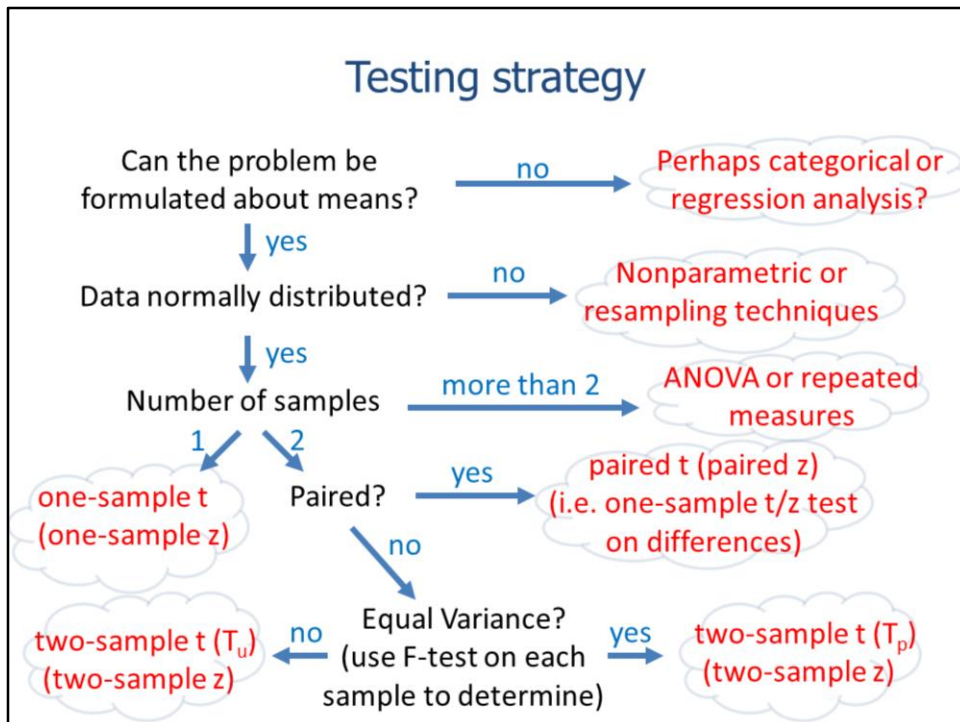


Testing for differences among datasets

Statistical analysis in python

1

This video will discuss some scipy tools that test for statistical differences among datasets.



This slide shows a general strategy for determining an appropriate statistical test to use for an analysis. This video will discuss scipy's tools testing for differences among datasets. These tools may require that the datasets tested fit a certain distribution – see the 12b lecture video for testing distribution types.

Kolmogorov-Smirnov test for two continuous samples

- Test of goodness of fit for two independent samples

```
>>> stats.ks_2samp(a = dataLst1, b = dataLst2)
```

(0.079, 0.411)

D-statistic

p-value

Note: Input datasets can have different lengths

- Null hypothesis is that the two distributions are identical
 - difference is significant if p-value is less than threshold

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html#scipy.stats.ks_2samp

3

The 2 sample KS test compares the distributions of two datasets. The test can be used for datasets with any type of distribution.

A significant p-value indicates that the two datasets are not identical.

One sample t-test

- Tests if mean is different than an expected value

```
>>> stats.ttest_1samp(dataLst, popmean = 0)
(array(-0.8800), 0.3809)
```

t-statistic p-value
(two tailed) expected value in
null hypothesis

- Difference is significant if p-value is less than threshold (e.g. 0.05)

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_1samp.html#scipy.stats.ttest_1samp

4

The one sample t-test tests if the mean of a dataset is different than an expected value. The test assumes the dataset has a normal distribution.

A significant p-value indicates that dataset mean is different than the expected mean.

Paired t-test

- t-test for two related samples - tests if the differences between pairs is significantly different from zero...

```
>>> stats.ttest_rel(dataLst1, dataLst2)
```

```
(array(-0.65126), 0.51517)
```

t-statistic

p-value

Note: input datasets must
have same lengths

- Difference is significant if p-value is less than threshold (e.g. 0.05)

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html#scipy.stats.ttest_rel

5

A paired t-test tests for differences between two related datasets. The two datasets are paired and must have the same number of data points. The test assumes the paired differences have a normal distribution.

A significant p-value indicates that the differences between the pairs is not zero.

Two sample t-test

- Test for difference between means of two independent samples...

```
>>> stats.ttest_ind(dataLst1, dataLst2, equal_var = True)
```

```
(array(-0.9198), 0.3578)
```

t-statistic

p-value

assume equal variance?
(default is True)

Note: input datasets can
have different lengths

- Differences between means is significant if p-value is less than threshold (e.g. 0.05)

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html

6

The two sample t-test tests for differences in the means for two independent datasets. The datasets do not need to have the same number of data points. The test can be done for datasets that either have equal variance or unequal variance. This test assumes both datasets have normal distributions.

A significant p-value indicates that the two datasets have different means.

