The following resources are associated:
'Opening and labelling data in SAS', 'Summarising categorical variables in SAS'

Testing proportions in SAS

Common Applications:

- Testing for a difference between the proportions observed in a sample of data and those in the wider population.
- Or testing whether the data fits a known probability distribution such as binomial

Data type: Binary (two values/groups) **Test:** Test of proportion

Data type: Nominal (3+ values/groups) **Test:** Chi-squared goodness of fit

Data: Visits to stats support are recorded and demographics of each student who visits are used to assess which groups of students are using or not using the service. In order to do this, proportions observed within the service data are compared to the University demographics.
Before running a test it is advisable to check how many people are in each group and ensure categories are classified by numbers rather than words to avoid errors in SAS. For example, Males are coded as 0's and females as 1's. Ethnicity has many options so check how many are in each group and consider combining categories to make larger groups. Here a new binary variable BAME has been created with White=0 and all other groups as 1. For disability, four categories were retained: 0=no known disability, 1=mental health condition, 2=learning difficulty or autistic spectrum disorder, 3=Physical health condition.

Data

Disability	Disability type	Gender	Female	Ethnicity	BAME	Discipline	Department
NO KNOWN DISABILITY	0	Male	0	WHITE	0	2	Psychology Sociology and Politics
MULTIPLE DISABILITIES	3	Female	1	WHITE	0	2	Psychology Sociology and Politics
NO KNOWN DISABILITY	0	Female	1	WHITE	0	2	Psychology Sociology and Politics
MENTAL HEALTH DIFFICULTIES	1	Female	1	WHITE	0	2	Psychology Sociology and Politics
NO KNOWN DISABILITY	0	Female	1	WHITE	0	1	Engineering and Mathematics
UNSEEN DISABILITY EG DIABETES	3	Male	0	Asian - Pakistani	1	1	Engineering and Mathematics

The value labels can be attached by setting up formats using **`proc format`** which can then be applied to any variable within a format line in procedure commands so that output displays labels rather than numbers. For example, set up a YesNo format which can be used with binary variables such as BAME or whether or not someone has a mental health.

```
proc format;
value YesNo 1='Yes' 0='No';
value dis 0='No disabiity' 1='Mental health' 2='Learning difficulty'
3='Physical health';
```

Read the data into SAS and use the data step to apply variable labels.

```
data visits;
set mydata.statssupport_visits;
label Disability_type='Type of disability';
run;
```

Tests for categorical variables such as test for proportions and chi-squared are contained in the **proc freq** procedure which also produces frequency tables.
Tables is the command used within **proc freq** to create a table with additional actions such as tests being carried out or different types of percentages being requested.
When testing one proportion use the binomial command and specify the population proportion **P** being tested (p=**P** level=**2**). For more than two use the chi-squared chisq command.

format will use the formats that were created earlier so that actual names are used within the tables instead of values. Put the variable name first followed by the format name defined earlier and a full stop to indicate it's a format not a variable. format bame YesNo.;

Testing one binary variable

Example 1: Are females more likely to use stats support compared to the university proportion?
57% of students at Sheffield Hallam are female and our gender variable has two levels; males = 0 and females =1. You test the proportion of the higher value which is female here (males =0, females=1) so the proportion we are testing against is 0.57.

```
proc freq data=visits;
tables Female/binomial(p=0.57 level=2);
format Female YesNo.;
run;
```

Female	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	347	30.04	347	30.04
Yes	808	69.96	1155	100.00
Frequency Missing = 34				

Binomial Proportion Female = Yes	
Proportion	0.6996
ASE	0.0135
95% Lower Conf Limit	0.6731
95% Upper Conf Limit	0.7260
Exact Conf Limits	
95% Lower Conf Limit	0.6722
95% Upper Conf Limit	0.7259

Test of H0: Proportion = 0.57	
ASE under H0	0.0146
Z	8.8943
One-sided Pr > Z	<.0001
Two-sided Pr > |Z|	<.0001

Effective Sample Size = 1155
Frequency Missing = 34

From the frequency table above, 70% of the users of stats support were female compared to 57% in the general population.

The first table in the test output gives a 95% confidence interval for the population proportion of stats support users of 67% - 73% which is considerably higher than the university %.

For very small sample sizes, you can use the exact limits instead.

The test for a proportion use a normal approximation so the p-values can be calculated using the Z distribution.

A test of proportion showed significant evidence (p<0.0001) of a difference between the proportions of females using stats support and the university population. Whilst 57% of Sheffield Hallam students are female, 70% [95% CI: 67%, 73%] of the users of stats support are female. This indicates that females are more likely to use statistics support than males.

Considerations: Not all courses at Sheffield Hallam study statistics and the users of the service are more likely to come from certain disciplines such as Psychology and Maths where more stats content is covered within the degrees. Therefore it may be unfair to compare across the whole course especially for aspects such as gender which vary by discipline. In Psychology for example, 75% of students are female. If you wish to test specific groups, use the where command to select a subgroup.

```
proc freq data=visits;
tables Female/binomial(p=0.75 level=2);
where Discipline=2;
```

Binomial Proportion Female = Yes	
Proportion	0.8588
ASE	0.0189
95% Lower Conf Limit	0.8218
95% Upper Conf Limit	0.8958

Test of H0: Proportion = 0.75	
ASE under H0	0.0235
Z	4.6341
One-sided Pr > Z	<.0001
Two-sided Pr > |Z|	<.0001

Sample Size = 340

Still significantly more females use stats support than males within the Psychology department.

