get arbitrarily far apart.  That doesn't happen, however, as they get further apart, the leash tends to pull them back together.  Each step is a fee choice, but collectively they do not get too far apart because there is a process that draws them back together.

In the real world of time series analysis, you cannot see any kind of retractable leash pulling the stock prices back together, but the fact that they do not drift arbitrarily far apart suggests some process is at work.  This process could be described as cointegration.

The mathematics of cointegration and the associated tests are the domain of time series analysis. They tend to be quiet complex and really do require specialist knowledge to use.

**Nonparametric tests**

These are really a category of tests that have been developed so that they do not rely on assumptions of distributions for particular parameters, such as assuming normality. These tests are often more robust – they apply more widely – but are generally less powerful.

Nonparametric tests are not really ‘tests’ but are a testing methodology.  They tend to be more robust because they use fewer assumptions (e.g. they generally do not rely on CLT or assumptions of normality) but have less power.

The good news is that there are many standard nonparametric tests built into software.  For example, the Wilcoxon signed-rank test (see Wikipedia) is a non-parametric test that can be used as an alternative to the paired t-test.

My experience suggests that practitioners in different domains tend to use and become knowledgeable about specific nonparametric tests that they use frequency.  You will probably encounter some of these tests in more domain specific courses, e.g. marketing analytics.

**Bootstrapping tests**

These are a relatively new, calculation-intensive category of tests that rely on resampling from data rather than arguments based on a CLT to create the distribution used to determine probability. They are quite robust and powerful but are not widely established in statistical software yet.

As with nonparametric tests, bootstrapping is a testing methodology, not a specific test.  They rely on using the existing sample data, a lot of computer power, and some clever design of sampling to calculate tests statistics.  They can be incredibly powerful, robust and flexible, which is great; unfortunately, they are not standard so to use them may require the design of a specific test for each situation.

I am a big fan of bootstrapping tests, but they are clearly beyond an introductory (or even second course) in statistics.  Later in the program, if you are interested, I can provide some lecture notes I have written on the topic.

I hope that helps a bit -- if you didn't follow it all, don't worry, this touches on some pretty advanced stuff.