Analysis of Variance (ANOVA)

- Tests for differences in means among 3 or more independent samples

```
>>> stats.f_oneway(dataLst1, dataLst2, ..., dataLstN)
```

```
(28.9207, 6.7335e-13)
```

F-statistic

p-value

Note: Input datasets can
have different lengths

```
>>> dataLsts = [dataLst1, dataLst2, ..., dataLstN]
```

```
>>> stats.f_oneway(*dataLsts)
```

} alternate
syntax

- Differences among means are significant if p-value is less than threshold (e.g. 0.05)

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html#scipy.stats.f_oneway

7

ANOVA tests for differences among the means of 3 or more independent datasets. The input datasets can contain different numbers of values; however, each dataset is assumed to have a normal distribution. All datasets are assumed to have an equal variance.

Note a 2-level list may be specified for the `f_oneway` test – this is particularly useful if the script will need to run ANOVA on an undetermined number of datasets.

A significant p-value indicates that at least one of the datasets has a mean that is significantly different from the others.

Bartlett test for equal variances

- Test for equal variances among 3 or more samples...

```
stats.bartlett(dataLst1, dataLst2, ..., dataLstN)
```

```
(189.003, 9.091e-42)
```

t-statistic

p-value

Note: input datasets can
have different lengths

```
>>> dataLsts = [dataLst1, dataLst2, ..., dataLstN]
```

```
>>> stats.bartlett(*dataLsts)
```

} alternate
syntax

- Difference in variances are significant if p-value is less than threshold (e.g. 0.05)

<http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bartlett.html#scipy.stats.bartlett>

8

The Bartlett test can be used to test for equal variances among 3 or more datasets.

A significant p-value indicates that at least one dataset has a variance that does not equal the others.

Wilcoxon signed rank test - a non-parametric paired t-test

- Tests for difference between two paired samples...

```
>>> stats.wilcoxon(dataLst1, dataLst2, zero_method = 'pratt')
```

(21219.0, 0.36719)

↑
sum of ranks

↑
p-value
(two-tailed)

↑
optional

↑
other treatment options:
'wilcox'; 'zsplit'

- Difference is significant if p-value is less than threshold
- If only `dataLst1` specified, it should be the difference between the pairs of the two datasets.

<http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html#scipy.stats.wilcoxon>

9

The Wilcoxon signed rank test is equivalent to a paired t-test except that it can be used on datasets that are not normally distributed (i.e. non-parametric). One or two datasets may be specified for the **wilcoxon** method. If two datasets are specified, then they should have equal numbers of values. If one dataset is specified, it should be the differences between the pairs of the two datasets.

A significant p-value indicates that the differences between the datasets are significant.

Wilcoxon rank sums test - non-parametric 2 sample t-test

- Test for differences between two independent samples...

```
>>> stats.ranksums(dataLst1, dataLst2)
```

```
(20884.0, 0.26079)
```

z-statistic

p-value
(two-tailed)

Note: Input datasets can
have different lengths

- Difference is significant if p-value is less than threshold (e.g. 0.05)

<http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ranksums.html#scipy.stats.ranksums>

10

The Wilcoxon rank sums test is a non-parametric version of the 2 sample t-test. It tests for differences between two independent datasets; the datasets do not need to be normally distributed.

A significant p-value indicates that the differences between the datasets are significant.

Kruskal-Wallis test - non-parametric ANOVA

- Tests for differences among 3 or more independent samples...

```
>>> stats.kruskal(dataLst1, dataLst2,...,dataLstN)  
(353.577, 1.665e-77)
```

H-statistic

two-sided p-value

Note: Input
datasets can have
different lengths

```
>>> dataLsts = [dataLst1, dataLst2,..., dataLstN]
```

```
>>> stats.kruskal(*dataLsts)
```

alternate
syntax

- Difference is significant if p-value is less than threshold (e.g. 0.05)

<http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html#scipy.stats.kruskal>

11

The Kruskal Wallis test is equivalent to a non-parametric ANOVA – it checks for differences among the means of 3 or more datasets. The datasets do not need to be normally distributed.

A significant p-value indicates that the differences between the datasets are significant.

Ansari-Bradley test for equality of scales - 2 samples

- Non-parametric
- Tests if scales (i.e. spread of distribution) are equal for two independent samples...

```
>>> stats.ansari(dataLst1, dataLst2)
```

```
(101669.0, 0.36977)
```

AB-statistic

p-value (2-tailed)

Note: Input
datasets can have
different lengths

- Difference in scale is significant if p-value is less than threshold (e.g. 0.05)

<http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ansari.html#scipy.stats.ansari>

12

The Ansari test tests if the spread of the distribution (e.g. standard deviation) is equal for two independent datasets.

A significant p-value indicates that the spread of distribution is different for the two datasets.

Levene test for equality of scales - more than 2 samples

- Non-parametric version of Bartlett test.
- Tests if scales (i.e. spread of distribution) are equal for 3 or more independent samples...

```
>>> stats.levene(dataset1, dataset2, ..., datasetN, center = 'median')  
(102.199, 1.991e-41)
```

W-statistic

p-value (2-tailed)

other options:
'mean', 'trimmed'

- Difference is significant if p-value is less than threshold (e.g. 0.05)

Note: Input
datasets can have
different lengths

<http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.levene.html#scipy.stats.levene>

The Levene test checks for equal scales for 3 or more datasets.

A significant p-value indicates that the spread of distribution is different for at least one of the datasets.