Carrying out a chi-squared goodness of fit test

If there are more than two categories to test against population proportions, use a goodness of fit test. Chi-squared tests compare the observed frequencies from the data with frequencies which would be expected if the proportions match the population values.
Example 2: Are students with disabilities more or less likely to engage with stats support?
In the university population, 72% of students do not have a registered disability or condition, 12% have a mental health condition, 6% a learning difficulty and 10% a physical disability.

A chi-squared goodness of fit test will test whether these proportions fit a distribution similar to the University proportions. It is carried out by adding `chisq` to the tables row within **`proc freq`**.
Add the percentages from the population without the % sign the order of the numerical categories using **`testp=(`**`72 12 6 10`**`)`**. If you have categories with labels they will appear alphabetically.

```
proc freq data=visits;
tables Disability_type/nocum chisq testp=(72 12 6 10)
                  plots=deviationplot;
                        format Disability_type dis.;
```

This table summarises the actual percentages from the data and the hypothesised percentages. Make sure each of the categories have a decent sample size with ideally no frequencies less than 5. You will need to exclude or combine categories with small frequencies for the chi-squared test to be valid. None of the differences between sample and population percentages are particularly large although more students with learning difficulties access stats support.

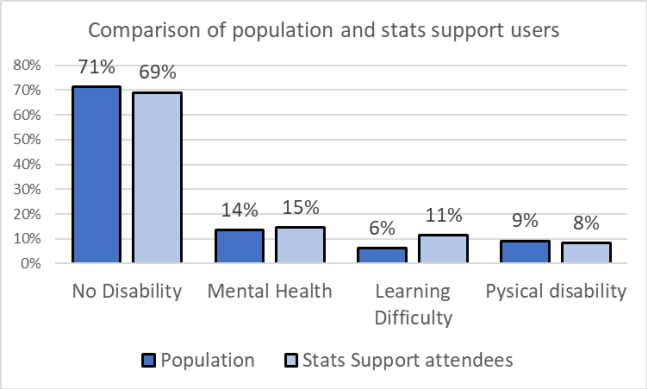
Type of disability			
Disability_type	Frequency	Percent	Test Percent
No disabiity	802	69.44	72.00
Mental health	104	9.00	12.00
Learning difficulty	124	10.74	6.00
Physical health	125	10.82	10.00
Frequency Missing = 34			

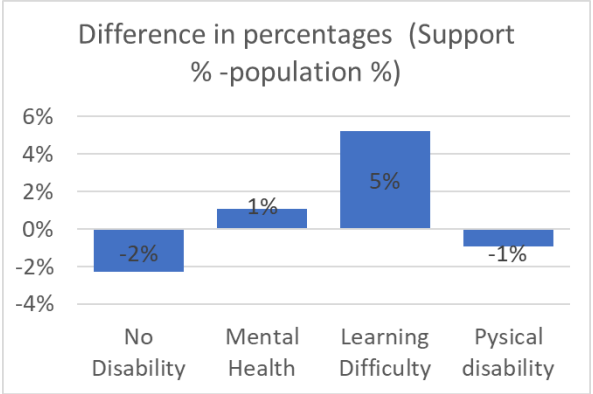
Chi-Square Test for Specified Proportions	
Chi-Square	53.6484
DF	3
Pr > ChiSq	<.0001

The Chi-squared goodness of fit test degrees of freedom is the number of categories -1 and the test statistic compares the observed and expected frequencies. The expected frequencies are calculated multiplying the hypothesised proportions by the sample size. Chi-squared tests are only valid when you have reasonable sample size, less than 20% of expected values are less than 5 and none have an expected count less than 1.

Report the results: *A chi-squared goodness of fit test showed significant evidence* $(\chi^2(3) = 53.6, p < 0.0001)$ *to suggest that the percentages from disability groups using stats support differ from the general population.*

For significant results follow up with an explanation of differences. Although you can get some automatic charts from SAS, Excel has more flexibility with how you present the data. You can compare the support and uni percentages next to each other or present differences between them.





Analysing data already grouped into a table

It is also possible to analyse summary data by stating the category numbers and frequencies. The `input` line states the variable names and the `datalines` requires the group codes and frequencies.

```
data SSmini;
input group frequency;
datalines;
0 802
1 104
2 124
3 125
run;
```

	group	frequency
1	0	802
2	1	104
3	2	124
4	3	125

The `weight` command must be used within any procedure to tell SAS that this is summary rather than raw data and that the `frequency` column indicates the number of individuals for a particular combination of groups. For example, there are 802 rows of students with no disability etc. Repeat the Chi-squared steps with the additional `weight` row to get the standard output and interpret in the same way.

```
proc freq data=SSmini;
tables group/nocum chisq testp=(72 12 6 10) ;
weight frequency;
format group dis.;
run;
```

Summarising multiple similar tests

If you are carrying out multiple tests of the same type, using a summary table or chart is preferable. Transfer summary statistics to Excel to allow multiple variables to be summarised in one chart or table as in the examples here. Then briefly describe key differences.

Demographic Data	Population	Stats Support attendees	Difference (Support-population)	P Value
BAME	29%	25%	-3%	0.0237
Male	26%	14%	-12%	<0.001
Disability	29%	42%	13%	<0.001
Mental Health	14%	15%	1%	
Learning Difficulty	6%	11%	5%	
Pysical disability	9%	8%	-1%	

The first chart contains the actual percentages of the stats support and University population for Psychology students and the second chart uses the differences between observed and hypothesised percentages with negative values indicating fewer are seen in stats support.